Étude du transcriptome des rétrovirus endogènes humains et implications fonctionnelles : applications à la recherche de marqueurs diagnostiques de cancers

Philippe Perot

Université Claude Bernard Lyon 1

UNIVERSITÉ DE LYON

N° d'ordre : 228-2012 Année 2012

THESE DE L'UNIVERSITE DE LYON
Délivrée par
L'UNIVERSITE CLAUDE BERNARD LYON 1

ECOLE DOCTORALE
Biologie **M**oléculaire, **I**ntégrative et **C**ellulaire

DIPLOME DE DOCTORAT
(arrêté du 7 août 2006)

TITRE :

ETUDE DU TRANSCRIPTOME DES RETROVIRUS ENDOGENES HUMAINS ET IMPLICATIONS FONCTIONNELLES : APPLICATIONS A LA RECHERCHE DE MARQUEURS DIAGNOSTIQUES DE CANCERS

Soutenue le jeudi 29 novembre 2012

par

PHILIPPE PEROT

Directeur de thèse : Dr. François MALLET

JURY :

Dr. Marc SITBON	Rapporteur
Dr. Pascale LESAGE	Rapporteur
Pr. Fabien ZOULIM	Examinateur
Dr. Gaël CRISTOFARI	Examinateur
Dr. François MALLET	Directeur de thèse

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. François-Noël GILLY**
Vice-président du Conseil d'Administration	M. le Professeur Hamda BEN HADID
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur Philippe LALLE
Vice-président du Conseil Scientifique	M. le Professeur Germain GILLET
Secrétaire Général	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Administrateur provisoire : M. le Professeur G. KIRKORIAN
UFR d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA.
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. FARGE

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. le Professeur F. De MARCHI
Département Biologie	Directeur : M. le Professeur F. FLEURY
Département Chimie Biochimie	Directeur : Mme le Professeur H. PARROT
Département GEP	Directeur : M. N. SIAUVE
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur A. GOLDMAN
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : Mme S. FLECK
Département Sciences de la Terre	Directeur : Mme la Professeure I. DANIEL
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. C. COLLIGNON
Observatoire de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. P. FOURNIER
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. C. VITON
Institut Universitaire de Formation des Maîtres	Directeur : M. R. BERNARD
Institut de Science Financière et d'Assurances	Directeur : Mme la Professeure V. MAUME-DESCHAMPS

Réalité, sortie de secours
Par Marc-Antoine Mathieu.
Mûr peint, rue de Beaulieu, Angoulême.

Liste des figures

Liste des tableaux

Principales abréviations

ACP	Analyse en composantes principales
ADN	Acide désoxyribonucléique
ARN	Acide ribonucléique
ALV	Avian leukosis virus
BaEV	Baboon endogenous virus
BLAST	Basic local alignement search tool
BLV	Bovine leukemia virus
bp	Base pair
CV	Coefficient de variation
cyt	Cytoplasmic tail
DC-SIGN	Dendritic cell specific intercellular adhesion molecule-3-grabbing non integrin
DRE	Digital rectum examination
en	Endogenous
EnCa	Endometrial carcinoma
Env	Envelope
ER	Endoplasmic reticulum
ERV	Endogenous retrovirus
ES	Embryonic cell
EST	Expressed sequence tag
Exo	Exogenous
FcEV	Felis catus endogenous retrovirus
FDR	False discovery rate
FOB	Fecal occult blood
FP	Fusion peptide
Gag	Group specific antigen
GCM	Glial cell missing
GDE	Gène différentiellement exprimé
GPI	Glycosylphosphatidylinositol
h	Human
HELLP	Hemolysis, elevated liver enzymes and low platelets
HERV	Human endogenous retrovirus
HERV-gDB	Human endogenous retrovirus genome database
HFV	Human foamy virus
HML	Human MMTV-like
HIV	Human immunodeficiency virus
HRM	High resolution melting
HTDV	Human teratocarcinoma-derived virus
HTLV	Human T-cell leukemia virus
IDDM	Insulino dependent diabetes mellitus
JSRV	Jaagsiekte sheep retrovirus
kb	Kilobase
KoRV	Koala retrovirus
LINE	Long interspersed element
LTR	Long terminal repeat
m	Mouse
MaLR	Mammalian apparent LTR-retrotransposon
MAO	Morpholino antisense oligonucleotide
MD-PCR	Multiplex degenerate PCR
MFSD2	Major facilitator superfamily domain-containing protein 2
MLV	Murine leukemia virus

MM	Mismatch probe
MMTV	Mouse mammary tumor virus
MPMV	Mason-Pfizer monkey virus
MPSS	Massive parallel signature sequencing
MS	Multiple sclerosis
MSRV	Multiple sclerosis associated retrovirus
NGS	Next generation sequencing
NO	Nitric oxide
OASIS	Old astrocytes specifically induced substance
OLISA	Oligo sorbent array
ORF	Open reading frame
PBMC	Peripheral blood mononuclear cell
PBS	Primer binding site
PBS	Phosphate buffer saline
PCR	Polymerase chain reaction
PcRV	Papio cynocephalus retrovirus
PE	Preeclampsia
PERV	Porcine endogenous retrovirus
PM	Perfect-match probe
Pol	Polymérase
PSA	Prostate specific antigen
RBD	Receptor-binding domain
RD114	A feline endogenous retrovirus
RMA	Robust multiarray average
RT	Reverse transcriptase
SAGE	Serial analysis of gene expression
SAM	Significance analysis of microarrays
SEP	Sclérose en plaques
SERV	Simian endogenous retrovirus
SINE	Short interspersed element
SIV	Simian immunodeficiency virus
SNP	Single nucleotide polymorphism
SNV	Spleen necrosis virus
SP	Signal peptide
SRV	Simian retrovirus
SSH	Suppression substractive hybridization
SU	Surface unit
TM	Transmembrane unit
tm	Transmembrane domain
URE	Upstream regulatory element
UV	Ultra-violet
WDS	Walleye dermal sarcoma
XMRV	Xenotropic murine leukemia-related virus

Introduction générale

Au début des années 1950, Barbara McClintock, se basant entre autres sur l'observation des patrons de couleurs de plants de maïs, présente la théorie que des éléments génétiques doivent avoir la propriété de se déplacer d'un chromosome à un autre. Longtemps taxée de ridicule, la notion de gènes sauteurs sera récompensée trente ans plus tard par le prix Nobel de médecine, décerné à la scientifique américaine pour sa découverte des transposons. Dans le même temps, la première moitié du siècle est marquée par une curiosité grandissante envers les 'agents pathogènes filtrants', caractérisés initialement sur des plants de tabac infectés par le virus de la mosaïque. Le lien entre éléments génétiques mobiles et virologie ne s'officialisera que plus tard, avec la co-découverte en 1970 par Temin et Baltimore de la transcriptase inverse des rétrovirus, complétant alors une vision de l'instabilité des génomes où des échanges de matériel génétique intrinsèques peuvent s'additionner aux transferts horizontaux de gènes venus de l'extérieur.

Généralement considérée comme un événement rare, l'infection d'une cellule de la lignée germinale par un rétrovirus exogène peut aboutir à un transfert vertical des gènes proviraux à la descendance. Si l'hôte survit aux effets de cette intégration, le provirus devient alors une séquence rétrovirale endogène (ERV) et, au fil des générations, se fixe dans une population par un mode de propagation mendélien. Au cours de l'évolution, la réplication d'un rétrovirus endogène parental par réinfections et rétrotranspositions successives peut conduire à l'augmentation du nombre de séquences paralogues au sein du génome hôte. Parallèlement, sous l'effet de la pression de sélection, des événements mutationnels (insertions, délétions, substitutions, recombinaisons) aboutissent à la formation de familles multicopies complexes. En 2001, le séquençage du génome humain a révélé que près de la moitié de notre ADN dérive d'éléments transposables, et que les rétrovirus endogènes en particulier représentent 8 % de la chromatine, soit un ensemble approximatif de 200 000 séquences individuelles (International Human Genome Sequencing Consortium 2001). D'un point de vue phylogénétique, une quarantaine de familles a pu être caractérisée mais il est estimé que notre patrimoine rétroviral en est composé d'une centaine, résultant probablement d'autant infections indépendantes ou de vagues de réinfections passées.

Aujourd'hui, chez l'homme, toutes les séquences HERV sont défectives pour la réplication, ce qui a longtemps valu aux rétrovirus endogènes le quolibet de 'junk DNA' [1]. Cette vision change pourtant progressivement, et notamment avec les démonstrations expérimentales que des protéines

[1] La traduction de junk DNA varie, selon les tempéraments, d'ADN poubelle à fatras génétique. Avec un peu de retenue on pourra s'en remettre à la vision de François Jacob qui voyait la biologie moléculaire comme un bricolage de l'évolution.

d'enveloppes domestiquées, appelées Syncytines en référence aux syncytium qu'elles forment, peuvent avoir une fonction physiologique dans le développement placentaire de certains mammifères. Par ailleurs, une littérature à présent abondante décrit des phénomènes de réactivation des HERV en contextes pathologiques, allant de la transcription à la synthèse de protéines et, parfois, à l'observation de particules, sans toutefois que des liens étiologiques clairs ne soient établis. Plus largement, les séquences LTRs présentes en abondance dans le génome, et qui sont des promoteurs rétroviraux naturels, peuvent contribuer à la régulation de l'expression des gènes cellulaires durant les phases de développement embryonnaire ou de différenciation cellulaire des tissus adultes.

L'étude de l'expression des rétrovirus endogènes humains est fortement contrainte par la nature répétée des séquences, ce qui se traduit souvent par des conclusions au niveau des familles HERV plutôt qu'au niveau des séquences individuelles qui les constituent. Nous avons donc cherché à mieux caractériser l'expression des rétrovirus endogènes humains à l'échelle du locus, afin de contribuer, d'une part, à la compréhension des rôles biologiques des HERV et, d'autre part, d'aider à l'identification de marqueurs de pathologies cancéreuses. Pour cela, une puce à ADN a été développée au laboratoire, permettant d'étudier l'expression individuelle de 5 500 loci HERV (dont plus de 4 000 LTRs) tout en rendant possible une lecture fonctionnelle de la transcription. Ainsi, la première approche de ce travail a consisté à établir des règles d'expression des HERV à partir d'un ensemble de tissus sains et cancéreux. Dans un second temps, nous avons cherché à replacer les connaissances et les outils fondamentaux en contexte clinique, en réalisant en particulier les premières étapes d'un projet de recherche de marqueurs moléculaires du cancer la prostate.

Le **chapitre bibliographique** de ce manuscrit s'organise en quatre parties.

Les notions de base de la rétrovirologie (cycle d'un rétrovirus, nomenclature) seront d'abord présentées sous l'angle du processus d'endogénisation dans une revue introductive rédigée en anglais. Au travers des exemples bien documentés des Syncytines, ce préambule aura pour but d'amener les différents niveaux de régulation de l'activité des HERV. Le cas illustré de la domestication de la Syncytine-1 permettra, en particulier, d'aborder les notions du contrôle de l'expression par les propriétés de séquences et d'environnement génomique, incluant une connaissance sur les facteurs de transcription, les stratégies d'épissages ainsi qu'un niveau de contrôle épigénétique. Une vue des adaptations protéiques et de leurs conséquences fonctionnelles (fusion, immunomodulation) sera également présentée, avant d'évoquer brièvement les conséquences d'une dérégulation de la Syncytine-1 dans le placenta et de sa réactivation ectopique associée à différentes pathologies.

Puis les contextes d'expressions des HERV seront étendus dans une seconde partie centrale. Les implications physiopathologiques de différentes familles de rétrovirus endogènes humains, et

parfois de loci individuels, feront l'objet d'un développement qui illustrera la complexité des modalités d'expression des HERV et de leur régulation.

Un paragraphe méthodologique et technologique de l'analyse transcriptionnelle en général, puis du transcriptome HERV en particulier, viendra ensuite souligner les apports et les limites de deux grands procédés de détection que sont la RT-PCR quantitative et les puces à ADN.

Enfin, les problématiques principales de la cancérologie clinique seront exposées, ainsi qu'une démarche pour la recherche de marqueurs moléculaires. Une vue simplifiée des besoins associés à la recherche de nouveaux marqueurs diagnostiques et pronostiques du cancer de la prostate amènera les enjeux de l'étude clinique poursuivis durant ce travail de thèse.

Le **volet expérimental** de la thèse s'articule en trois temps.

D'abord, les outils d'analyse développés au laboratoire et utilisés pour l'étude du transcriptome HERV seront détaillés. Les étapes clés de création et d'amélioration de la puce à ADN haute densité du laboratoire dédiée aux séquences de rétrovirus endogènes humains y seront explicitées, où les particularités de son procédé analytique seront en particulier mis en avant. Une méthode utilisée pour la validation des résultats d'expression et qui associe un procédé de fusion à haute résolution (HRM) à une étape de RT-PCR sera également décrite.

Les résultats biologiques de l'expression des HERV obtenus à partir d'un panel de tissus sains et cancéreux seront alors présentés et associés à un premier niveau de discussion. L'expression globale ainsi que la mise en évidence de profils d'expression tissu-spécifiques au niveau locus et d'éléments réactivés en contextes cancéreux dressent une première esquisse d'un paysage transcriptomique HERV. A cela, l'identification de séquences LTRs promotrices et polyA, en lien avec un environnement génomique, enrichie la vision des interactions virus/hôte. A ce stade, une validation des résultats d'expression a été réalisée au travers d'une étude de la variabilité inter individuelle de séquences HERV-H identifiées dans le cancer du côlon.

Une fois l'intérêt de l'utilisation du répertoire HERV en oncologie renforcé, nous avons cherché à positionner les outils d'étude et les connaissances acquises dans le contexte clinique d'un travail sur échantillons complexes. Des prélèvement d'urines après massage prostatique ont été recueillis auprès de patients atteints de cancer de la prostate dans le but de mettre au point un protocole de traitement de l'échantillon adapté à un projet de recherche de marqueurs transcriptomiques. Nous avons alors pu conduire une étude pilote sur 45 patients dans le but d'identifier des marqueurs diagnostiques et pronostiques du cancer de la prostate. L'utilisation de référents méthodologiques à différents niveaux a permis de mettre au point puis d'appliquer des critères qualitatifs aboutissant à l'identification sur puce à ADN, et à la confirmation par RT-PCR quantitative, d'une expression différentielle de quelques séquences HERV, associée au groupe de patients cancéreux.

La **discussion générale** prolongera la remise en perspective des résultats de ce travail par rapport aux connaissances actuelles et fera des suggestions pour aller plus loin dans la compréhension de la régulation de l'expression des HERV. Une réflexion sur les niveaux de difficultés associés à la recherche translationnelle sera proposée, ainsi qu'une vision élargie de la place des outils du laboratoire dans les problématiques technologiques et scientifiques qui se posent à l'heure actuelle.

En **annexes**, le lecteur trouvera (i) une revue sur les mécanismes comparés de la fusion rétrovirale et des Syncytines (ii) une démarche analytique type d'analyse de puces à ADN et (iii) des compléments techniques sur les études de mises au points de protocoles de transcriptomique à partir d'échantillons urinaires.

Viruses: Essential Agents of Life
Witzany, Günther (Ed.)
ISBN 978-94-007-4898-9
November 2012

Springer
the language of science

From Viruses to Genes: Syncytins

Philippe Pérot[1], Pierre-Adrien Bolze[1,2] and François Mallet[1§]

[1] Laboratoire Commun de Recherche Hospices Civils de Lyon – bioMérieux, Cancer Biomarkers Research Group, Centre Hospitalier Lyon Sud, 69495 Pierre-Bénite cedex, France.
[2] Université Claude Bernard Lyon 1, Hospices Civils de Lyon, Centre Hospitalier Universitaire Lyon Sud, Centre de Référence des Maladies Trophoblastiques, 69495 Pierre-Bénite cedex, France.
§ Corresponding author; Email address: francois.mallet@biomerieux.com

On behalf of past and present members of the Mallet's group, we would like to dedicate this chapter to the memory of our colleague and friend Olivier Bouton who substantially contributed to the human and scientific adventure that was the MSRV/HERV-W/ERVWE1 discovery.

Abstract

The content of 5-90 million years old retroviruses and even older retrotransposons of animal genomes and the wide variety of modern retroviruses infecting the same range of species suggest that these elements can be assimilated to shuttle across evolution. A snapshot taken a few decades ago showed us the capture of cellular proto-oncogenes by infectious elements, representing the dark side of the communication between the worlds of viruses and animals. Another snapshot we took more recently shows multiple captures by animal genomes of envelope genes originating from infectious retroviruses, illustrating a phenomenon of convergent evolution. This could be seen as the bright side of these relations as those envelopes were shown to be involved in the earlier steps of human development, i.e. fusion of placental syncytiotrophoblastic layer, therefore they were dubbed Syncytins. Sequencing of more and more animal genomes allowed comparative genomic analyses that revealed how these envelopes have been domesticated in human, mouse, goat, rabbit, etc. More generally, we illustrate in this chapter how close are the viral and animal genome worlds and, focusing mainly on the hominoid ERVWE1 locus encoding Synctin-1, how the different proviruses encoding Syncytins have been domesticated to achieve placental functions. Influence of the chromosomal integration context, the epigenetic control and the splicing strategy upon transcription, and protein maturation processes as well will be discussed in order to illustrate what makes these nowadays genes different from their ancestral infectious counterpart. The price to pay for this beneficial invasion will be illustrated by the possible implications of Syncytin-1 in a wide range of diseases. Last, the apparent stringency of placental regulation will await to be challenged as regard to the evidence of expression in other physiological fusogenic contexts such as myoblasts and osteoclasts.

Keywords: Retrovirus, endogenous retrovirus, Syncytins, domestication

I.1.1. Introduction

The most convincing clue of filiation between all living things is likely the sharing of nucleic acid molecules. Thus, the RNA world is for biologists quite similar to the primordial soup for astrophysicists: some kind of constructive thought experiment to travel back in time. According to the prevailing theory, primitive cells were possibly very simple membrane structures containing RNA nuons (i.e. any distinct nucleic acid) that have undergone escape and uptake of molecules (Brosius and Gould 1992). Cell division and fusion ensured the dynamic to trying out new nuons, and consequently genetic exchanges became early one core evolutionary force (Brosius 2005). Retroviruses can be seen as RNA shuttles ensuring genetic exchanges from one species genome to another. How old are retroviruses is a difficult issue, but since Howard Temin formulated the initial hypothesis that retroviruses evolved from cellular moveable genetic elements (Temin 1980), knowledge on genomic oncogenes capture and the resulting emergence of infectious transducing retroviruses has made significant steps forward (Pedersen and Sørensen 2010). Another type of capture exists between retroviruses of distant species, consisting in the swapping of envelopes between species living in the same environment or linked by the food chain. For example, the RD114 infectious endogenous virus comes from two genetic recombinations resulting in two *env*-captures. First, the SERV *env* was captured by the PcRV leading to the BaEV. Second, the acquisition of BaEV *env* by FcEV led to the emergence of RD114 virus (Kim et al. 2004).

The strongest candidates for developmentally regulated cellular fusogens in mammals are Syncytins, a family of single-pass transmembrane envelope proteins, which contribute to cell-cell fusion leading to placental syncytiotrophoblast at least in higher primates, rodents, lagomorphs and sheeps (Pérot et al. 2011). They consist of domesticated endogenous retroviral envelope glycoproteins whose fusion properties depend on the initial recognition of a specific receptor. Syncytins appear to group relatively distinct actors that may exhibit common characteristics leading to membrane fusion and hence are good illustrations of the various evolutionary pathways taken to establish similar but different structures with convergent roles.

During the first 10 years of the 21st century, decoding the genomes, notably that of mammals, showed that besides the Syncytins, there were many other ERV envelopes genes for which no function has yet been assigned. For example, deciphering the human genome (International Human Genome Sequencing Consortium 2001) permitted to identify 16 almost intact envelopes ORFs (Blaise et al. 2003; Blaise et al. 2005), in addition to the well described two human Syncytins. Beyond theses envelope proteins, without unequivocal evidence of infectious agent, retroviral particles were observed in physiological (Lyden et al. 1994) and pathological situations in man (e.g. Boller et al. 1993; Perron et al. 1989), suggesting that endogenous elements can reach a similar complexity level as infectious retroviruses. The degrees of difficulty vary to understand the origin of these retroviral elements whether we consider the simplest level of complexity through a single gene or the ultimate degree of complexity brought with the particles. Outstandingly, the current knowledge on Syncytins embraces at least three levels of complexity. First, in term of architecture, the placenta is probably the more variable organ within mammals (Bernirschke K, Comparative placentation at http://placentation.ucsd.edu). Second, Syncytins recognize specific and highly function-divergent and unrelated receptors as observed in human (Blond et al. 2000; Esnault et al. 2008; Lavillette et al. 2002) and rodents (Dupressoir et al. 2005). This illustrates that the proteins involved in cell-cell fusion, such as Syncytins partner receptors, are likely to play pleiotropic roles in other cellular processes like the transport of small molecules, but also the modulation of membrane structures, and that their functions are being achieved through the coupling of these proteins to different upstream and downstream effectors. Third, Syncytins were shown to exhibit other functions than fusion, such as immunomodulation (Mangeney et al. 2007), receptor interference (Blond et al. 2000; Ponferrada et al. 2003) anti-apoptosis (Knerr et al. 2007; Strick et al. 2007) and cellular proliferation (Larsen et al. 2009; Strick et al. 2007), these functions being not shared by all these proteins. Regarding retroviral particles, they could derive either from a single locus or from several loci *via* transcomplementation processes. This is supported by the presence among the hundreds thousand retroviral elements of the mouse genome (Mouse Genome Sequencing Consortium 2002) but also of the human genome (International Human Genome Sequencing Consortium 2001) of almost complete proviruses and significant number of still intact coding sequences for *gag* and *env* genes (Villesen et al. 2004).

In this chapter we seek to illustrate how one element may move from an infectious virus to become an entity transmitted as a gene. We will show how the study of the placenta allowed to overcome conceptual leaps in understanding the role of HERVs, given that this tissue is historically a privileged place for HERVs expression as well as proteins and particles detection. Then we will describe in detail what are the domestication mechanisms of the Syncytin-1, including the genomic integration context, the control of the transcription and the protein maturation, to see what makes these nowadays genes different from infectious retroviruses. Ultimately, in an attempt to decode the underlying regulatory mechanisms, we will look at the expression and functions of human Syncytins in pathologies and specify how they behave outside of the domestication scene.

I.1.2. When Rous met Mendel

I.1.2.1. From viruses to genomes

I.1.2.1.1. The retroviral life cycle

The rare event that represents the infection of a germ line cell by an exogenous retrovirus leads to the integration into the host genome of a retroviral DNA, or provirus, that becomes part of the genetic heritage of the host. Therefore, this endogenous provirus is transmitted to the next generation in a Mendelian way. The parental infectious retrovirus is a diploid RNA virus whose 8 to 10 kb compact genome consists of four major genes *gag*, *pro*, *pol* and *env* encoding the proteins required for its replication life cycle, and flanked by 5' R-U5 and 3' U3-R untranslated regions. The *gag* gene encodes matrix, capsid and nucleocapsid proteins necessary for viral RNA encapsidation and particle formation. The *pro* gene encodes a protease required for the cleavage of Gag-Pro-Pol and Gag polypeptidic precursors and also, in the case the Gammaretroviridae genera, for the final Env maturation step leading to fusion competency. The *pol* gene encodes the major viral enzyme machinery, including two enzymes, an RNA-dependent DNA polymerase named reverse transcriptase and an integrase, both sequentially required for the successful conversion of the viral RNA into the proviral integrated DNA form. Last, the *env* gene encodes viral envelope glycoproteins that confer virus infectivity, i.e. receptor recognition *via* the subunit named SU and virus-cell membranes fusion *via* the subunit named TM. In addition, the TM subunit contains motives that is likely to confer to retroviruses immunosuppressive properties. The **Figure I-1 a** shows schematically the replication cycle of a simple infectious retrovirus in order to point out which HERV components, either proteins or regulatory elements involved in transcription initiation and termination, can fulfil a function by their contribution to a physiological or pathological role. Briefly, following the entry into the cell, the viral RNA is reverse transcribed into DNA by the viral RT using a cell specific tRNA as a primer which hybridizes with the PBS region located at the R-U5 and *gag* junction of the retroviral genome. The resulting double-stranded DNA, which contains at each end a non-coding LTR sequence derived from R-U5 and U3-R viral sequences (see locations **in Fig. I-1 b**), is integrated into the genome of the host cell through the action of the viral integrase. The expression of the proviral DNA is then becoming dependent on the host cell machinery that provides the transcription factors required to activate the 5'LTR. The 5'LTR plays the role of promoter and enhancer sequence conferring tissue-specific expression. The distal 3'LTR contains the polyadenylation signal terminating the transcription.

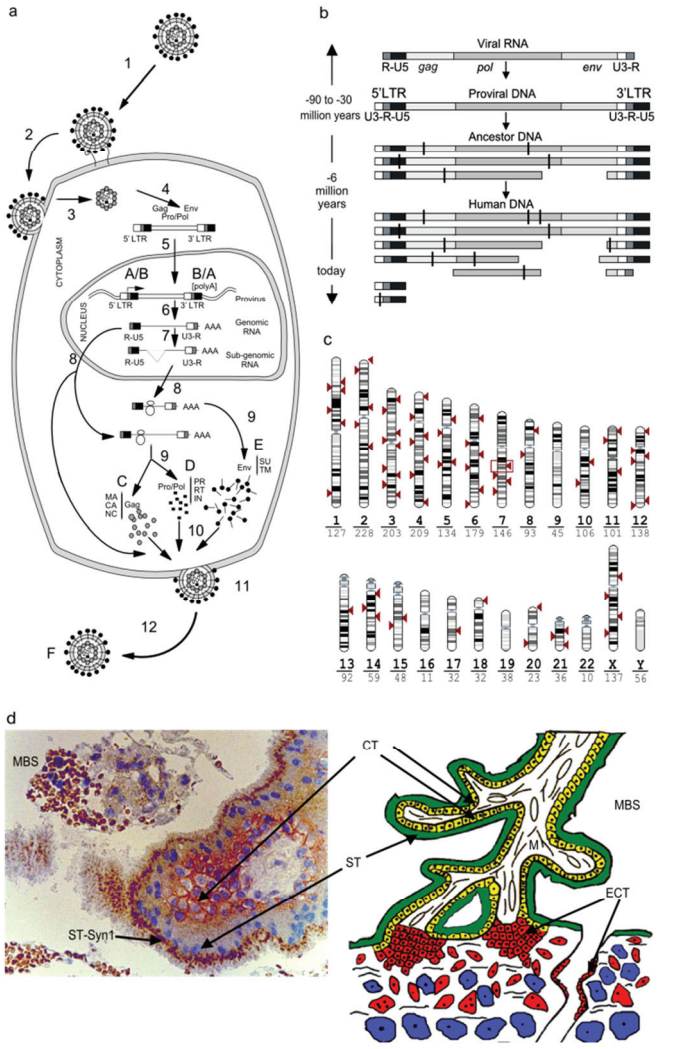


Figure I-1 From the ancestral infectious retrovirus to the contemporary human endogenous retroviruses family. (**a**) Schematic representation of the replication life cycle of an infectious retrovirus, illustrating the functions achieved by the various retroviral elements during the steps of the cycle, and allowing the identification of endogenous retroviral elements which may be involved in a physiological or pathological function. -1- Binding to a cell receptor *via* the viral envelope surface subunit (SU) (Env, ●–), -2- fusion of viral and cell membranes *via* the viral envelope transmembrane subunit (TM), leading to -3- the entry of the capsid in the cytoplasm, -4- the conversion of the viral RNA to cDNA and DNA by the reverse transcriptase (RT, ■) -5- the integration of the provirus flanked by two identical LTRs in the cellular DNA (provirus). -6- Transcription (5'LTR promoter function) controlled by host cell transcription factors and -7- production of genomic and subgenomic (spliced) mRNA -8- transported to the cytoplasm, -9- translation and production of the polyproteic capsid (Gag,◍) and enzymatic machinery (RT, integrase, protease) from the genomic transcript, and envelope, from subgenomic transcript, -10- assembly of the genomic RNA and viral proteins leading to -11- the budding and -12- release of virions which can infect new cells. Identification of elements that may be involved in a function in endogenous retroviruses. -A- Promoter function (U3 region of LTRs) can lead to the synthesis of RNA coding for retroviral proteins (5'LTR) or non-retroviral (solo LTR or 3'LTR). -B- Polyadenylation signal (R region of LTRs). -C- Gag proteins can form particles (intra or extra cellular) able to encapsidate genomic-like RNA that can be reverse transcribed (via RT) and re-integrated. -D- Reverse transcriptase (RT) activity that may contribute to the re-integration of retroviral (RT-HERV and RT-LINE) or cellular (RT-LINE) genes deleted from their introns (pseudogenes). -E- Envelope protein can be expressed at the cytoplasmic membrane (intra and extra cellular portions), interact with a receptor and fuse cell-cell membranes, module immunity *via* an immunosuppressive motif. -F- Some HERV loci produce enveloped (or not) particles that can potentially deliver a distant signal in the body; genomes identified so far, however, are all defective for replication. (**b**) Constitution of a HERV family. The proviral DNA, integrated several millions years ago into the DNA of a germ cell, spread mainly by reinfection and retrotransposition and the different offspring loci went through a mutagenic process (symbolized by a vertical line or a deletion) during evolution. No contemporary copy is infectious as shown by i) the frequency of *env* gene deletion, ii) the absence of the U3 region in 5'LTRs on certain entities, iii) the existence of entities deleted from the majority of their sequences and iv) solo LTRs. (**c**) Representation of the chromosomal distribution of genetic entities of the HERV-W family. The position and the number of elements per chromosome are shown; it should be noted on chromosome 7 (arrow) the presence of the ERVWE1 locus containing the unique complete Env open reading frame, i.e. Syncytin-1. (**d**) Immunohistochemical detection of Syncytin-1 protein (SC-Syn1) at the apical syncytiotrophoblast (ST) microvillus membrane of a ten-week gestation normal placenta. Note that desmoplakin, a protein of the desmosomiale plaque involved in intercellular junctions, is absent from the syncytiotrophoblast fused tissue and lines the plasmatic membranes of the cytotrophoblasts (CT). Red blood cells in the maternal blood space (MBS); extravillous cytotrophoblast cells (ECT).

During the evolution, the founding-captured HERV provirus, and latter its son elements, is replicated by mechanisms that essentially rely on transcription, i.e. reinfection and retrotransposition. Due to the general absence of selection pressures, most of the elements contain disruptive mutations, like substitutions, insertions and deletions, in at least one of the structural genes of the provirus. Thus, the preferential loss of the *env* gene of many HERV elements is a common phenomenon that may reflect the unnecessary requirement of this gene once the barrier of species is crossed over (Boeke and Stoye 1997). Nevertheless, open reading frames can persist and lead to protein synthesis or even non-infectious particles. In addition, each family contains numerous solitary LTRs, resulting from the loss of full coding sequences by recombination between two flanking LTRs (Lower et al. 1996). All these mechanisms lead to complex multicopy families each consisting of heterogeneous elements (**Fig. I-1 b**). More, all loci of the contemporary HERV families are defective for replication, which means they have lost their infectious properties and are engaged in a vertical mode of transmission exclusively. However, the processes of spread within the genomes has generated, in addition to a significant level of complexity, a generally wide distribution of the sequences as illustrated by the chromosomal location of the elements belonging to the HERV-W family (Blond et al. 1999) (**Fig. I-1 c**). Among these HERV-W elements, there is one located on chromosome 7, ERVWE1, that became a bona fide gene (**Fig. I-1 c**) (Mallet et al. 2004) and producing a retroviral envelope glycoprotein involved in hominoid placental physiology, as illustrated in **figure I-1 d**.

I.1.2.1.2. Forgotten territories seeking an identity

The definition of a precise nomenclature for animal ERV families is a difficult task in the absence of function or obvious pathology associated with these retroviruses, as opposed to infectious retroviruses. Yet, the development of a systematic nomenclature was tentatively proposed (Blomberg et al. 2009; Mayer et al. 2011). Retroviral classification was historically based on virion morphology observed by electronic microscopy during maturation and assembly of particles (Coffin 1992). Accordingly, retroviruses were designated A-, B-, C- or D-type. The current-usage classification of HERV is based on the PBS sequence located downstream of the 5'LTR, or its similarity to the infectious retroviruses PBS, which is recognized by a specific tRNA. A code based on the letter that refers to the amino acid recognized by the tRNA is applied to become a suffix, e.g. HERV-H exhibits a PBS which is recognized by a histidine (H) tRNA, the HERV-W PBS is homologous to the PBS of the avian retrovirus tryptophan (W) tRNA. However, some names sometimes coexist such as ERV-3 known as HERV-R, or ERV-9 which also share an arginine (R) PBS. This nomenclature can also be misleading, for instance the super-family HERV-K contains 11 phylogenetically distinct sub-groups referred to as HML-1 to HML-11 (Blikstad et al. 2008; Subramanian et al. 2011), what can let think that all the members share the same PBS, yet HML-5 members are not primed by a lysine (K) tRNA as the name would suggest, but must likely by a methionine (M) or isoleucine (I) tRNA (Gifford and Tristem 2003; Lavie et al. 2004). The International Comity of Taxonomy has now established seven genera of Retroviridae, grouping exogenous and endogenous retroviruses as well, based on sequence homologies of the *pol* region: Alpharetoviruses thus correspond to the avian type C retroviruses (ALV), Betaretroviruses to type B (MMTV, HERV-K) and D (SRV-1) retroviruses, Gammaretroviruses to mammalian type C retroviruses (MLV, HERV-E, HERV-W), Deltaretroviruses to the ancient group composed of HTLV and BLV, Epsilon retroviruses to the WDS viruses family, Lentiviruses contains HIV and SIV and Spumaviruses include HFV and HERV-L (van Regenmortel et al. 2000).

A complete view of the (H)ERV landscape was expected from the publication of several mammalian genomes, including human (International Human Genome Sequencing Consortium 2001) and mouse (Mouse Genome Sequencing Consortium 2002). Although the human and mouse genomes contain essentially different retroviral families, they both display a huge but similar amount of endogenous retroviruses, reaching 8.5% and 9% of their euchromatin, respectively. More precisely, the human genome contains 203,000 copies resulting from about 100 independent infectious events, although only about 40 groups have been studied (Benit et al. 2001; Jurka et al. 2005; Mager and Medstrand 2003; Tempel et al. 2008; Tristem 2000), but it also contains some 240,000 MaLR elements which are considered as the ancestor of infectious retroviruses. It is crucial to appreciate how these hundreds of thousands of retroviral sequences constitute a mass significantly greater than all the human genes, a number currently estimated to range between 20,000 and 25,000 (International Human Genome Sequencing Consortium 2004).

I.1.2.2. Crossing the border

'Just leave the woods
and you'll improve your lot'

Interplay between the primitive virus world and the eukaryotes domain could be observed at the *env* level. Thus, infectious retroviruses appear to have burst, getting out from the genomes of our far ancestors by transcomplementation of cellular retrotransposons with viral envelopes genes (Malik et al. 2000). In turn, endogenous retroviral sequences progressively undergo genetic drift and they spread throughout the genomes as describe previously. As a consequence the separation between endogenous and exogenous retroviruses sometimes is really thin.

I.1.2.2.1. KoRV: ongoing endogenisation

Koala retrovirus provides a unique opportunity to study the process of ongoing endogenisation as it still appears to be spreading through the koala population (Miyazawa et al. 2011; Tarlinton et al. 2008; Tarlinton et al. 2006). Interestingly, infectious viral particles are produced by the endogenous form of KoRV and high levels of viraemia have been linked to neoplasia and immunosuppression (Tarlinton et al. 2005). It remains unclear how the host can react when exogenous and endogenous forms of a virus are coexisting within the genome and his environment. Studies on koala might answer this question.

I.1.2.2.2. HERV-K: almost infectious

In human, the most recent HERV family to have entered the genome, HERV-K, contains tens of almost complete but mutated proviruses that allow the expression of viral proteins which appear able to form retroviral particles. However, due to genetic drift, no complete proviruses able to produce replication-competent and infectious viral particles have been detected. The HERV-K113 locus is the youngest element that belongs to the HERV-K super-family and is still not fixed in the human population as it is only detectable in up to 30% of individuals, depending of ethnicity (Moyes et al. 2005; Turner et al. 2001). HERV-K133 contains intact ORFs for all the viral proteins but does not produce any particles despite *in vitro* potential (Boller et al. 2008; Lee and Bieniasz 2007). Trans-complementation between different HERV-K (HML-2) proviruses could theoretically produce infectious particles, although not demonstrated to date. Interestingly, the infectious potential of HERV-K particles was artificially restored by generating a consensus HERV-K (HML-2) provirus named Phoenix supposed to be the HERV-K family progenitor (Dewannieux et al. 2006). This consensus contained at least 20 amino acid changes on the overall sequence as compared to individual proximal HERV-K loci. By electronic microscopy, this resurrected HERV-K forms viral particles in transfected cells. The budding of its particles is similar to γ- , δ- retroviruses or lentiviruses with no particles preassembling into the cytoplasm.

I.1.2.2.3. MSRV: full story, lack of evidence

MSRV is closely related to the HERV-W family including the Syncytin-1 encoding ERVWE1 locus which is the only W-locus bearing a full-length envelope. The sequencing of ERVWE1 envelope confirmed that the MSRV envelope was not encoded by the ERVWE1 locus (Mallet et al. 2004). It was thus proposed that MSRV particles, if not derived from an as yet uncharacterised exogenous retrovirus, may result from transcomplementation of dispersed HERV-W copies activated simultaneously (Dolei 2005), what appears poorly probable as regards to the HERV-W elements identified in the human genome consensus, unless there is a particular polymorphism uncharacterized to date. An alternative hypothesis would be that point mutations may counterbalance the genetic drift of one or more almost complete HERV-W sequences, reverting them to coding proviruses in multiple sclerosis. In particular, a HERV-W locus located on chromosome Xq22.3 harbors an almost complete ORF for full-length envelope protein but is interrupted by a stop-codon. The reversion of the stop codon in artificial systems led to the successful expression of a reconstituted full-length HERV-W envelope protein sharing very similar post-translational features with the Syncytin-1 (Roebke et al. 2010). Xq22.3 sequencing from blood of 6 MS patients showed that the stop codon is still present at the germinal level (Bouton and Mallet, unpublished data), even if the presence of punctual somatic mutation within acute demyelinating lesions cannot be definitively excluded. Nevertheless, although this locus lacks a 5'LTR promoter element and thus needs an upstream control element unidentified to date (Nellaker et al. 2006), the transcription of the Xq22.3 truncated envelope has been reported several time in PBMC (Laufer et al. 2009; Nellaker et al. 2006) what is a prerequisite for recombination triggering. So it could not be formally excluded that MSRV/HERV-W genome, associated with particles, may result from very complex recombination events involving several loci on distinct chromosomes (Laufer et al. 2009; Roebke et al. 2010).

I.1.2.3. How functions were imagined… then found

I.1.2.3.1. Hypothesis that came from the exosphere

Notwithstanding the often abundant imagination of the scientific community, HERV and transposable elements were first considered as 'junk', 'selfish' or 'parasite' DNA, without any physiological effect. Yet, given the distribution and the nature of these retroviral elements, several functions have gradually been conceived using knowledge gained from retrovirology in decades, and demonstrations have followed that HERV act on their hosts by different mechanisms: (i) HERV may be involved in genomic plasticity during evolution as recombination sites within or between chromosomes (Hughes and Coffin 2001) (ii) they can produce recombination-induced germinal or somatic mutations giving rise to the loss of function of a cellular gene (Blanco et al. 2000; Kamp et al. 2000; Sun et al. 2000) (iii) individual or proviral LTR can modulate the expression of adjacent cellular genes (Cohen et al. 2009; Long et al. 1998; Schulte et al. 1996; Ting et al. 1992) (iv) the expression of HERV proteins with conventional retroviral functions, like fusion, immunomodulation or RNA nuclear export, can influence physiological or pathological conditions of the host (Blond et al. 2000; Boese et al. 2000; Magin et al. 1999; Mangeney et al. 2007).

I.1.2.3.2. Barriers to reach the inner knowledge

The analysis of transcriptional expression of HERV is extremely difficult due to the multicopy nature of these families, although locus-specific approaches like microarrays or PCR coupled with sequencing are developed in some laboratories (Flockerzi et al. 2008; Gimenez et al. 2010; Liang et al. 2012). Indeed, the expression of a family in a given tissue does not generally reflect the expression of all the elements of this family, but rather results in the activation of a limited number of retroviral copies including the case of a single locus. The major determinants of such differences are related to the particular site of integration, methylation status of the LTR, and the susceptibility to cellular factors and environmental stimuli as well. Expression of HERV in reproductive and embryonic tissues may reflect a contribution to the genetic diversity or the physiological development. In contrast, expression of HERV later in the life of the organism may have adverse consequences. Increased HERV transcripts in cancers and autoimmune diseases confirmed that their activity is altered in pathological conditions. Thus, HERV have frequently been proposed as causative cofactors in such pathologies (Löwer 1999). Nevertheless, it remains complex to demonstrate conclusively whether the expression or re-expression of retroviral sequences is a cause or a consequence of the biological context in which it is detected, e.g. HERV-W and HERV-H in multiple sclerosis (Christensen 2010).

I.1.2.3.3. The placenta, where everything converges

In the seventies, electron microscopy has described the presence of virus related particles in placental chorionic villous tissues of humans and primates (Kalter et al. 1975). Further studies then revealed some retroviral characteristics of these particles such as ultrastructural features and RT activity (Lyden et al. 1994). Apart from particles, mRNA expression from different HERV families have been reported early in the placenta (Lower et al. 1996) and was followed by the detection of retroviral envelopes using immunohistochemical techniques in human (Venables et al. 1995) and in baboon (Langat et al. 1999). Together with osteoclasts, skeletal muscles and sperm-oocyte fusion, the placenta is a tissue where cells fuse in physiological conditions what may strongly suggest the involvement of retroviral fusogens (Pérot et al. 2011). More, the ability of HERV and HERV-related sequences to modulate the immune system (Mangeney et al. 2001; Rolland et al. 2006) seemed very adapted to provide new insights on feto-marternal tolerance mechanisms.

I.1.2.3.3.1. Placenta and LTRs

The study and the description of transcriptional mechanisms that involve non-retroviral cellular genes and their neighboring retroviral LTRs were fruitful in the placenta. Indeed, the integration of retroviral elements near some genes provided them various cell expression tropisms, depending on the activation of native non-retroviral or additionally-acquired retroviral promoters. PTN, CYP19A1, NOS3, INSL4 and IL2RB are some significant examples discussed here of genes whose expression in the placenta is solely due to the presence of a promoter LTRs and thus can be regarded as exaptation events (see also Cohen et al. 2009). Expression of pleiotrophin (PTN) in the central nervous system during the perinatal period is controlled by a non-retroviral promoter, whereas expression in the normal trophoblast is controlled by a promoter HERV-E LTR inserted upstream of the first exon (Schulte et al. 1996). CYP19A1, encoding the P450 aromatase, uses a MER21A LTR as placenta-specific promoter in addition to several non-LTR promoters in other tissues (Conley and Hinshelwood 2001; Kamat et al. 1998; Sun et al. 1998; Toda et al. 1996; van de Lagemaat et al. 2003). The nitric oxide synthase 3, NOS3, which mediates VEGF-induced angiogenesis, uses one LTR of the HERV-I family as a predominant promoter in the placenta (Cohen et al. 2009). The 3'LTR of a HERV-K element inserted near the INSL4 (insulin-like growth factor) gene has been exapted as primary promoter and regulates the placental specific expression of this gene during the formation of the syncytiotrophoblast (Bieche et al. 2003; Rawn and Cross 2008). Finally, the cytokine receptor subunit β, IL2RB, involved in the activation of T and NK cells, has been described more recently to rely on the alternative promoter function of a THE1D LTR in the placenta specifically (Cohen et al. 2011). Additional somewhat less major contributions of promoter LTRs to placental gene expression can also be mentioned, like LTRs of the HERV-E family playing a role in the expression of the endothelin receptor B EDNRB (Medstrand et al. 2001) and the Optiz syndrome-associated midline 1 MID1 genes (Landry et al. 2002), given that transcripts from native promoters in these cases are also detected in numerous tissues. Overall, the ability of the LTRs to act as enhancer elements should not be neglected, as it was reported in the case of leptins. The leptin gene, LPT, is expressed in mice adipocytes but the insertion of a MER11 LTR in the natural promoter of LPT confers an activating effect in the human placenta (Bi et al. 1997). Although it remains difficult to identify enhancer LTRs due to the genomic distances that can theoretically disconnect them from their target gene, it is likely that a large number of retroviral enhancers exist throughout the human genome.

I.1.2.3.3.2. Placenta and Syncytins

The keen interest for endogenous retroviral proteins expression in placentas is fed by *in vivo* or *ex vivo* demonstrations that directly link retroviral envelopes with fusion events during the development of many eutherian species. The molecular characterization of the HERV-W family relied on the isolation of placental cDNA clones,

including one complete *RU5-env-U3R-polyA* sequence containing an *env* full length viral ORF (Blond et al. 1999). In 2000, protein truncation tests confirmed that this *env* ORF was unique among the genome and has the coding ability for a putative envelope gene, in association with a functional U3 promoter (Voisset et al. 2000). One year later, Blond and Mi concomitantly associated a HERV-W envelope protein with fusion events in TE671 and BeWo cells, and the name Syncytin was proposed by Mi in reference to the resulting syncytia (Blond et al. 2000; Mi et al. 2000). The amino-acid transporters hASCT2 (Blond et al. 2000) and hASCT1 (Lavillette et al. 2002) were identified as the receptors/fusion partners of Syncytin-1. Note that the connexin 43 was also demonstrated to play an important role in the fusion by interacting with hASCT2 in the syncytiotrophoblast basal membrane (Dunk et al. 2012). Heidmann and colleagues conducted a genome wide screening that identified a second envelope protein, belonging to the HERV-FRD family, and expressed exclusively in the human placenta. They named Syncytin-2 this second fusogenic *env* protein (Blaise et al. 2003) whose receptor is the carbohydrate transporter MFSD2 (Esnault et al. 2008). A similar *in silico* approach was done then in the murine genome, identifying the two coding envelopes genes present as unique copies and with a placenta specific expression Syncytin-A and Syncytin-B (Dupressoir et al. 2005) but without receptor identification to date, and later in the rabbit genome, identifying the Syncytin-Ory1 gene that unexpectedly shares ASCT2 as a receptor (Heidmann et al. 2009). If the situation is much more different in the ovine genome, where approximately 27 copies of endogenous betaretrovirus (enJSRVs) were detected, RT-PCR and *in situ* hybridization clearly indicate a conceptus (embryo/fetus and extra embryonic membranes) localization of enJSRVs *env* transcripts during gestation (Dunlap et al. 2006). JSRV uses the GPI-anchored cell surface HYAL2 protein to enter the cells (Arnaud et al. 2008) (**Fig. I-2 a & b**). Retroviral envelope sequences have also been detected in the placenta of cat, dog, guinea pig, as well as in bovine binucleate cells (Baba et al. 2011; Koshi et al. 2011; Vernochet et al. 2011), although the demonstration of a function remains to be established to date. Although the role of Syncytins in human placentation awaits a definitive demonstration (e.g. infertility associated mutation), knock-out gene experiments in mice clearly achieved this goal in rodent model and demonstrated for the first time the critical role of Syncytin-A in placenta morphogenesis. Using a homologous recombination strategy, Syncytin-A null mouse embryos exhibited growth retardation with an altered placenta labyrinth architecture and died in utero (Heidmann et al. 2009), while Syncytin-B null placenta displays impaired formation of syncytiotrophoblast layer II (Dupressoir et al. 2011). This is consistent with previous *in vitro* works that used specific antibodies and antisense oligonucleotides to show a decrease in syncytia cell formation after Syncytin-A inhibition (Gong et al. 2007).

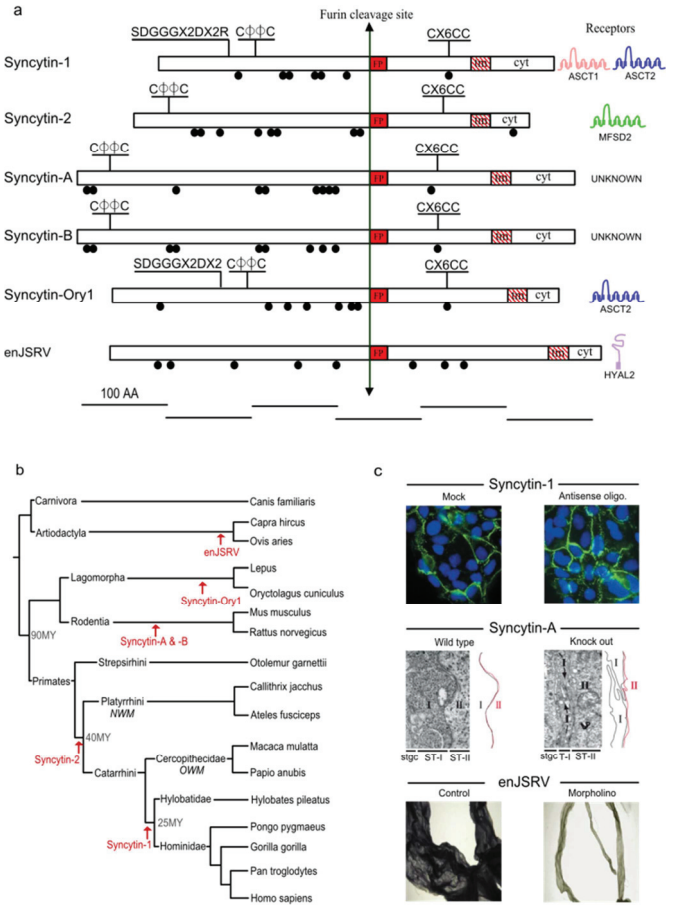


Figure I-2 Structure, phylogeny and fusion capacities of Syncytins involved in placenta development. (**a**) Envelopes structures of Syncytins and schematic representation of their cognate receptors. FP: fusion peptide; tm: transmembrane domain; cyt: cytoplasmic tail. Black dots indicate the predicted N-glycosylation sites. SDGGGX2DX2R, consensus motif conserved in type D retroviral interference group, is indicated in human Syncytin-1 and rabbit Syncytin-Ory1. (**b**) Phylogenetic tree depicting the conservation among species of the six envelope-open reading frames harbouring retroviral canonical motifs (branches of the tree are only illustrative). NWM: new world monkeys; OWN: new world monkeys. (**c**) 1st column: Assays reporting the biological effect of Syncytins. 2nd column: *Ex vivo* or *in vivo* specific inhibition of Syncytins expressions. From top to bottom: Syncytin-1-induced human primary trophoblasts fusion and differentiation results in syncytia formation *ex vivo* (1st column). Inhibition by specific antisense oligonucleotide largely reduces syncytia formation (2nd column). Electron micrograph of Syncytin-A wild type mouse placenta shows tight apposition of the syncytiotrophoblast I and II layers (ST-I; ST-II); stgc: sinusoidal trophoblast giant cells (1st column). Syncytin-A knock out mouse embryo interhemal domains shows unfused trophoblast I cells (T-I) (2nd column). Micrograph of the normal development of a sheep conceptus (1st column). Retarded growth of a sheep conceptus recovered after an envelope enJSRV morpholino antisense oligonucleotide injection (2nd column).

In addition, the endogenous retroviruses of sheep, enJSRVs, play a fundamental role in sheep conceptus growth and trophectoderm differentiation *via* their envelope glycoproteins. Indeed, *in vivo* experiments using an enJSRV envelope-specific morpholino injection trigger the loss of pregnancy by day 20 after injection (Dunlap et al. 2006). These kind of *in vivo* experiments obviously cannot be performed in human. Yet, primary cultures of human villous cytotrophoblasts cells give a unique opportunity to study placenta cells as closely related as possible to tissue environment. Thus, by using specific antisense oligonucleotides and siRNA strategies, expression of Syncytin-1 mRNA and protein as well as the syncytium formation by cell fusion events were dramatically reduced (Frendo et al. 2003; Vargas et al. 2009). In addition to that, Vargas and colleagues compared these results using the same targeting strategy against Syncytin-2, and interestingly

showed that Syncytin-2 inhibition in primary cells culture also leads to a decrease in fusion index that is more important than for Syncytin-1 (Vargas et al. 2009). They concluded that Syncytin-2 could also be a major determinant of trophoblast cell fusion, and in a coherent vision this underlines there should be more than one HERV envelopes proteins acting upon trophoblast cell fusion in human. Those parallel procedures demonstrating the involvement of human, mouse and sheep Syncytins in placenta development are illustrated in **Figure I-2 c**.

I.1.3. Domestication inside, the case of ERVWE1 and genemates

We now discuss the domestication processes through the better exemplified case of retroviral envelope gene, the ERVWE1/Syncytin-1 (**Fig. I-3 a**), at different regulation levels: insertion, transcription, maturation and function.

I.1.3.1. Sequence features

I.1.3.1.1. LTR and MaLR provide together a bipartite control element

Like any conventional retroviruses, endogenous retroviruses may display all the signals required for the transcription initiation and regulation within their LTRs. The U3 region of the ERVWE1 5'LTR possesses the promoter activity. The core promoter domain within the U3 region contains the CAAT box and the TATA box located upstream of the CAP site, marking the beginning of the R region (Prudhomme et al. 2004). Mutant analyses have confirmed the functional role of these boxes. Moreover, the 5' end of the U3 region harbors multiple binding sites contributing to overall promoter efficiency including GATA, Sp-1, AP-2, Oct-1, and PPAR-γ/RXR. Although Sp-1 and Ap-2 binding sites remain putative, they have been found to be essential for LTR activity (Prudhomme et al. 2004). It is noteworthy that Syncytin-1 regulation elements not only include the 5'LTR but also a so-called Upstream Regulatory Element (URE), a cellular 436 bp sequence located immediately upstream the Syncytin-1 proviral integration site, that define together with the 5'LTR a bipartite control element (Prudhomme et al. 2004) (**Fig. I-3 b & c**). This URE is composed of two main domains : (i) a distal regulatory region, including the previously mentioned putative binding sites found in the promoter core, as well as binding sites for the NF-κB and AP-1 important for the stimulation by TNFα, IFNγ, IL-1β, IL-6, and the inhibition by IFNβ (Mameli et al. 2007) (ii) a MaLR ancestor retrotransposon with binding sites for glucocorticoid and progesterone receptors, that features a trophoblast specific enhancer with putative sequences for ubiquitous Ap-2 and Sp-1, but also placenta specific GCMa binding sites (Prudhomme et al. 2004). Sequencing of a human panel showed that the 780 bp ERVWE1 5'LTR exhibits an unusually low polymorphism of one variable site every 18.0 kb as compared to a variability of one in 0.47 kb and one in 0.31 kb described for noncoding sequences and repeated sequences, respectively (Nickerson et al. 1998). This highlights a strong selection pressure in this region (Mallet et al. 2004). Moreover, comparative tests showed the conservation of the MaLR-LTR tandem from human to gibbon, and the juxtaposition of the MaLR of Hominidae with their related LTRs induces a drastic increase of the transcriptional activity in human trophoblastic cells (Prudhomme et al. 2004).

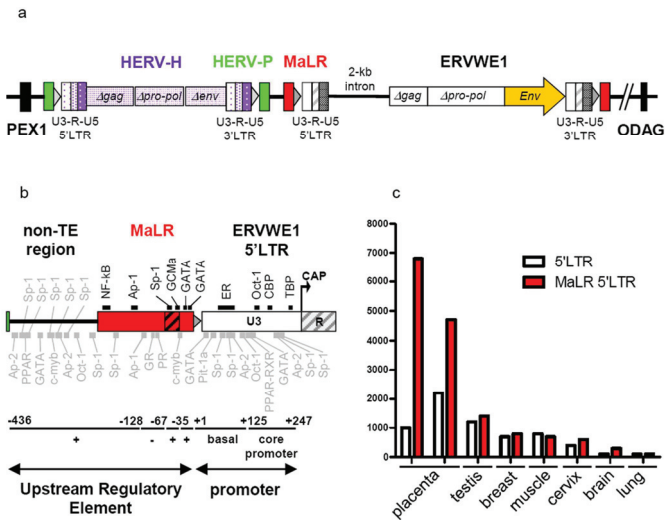


Figure I-3 Schematic representation of the retroviral-enriched PEX1-ODAG intergenic region and functional analysis of the bipartite element (MalR and ERVWE1 LTRs) which controls Syncytin-1 placental expression. (a) Flanking black boxes correspond to the 24th exon and the 5th exon of the PEX1 and ODAG genes, respectively. Host nonretroviral DNA is represented by a line. LTR elements are depicted as red boxes (MaLR LTR), green boxes (ERV-P LTR), purple tri-partite boxes (HERV-H provirus) and grey tri-partite boxes (ERVWE1 provirus). The U3, R, and U5 regions of HERV-H and ERVWE1 proviruses are labelled. The Env ORF is indicated by an orange arrow. (**b**) ERVWE1/Syncytin-1 transcriptional regulatory element: ERVWE1/Syncytin-1 expression is regulated by a bipartite element consisting of a cyclic AMP-inducible LTR retroviral promoter (ERVWE1 5'LTR U3 region) adjacent to an upstream regulatory element (URE) of composite origin. This URE consists of a 208 bp non-retroviral, non-repeated/transposable cellular sequence (non-TE region) and a 228pb MaLR LTR containing a trophoblast specific enhancer (TSE). True (top black boxes) and putative (bottom grey boxes) transcription factor binding sites along ERVWE1 5'LTR and URE are indicated. The positive (+) or negative (-) involvement of regulatory domains in placental tissue is annotated below the schematic representation. The CAP transcription initiation site (arrow) is located at the 5' end of the R region. (**c**) Trophoblast specific enhancer role of the MalR element. The ERVWE1 5' LTR (5' LTR, white bars) and the MalR - ERVWE1 5' LTR bipartite element (MalR 5'LTR) (red bars) were used to transfect 8 human cell types (BeWo, Jeg-3, N-Tera-2, HBL-100, TelCeB6, HeLa, U373, and LC5) corresponding to seven organs. Luciferase relative activities from at least three independent experiments are shown, illustrating that MalR element defined placenta tropism. Note that MalR cloned upstream from a heterologous SV40 promoter or in a reverse orientation far downstream from the ERVWE1 5' LTR increased both promoters efficiencies in BeWo cells what confirmed enhancer function (not illustrated).

I.1.3.1.2. Cross-species transcription regulation exemplified with GCM

Glial cell missing is a transcription factor family that has gradually attracted the attention of placenta researches. Originally isolated from a *Drosophila melanogaster* mutant line, two GCM homologues (GCMa and GCMb) have then been reported in mice, rats and humans (Keryer et al. 1998). GCMa is characterized by a zinc-coordinating DNA binding domain of β-sheets that recognizes an octomeric GCM binding motif 5'-ATGCGGGT-3' (Cohen et al. 2003). GCMa is primarily expressed in the placenta in humans and highly expressed in the labyrinthine trophoblast cells in mice

(Basyuk et al. 1999). Two binding sites by which GCMa can specifically transactivate Syncytin-1 have been described (Yu et al. 2002) and functional GCMa-binding sites were also identified in Syncytin-2 and MFSD2 promoters (Liang et al. 2010). Moreover, GCMa regulation has been linked to AMPc, protein kinase A signaling pathways (Chang et al. 2005; Chang et al. 2011; Knerr et al. 2005) and hypoxia levels (Klase et al. 2009). In agreement with these observations, the Syncytin-1 5'LTR core promoter is cAMP-inducible (Prudhomme et al. 2004). Interestingly, a microarray approach that aimed to identify GCMa target genes reported Syncytin-A to be downregulated in murine GCMa-deficient placenta (Schubert et al. 2008). siRNA GCMa inhibition in BeWo cells led to a decrease in syncytialization upon fusion events (Baczyk et al. 2009), and a reduced placental GCMa expression has been reported as a causative factor in defective syncytiotrophoblast differentiation in human preeclampsia (Bainbridge et al. 2012). More, GCMa have been identified as an upstream regulator of the connexin 43, a partner-protein engaged alongside with hASCT2 in cell-cell fusion (Dunk et al. 2012). Altogether, these data argue that GCM acts as a major regulator in the humans and mice Syncytins expression as well as in placenta maintenance and development.

I.1.3.1.3. Splicing strategies

ERVWE1 is a 10.2 kb-long locus located on chromosome 7q21.2 that can produce different spliced transcripts depending of the context (**Fig. I-4 a**). Historically, three single-spliced transcripts were detected in the placenta (Blond et al. 1999; Gimenez et al. 2010; Mi et al. 2000; Smallwood et al. 2003). The first mono-spliced transcript, 7.4 kb-long, contains the *gag*, the *pro/pol* pseudogenes and the *env* gene. The second one, 3.1 kb-long, strictly includes the open reading frame for the envelope protein Syncytin-1. Additionally, early northern blot experiments also detected a largely-spliced 1.3 kb transcript in the placenta (Blond et al. 1999). Alongside with the placenta, the testis exhibits different ERVWE1 mRNAs. The 7.4 kb form was seen early by northern blot in normal testis samples (Mi et al. 2000) and later by RT-PCR (Gimenez et al. 2010), and the 3.4 kb mRNA, which embarks the envelope-coding capacity, have been detected by RT-PCR in seminoma samples (Gimenez et al. 2010; Trejbalova et al. 2011). Note that the genomic full-length transcript of ERVWE1, from R (5'LTR) to R (3'LTR) has never been observed by northern blot although it was recently detected by RT-PCR in seminoma (Trejbalova et al. 2011). All these observations are in line with complex retroviruses transcription patterns such as MMTV or HTLV, for which several genomic and subgenomic transcripts derive from a single locus by alternative splice variations. Although the biological significance of non-coding splice forms of ERVWE1 in cancers is subject to discussions, the fact that the 3.1 kb *env*-coding RNA, in normal tissues, is constricted to the organ in which a physiological function exists but can re-appear in particular cancers argues that splicing variations may represent one additional level of control to the expression of domesticated retroviral sequences, balanced by other epigenetic mechanisms (Trejbalova et al. 2011).

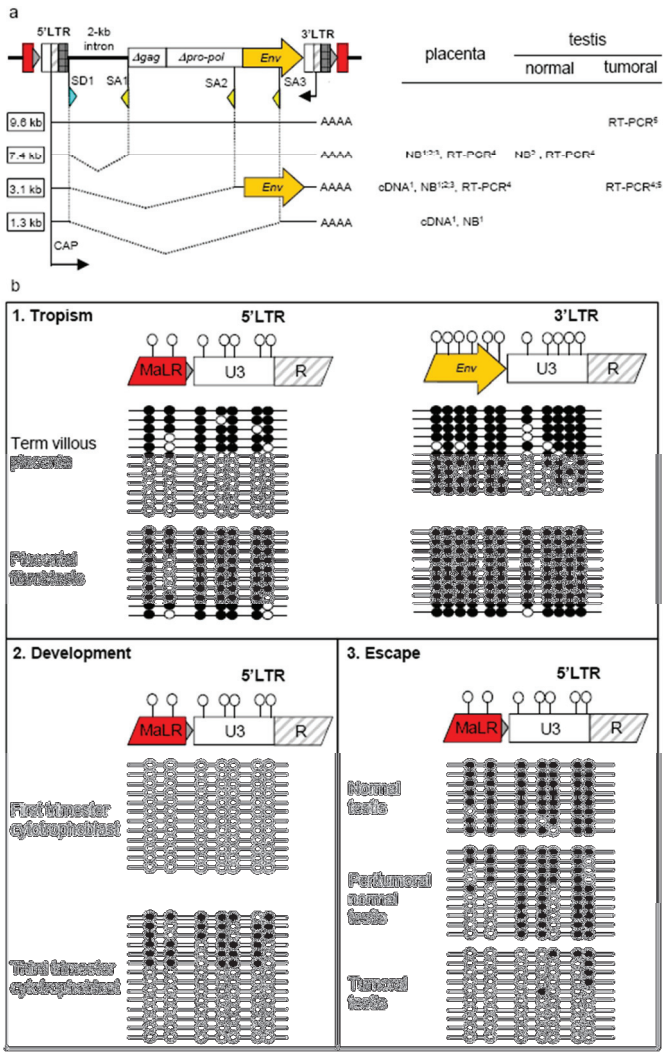


Figure I-4 Transcriptional and epigenetic control of Syncytin-1. (a) ERVWE1 splicing strategy in placenta and normal and tumoral testis. Left panel: the CAP transcription initiation site (right arrow) is located at the 5' end of the R region of the 5'LTR. The polyadenylation signal (left arrow) is located toward the 3' end of the R region belonging to the 3'LTR. ERVWE1 appears to produce four single-spliced transcripts, a genomic 9.6 kb, the subgenomic 7.4-kb and 3.1-kb mRNAs and the fully-spliced 1.3-kb mRNA. Only the 3.1-kb variant is responsible for Syncytin-1 translation. Splice donor (SD) and acceptor (SA) sites are indicated by blue right and yellow left arrows, respectively. SD and SA were identified by screening a placental cDNA library. Right panel: these four transcripts have been evidenced either by Northern blot (NB), RT-PCR or as almost complete cDNA clones in the tissues mentioned at the top of the table. References: [1] (Blond et al. 1999), [2] (Mi et al. 2000), [3] (Smallwood et al. 2003), [4] (Gimenez et al. 2010) and [5] (Trejbalova et al. 2011). (b) Comparative epigenetic control of 5' and 3' ERVWE1 LTRs in placenta (**1. Tropism**) and convergent modulation of bipartite element MalR-ERVWE1 during gestation (**2. Development**) versus potentially sequential in tumoral context (**3. Escape**). Promoter regions are indicated as boxes and CpG schematized by circles. MalR, LTR containing trophoblast specific enhancer, U3, ERVWE1 LTR promoter, R, transcription initiation site. CpG methylation is determined by bisulfite sequencing PCR in the indicated tissues. Each line represents an independent molecule. Methylated CpGs are schematized by black circles and unmethylated CpGs by white circles. (1. Tropism) Schematic representation of MaLR[LTR]–ERVWE1[5'LTR] and ERVWE1[env-3'LTR] analyzed regions. Methylation analysis was performed in villous trophoblast of term placenta and in placental fibroblasts from chorionic villi of a first trimester placenta. Each line represents an independent clone. Methylated CpG are schematized by black circles, unmethylated CpGs by white circles. (2. Development) Schematic CpG methylation dynamics of envelope-coding HERV 5'LTRs in cytotrophoblasts during pregnancy. Methylation MaLR[LTR]–ERVWE1[5'LTR], was in cytotrophoblasts (CT) at different times of gestation, i.e. CT of first trimester placenta from legally induced abortion and term placenta from healthy mother. Partial apparent remethylation may suggest an imprinting scenario (3. Escape) MalR LTR and ERVWE1 5'LTR global methylation comparison in normal, peritumoral and tumoral testis. Half of the molecules are hypomethylated in the MalR domain for the peritumoral tissue, suggesting a preferential route for hypomethylation.

I.1.3.1.4. Above the battlefield: epigenetics

I.1.3.1.4.1. Methylation of the LTRs

Methylation pattern studies of the ERVWE1 5'LTR revealed an inverse correlation between CpG methylation and locus expression indicating that demethylation of the 5'promoter is a prerequisite for the Syncytin-1 expression in trophoblasts cells (Matouskova et al. 2006). In an attempt to gain epigenetic characterization, Gimenez and colleagues (Gimenez et al. 2009) compared the methylation profiles of different HERV-W LTRs, including ERVWE1 5'LTR and 3'LTR, in villous placenta and in various non-trophoblastic cells (Fig. I-4 b). They showed that ERVWE1 5'LTR has the lowest methylation rate in villous placenta compared to others HERV-W LTRs, whereas all these LTRs including ERVWE1 5'LTR were broadly methylated in non-trophoblastic cells, a result reinforced by others (Macaulay et al. 2011). More, ERVWE1 5'LTR and 3'LTR, that both belong to the same locus, shared different methylation pattern since the 3'LTR remained highly methylated in villous placenta. Differential methylation between 5'LTR and 3'LTR is known for HTLV-1 (Koiwa et al. 2002) and HIV-1 (Ishida et al. 2006) during stages of viral latency but in a mirror scenario in which the 5'LTR is methylated as opposed to the 3'LTR demethylation. This suggests different methylation features whether we consider exogenous (pathogenic) or endogenous (domesticated) proviruses albeit with a common strategy that likely prevents the spread of methylation from one LTR to the other like the use of boundary elements as hypothesized for HTLV-1 (Koiwa et al. 2002). In the case of the ERVWE1 3'LTR, this could be crucial to keep active the Syncytin-1 promoter as well as to safeguard the use of the 3'LTR as a competitive or disrupting alternative promoter. Conversely, ERVWE1 5'LTR and MaLR behave quite similarly, e.g. in term villous placenta where all but one clone present a similar methylation profile whether we consider the MaLR or the 5'LTR. Thus, although belonging to distinct LTR types, these two elements could be linked and be involved jointly in the regulation of ERVWE1/Syncytin-1, what seems consistent with previous co-optation demonstrations (Bonnaud et al. 2005; Prudhomme et al. 2004) and the perspective of a bona fide gene (Mallet et al. 2004).

Changes of methylation patterns within ERVWE1 during pregnancy were also studied by a comparison involving first and third trimester samples (Gimenez et al. 2009). Methylation of the ERVWE1 5'LTR reaches 40% at term while completely absent at the beginning of the pregnancy. Thus, the selective and temporal unmethylation of the ERVWE1 locus in placenta during the first trimester may allow Syncytin-1-mediated cell differentiation and fusion, while, in contrast, increased methylation at term may limit Syncytin-1 production and consequent cell fusion or putative anti-apoptotic protection (Knerr et al. 2007) in accordance with cytotrophoblast limited fusion and higher apoptosis rate (Chen et al. 2011). Interestingly, ERVFRDE1/Syncytin-2 and ERV3 proviruses which are involved in fusion and immunomodulation, or proliferation, respectively (Andersson et al. 2005; Blaise et al. 2003; Kato et al. 1987; Mangeney et al. 2007) exhibit different and independent methylation patterns than the ERVWE1 locus in the placenta, what may reflect complementary and ordered physiological functions for these three provirus sequences (Gimenez et al. 2009).

A CpG hypomethylation status of the domesticated ERVWE1 5'LTR was reported in seminoma samples although at different extent, what may result from the small number of tumour samples or various degrees of differentiation (Gimenez et al. 2010; Trejbalova et al. 2011). The work on normal testis in addition to tumoral and peritumoral samples tends to show a switch from methylated to unmethylated DNA induced by a permissive escape of the MaLR (Fig. I-4 b / 3.Escape). This illustrates the need for the host to develop strong epigenetic defences in order to turn HERV sequences into domestic partners that can play physiological roles.

I.1.3.1.4.2. Changes in the histone code

Together with methylation levels, histones marks begun few years ago to change the appreciation of how chromatin is organized in normal development, cellular reprogramming and cancers (for an overview, see Baylin and Jones 2011). Recent works have investigated epigenetics hallmarks of the ERVWE1/Syncytin-1 and ERVFRDE1/Syncytin-2 loci surrounding their 5'LTR in BeWo and HeLa cells, and showed that the level of H3K9 trimethylation correlates perfectly with the CpG methylation of both proviruses (Trejbalova et al. 2011). The authors also associated the high density of H3K36 trimethylation along the intron-exon boundary of the Syncytin-1 envelope with high expression and efficient splicing form of the envelope gene. If these findings suggest at least partial redundancy within levels of control, histones modifications can also be seen as additive and multi-layers adaptations of the cell to guard against unfavorable effects of HERV elements. In line with this idea, we showed for instance that tissue specificity of the URE does not completely prevent weak and basal expression of ERVWE1 5'LTR in non-placental cell lines (Prudhomme et al. 2004).

Despite little is known about the general mechanisms of H3K9me3-dependant repression of ERVs sequences, different partner 'readers' proteins have been described to bind methylated lysines and to establish silent chromatin state in mouse, like isoforms HP1α, HP1β and HP1γ, but also the chromodomain proteins CDYL CDYL2, CBX2, CBX4, CBX7 and the M-phase phosphoprotein 8 protein (MPP8). In an attempt to clarify the role of these readers, Maksakova and coworkers recently performed experiments in mice embryonic stem cells and demonstrated that neither the depletion of HP1s nor the knockdown of the remaining known H3K9me3 readers lead to significant proviral reactivation (Maksakova et al. 2011). This suggests H3K9me3 might directly maintain ERVs in silent state in mice

embryonic stem cells, and consequently this invites to take an interest in what would happen in early stages of development.

I.1.3.2. Protein properties

I.1.3.2.1. Physiological cell-cell fusion requires crucial sequence adaptations

The fusogenic form of viral envelope glycoproteins is the outcome of a succession of maturation events. More precisely, class I fusion proteins are synthesized as glycosylated precursors in the lumen of the endoplasmic reticulum and are first modified by the cotranslational addition of N-glycans to the polypeptidic chain as well as by disulfide bond formation. After that a trimerization step involving a leucine zipper-like motif $LX_6LX_6NX_6LX_6L$ occurs in the ER before a proteolytic cleavage involving the cellular furin-like endoprotease gives rise to the two subunits SU and TM in the Golgi apparatus. Then a disulfide bond is established between SU and TM using the CΦΦC and the CX6CC motives, respectively. The final maturation step for γ–retroviruses envelopes, during viral budding, involves a 16-amino-acid carboxy-terminal peptide of TM, named R peptide, which is proteolytically cleaved by the viral protease what enables envelopes to ultimately trigger membrane fusion, as described by mutagenesis experiments (Yang and Compans 1996). Note that the cytoplasmic tails of most retrovirus envelope glycoproteins contain a YXXΦ (Φ is an amino acid with a bulky hydrophobic side chain [Leu, Ile, Phe, Val, or Met]) tyrosine-based sorting signal which plays a key role in subcellular distribution and adaptin-mediated endocytosis of plasma membrane-bound glycoproteins. Interestingly, the YXXΦ motif is located in the R peptide for MLVs and Mason-Pfizer monkey virus Env but is missing in Syncytin-1. These conserved motives are depicted in **Figure I-5 a**. Altogether, these maturation steps are essential for the acquisition of the envelope protein's fusogenic activity and therefore virion infectivity as illustrated in **Figure I-5 b**.

We illustrate how Syncytins used various strategies that diverge from envelopes of infectious retroviruses to adapt to their physiological functions in **Figure I-5 c**. Surprisingly, sequences comparison of the Syncytin-1 locus with all other HERV-W envelope elements revealed a 12-bp (corresponding to four LQMV amino acids) deletion in its cytoplasmic tail (Bonnaud et al. 2004) just downstream from the R-like ERVWE1 counterpart region. Moreover, insertion of these four amino acids into Syncytin-1 tail completely abolished the fusogenic potential (Bonnaud et al. 2004). This result argues that Syncytin-1 is constitutively fusion competent, as opposed to exogenous retroviruses envelopes, and is coherent with a domestication point of view since no viral protease open reading frames exist anymore in the human genome (Voisset et al. 2000).

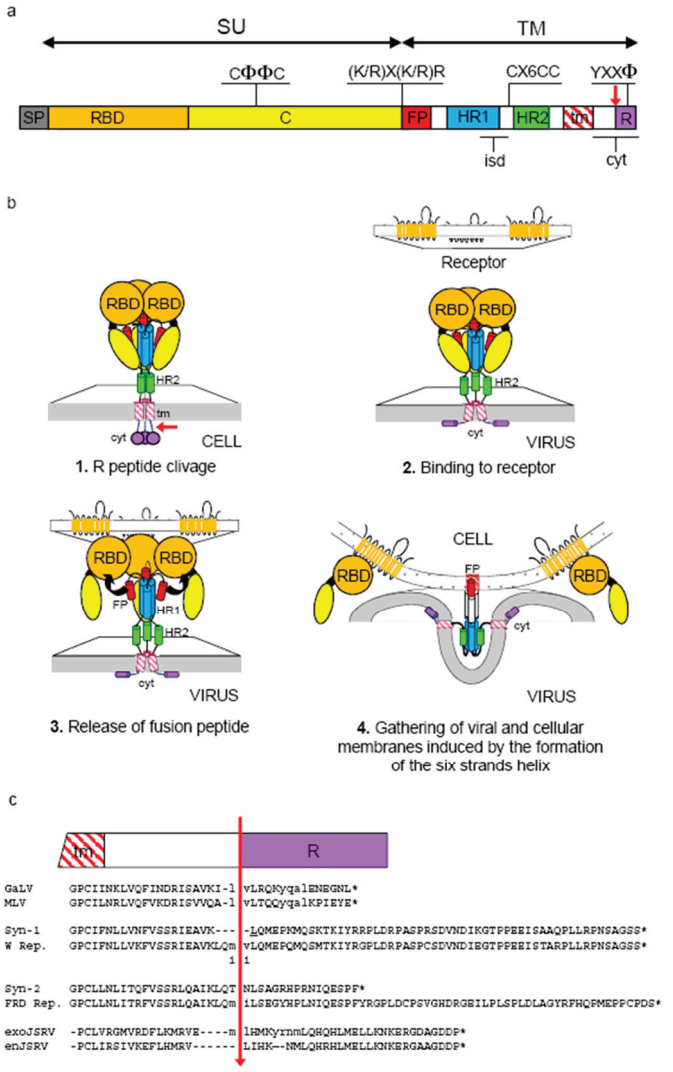


Figure I-5 Structure and maturation of retroviral envelope leading to virus-host cell membrane fusion and comparative evolution of Syncytins cytoplasmic tails. (a) Schematic portrait of an envelope prototype. SP, signal peptide (grey). SU, surface unit contains RBD, receptor binding domain (orange) and C, C-terminal domain (light) with CΦΦC motif (Φ=L,I,V,F,M or W), (K/R)X(K/R)R, furin cleavage site (red box). TM, transmembrane unit contains FP, fusion peptide (red); leucine zipper motif ($LX_6LX_6NX_6LX_6L$) with HR1 (blue) and HR2 (green) heptad repeats followed by the CX6CC motif; tm, trans-membrane anchorage domain (red, hatched); The ectodomain part of the TM contains a so-called immunosuppressive domain labelled isd, (QNRX2LDXLX5GXC); cyt, cytoplasmic tail with C-terminal R peptide (blue) containing YXXΦ motif. (**b**) Schematic representation of the maturation and conformational changes leading to virus-cell membrane fusion, beginning with the fusion competency acquisition of the envelope glycoprotein (1) based on R peptide release by viral protease and ending with the gathering of viral and cellular membranes (4) induced by the anchorage of the fusion peptide into the cell membrane. Red arrow symbolizes the R peptide cleavage. (**c**) The first five amino acids correspond to the transmembrane domain. Experimentally determined (GaLV, MLV, exoJSRV) and putative (W Rep. and FRD Rep.) protease cleavage site (black line) and YXXΦ signaling motif are indicated in lowercase. Comparison of the Syncytin-1 protein (Syn-1) with the HERV-W family consensus sequence obtained from Repbase (W Rep.) shows a four amino acids deletion (LQMV) in the domesticated fusogenic protein, overlapping the ancestral viral protease cleavage site. The underlined leucine indicates a C-terminal truncation mutant exhibiting hyperfusogenic activity and significant pseudotyping capacity. Comparison of the Syncytin-2 protein (Syn-2) with the Repbase FRD consensus sequence (FRD Rep.) shows a stop codon that shortens the Syncytin-2 cytoplasmic tail and no evidence of viral protease cleavage site. Alignment of enJSRV and exoJSRV shows the placenta-expressed enJSRV has accumulated mutations surrounding the protease cleavage site and lacks downstream tyrosine (Y) residue. Genebank accession numbers: MLV: M14702; GaLV: AF055060, Syncytin-1: GQ919057, Syncytin-2: HEU27240, enJSRV: enJS56A1 and exoJSRV: AF105220.

Furthermore, the role of the cytoplasmic domain of Syncytin-1 has been systematically investigated by producing a series of C-terminal truncated variants, leading to the conclusion that residues adjacent to the membrane domain are required for optimal fusion probably by forming a helical structure, while final C-terminal residues more likely act as a

fusion inhibitor domain (Cheynet et al. 2005; Drewlo et al. 2006). Remarkably, a truncation mutant which shortens the cytoplasmic tail precisely at the site of the LQMV-deletion motif exhibits higher fusogenic properties than the wild-type protein (Cheynet et al. 2005). Even if no work on Syncytin-2 has been done in an exhaustive way to assess the fusogenic properties modulation of its cytoplasmic tail, we identified a stop codon in the cyt of Syncytin-2, as opposed to the RepBase prototype, resulting in a shortening of the tail. More, the protease cleavage site appears absent as regard to the FRD family consensus genome. Studies on the cytoplasmic tail of JSRV envelope protein first focused on the VR3 region that was described as the least conserved region between exogenous and endogenous forms. The VR3 region includes the putative membrane-spanning domain as well as the cytoplasmic tail, and series of envelope chimeras revealed that mutations in a YXXM motif of the cytoplasmic tail of JSRV *env* were sufficient to inhibit or modulate its transforming abilities (Hull and Fan 2006; Palmarini et al. 2001). Further mutational amino acid substitutions have proven the tyrosine residue to be essential for transformation of exogenous JSRV. We observed that the VR3 region of all exogenous stains of JSRV sequences exhibit this tyrosine residue whereas all the enJSRVs envelopes described so far lacked this motif critical for JSRV transformation (**Fig. I-5 c**).

I.1.3.2.2. Immunomodulation, that makes the switch

Given that the placenta is an extra-embryonic tissue, half paternal and half maternal genetically inherited, the past decades have gathered reproductive immunologists researches to solve the fetal semi-allograft problem. Regulatory T cells are responsible for the establishment of tolerance by modulating the immune response, and uterine natural killer cells direct placentation by controlling trophoblast invasion (Munoz-Suano et al. 2011). As an example the contact zone between mother uterus and fetus extravillous cells of spiral arterioles appears to be one of these predictive immunological conflict zones, where Syncytin-1 has also been shown to be expressed (Malassine et al. 2005). So, beside extravillous cytotrophoblast-expressed HLA-G immunoregulatory proteins, immunomodulation properties of retroviruses sequences (Mullins and Linnebacher 2011; Rolland et al. 2006; Wang-Johanning et al. 2008) may contribute to answer the tolerance during pregnancy. Thus, a first mechanism of Syncytins-mediated immunosuppressive activity may be due to the presence of a putative immunosuppressive region conserved among murine, feline, and human retroviruses (Cianciolo et al. 1985), depicted hereafter for MPMV, SRV, SNV and BAEV so-called immunosuppressive retroviruses: LQNRRGLDLLTAE**Q**GGICL**A**. The analysis of this domain for the human and mouse Syncytins in a mouse model of transplant rejection has revealed an immunosuppressive activity for Syncytins-2 and -B but not for the Syncytins-1 and –A (Mangeney et al. 2007). More precisely, two amino acids have been described as commutator points that can be alternatively turned 'on' or 'off' in substitution experiments and trigger a switch from immunosuppressive to non-immunosuppressive activity (see bold letters in sequence above). This suggests a possible co-operation in tandem of the Syncytins pairs in Primates and Muridae, with complementary fusion and immunosuppression functions adapted to cellular and physiological contexts. Additional recent findings also suggest that envelopes coming from the HERV-K family may contribute to placentogenesis or provide immune protection to the fetus (Kammerer et al. 2011), what reinforces the idea of complementary functions within HERV envelope partners in the placenta. A second potential mechanism links immune response and amino acid balance. During pregnancy, maternal tryptophan is required for the T lymphocytes activation and 'immunosuppression by starvation' is the consequence of tryptophan depletion experiments (Mellor and Munn 1999). Besides, a tryptophan-catabolizing enzyme, the indoleamine 2,3-dioxygenase, is particularly expressed in the syncytiotrophoblast. Thus, the lymphocyte regulation appears to be strongly mediated by the ability of the apical membrane to incorporate the tryptophan into the syncytiotrophoblast (Kudo et al. 2001). In other words, the tolerance towards the allograft is locally conditioned by the CD98/LAT1 tryptophan transporter and the resulting amino acid balance changes. If we consider that the Syncytin-1 interacts with amino-acid transporters from one side (Blond et al. 2000; Lavillette et al. 2002) and with the TLR4 and the pathogen-recognition receptor DC-SIGN *in vitro* from the other side (Cheynet et al. 2005; Rolland et al. 2006), the modulation of the immune system *via* amino-acid balances and TLR4 stimulations would become an axis of understanding the tolerance, articulated around the Syncytins. Even if the Syncytin-1 and Ory-1 ASCT2 receptor only mediates the transport of small amino acids, and consequently probably not tryptophan, considerations about balance changes that could impact the immune system response are maybe not so far. On one hand, infection of cells with Syncytin-1 phylogenetically-related RD114/simian immunosuppressive type D retroviruses results in impaired amino acid transport, a mechanism proposed to mediate virus immunosuppression (Rasko et al. 1999). On the other hand the glutamine, a small amino-acid accepted by both ASCT1 and ASCT2 transporters, was shown to influence the balance within the T lymphocytes sub-populations, potentially influencing the host response (Chang et al. 1999). Interestingly, recent findings suggest that Syncytin-1 is shed from the placenta into the maternal circulation in association with microvesicles, and modulates immune cell activation (Holder et al. 2012). Surprisingly, similar effect was demonstrated with a recombinant protein encompassing amino acids 116-225, i.e. excluding the TM immunosuppressive domain but conserving part of the SU subunit which includes most of the $SDGGGX_2DX_2R$-conserved motif previously seen to be directly involved in Syncytin-

1/hASCT2 receptor recognition (Cheynet et al. 2006).

I.1.3.3. A price to pay

'That said, the wolf ran off, and he is running still'

As illustrated above, the multiple levels of control exemplified with the Syncytin-1 may suggest that Syncytins expression is tightly regulated to be constrained to the placenta, where physiological functions take place. However, diseases of the placenta have been linked with Syncytins derugulations, and various pathological contexts have reported Syncytin-1 expression. We give here an overview of the price to pay to the domestication of such retroviral elements.

I.1.3.3.1. Syncytins and diseases of the placenta

Pre-eclampsia and HELLP syndrome are disorders associated with abnormal placentation, including defects in syncytiotrophoblast formation. Numerous studies have associated PE and HELLP with Syncytin-1 and Syncytin-2 significant reduction (Chen et al. 2008; Chen et al. 2006; Knerr et al. 2002; Lee et al. 2001; Strick et al. 2007). Syncytin-2 expression was more importantly impaired than Syncytin-1 in severe pre-eclampsia (Vargas et al. 2011). Interestingly, a redistribution of Syncytin-1 within the syncytiotrophoblast polarized cell layer was observed for patients with PE (Lee et al. 2001). Though, cultured cytotrophoblast cells from PE and HELLP showed higher apoptotic rates (Strick et al. 2007). A reduced Syncytin-1 expression has also been reported in placenta from intra uterine growth restriction and was associated with an overall disorganized syncytiotrophoblast layer with fewer nuclei (Ruebner et al. 2010).

I.1.3.3.2. Expression of Syncytin-1 in autoimmune diseases and cancers

Syncytin-1 is expressed in astrocytes, glial cells and activated macrophages in brain regions affected by multiple sclerosis. Syncytin-1 expression in astrocytes mediates neuroimmune activation and death of oligodendrocytes by inducing the release of cytotoxic redox reactants (Antony et al. 2004). In astrocytes, Syncytin-1 induces the expression of OASIS, an endoplasmic reticulum stress sensor, which in turn increases the expression of inducible NO synthetase and concurrent suppression of cognate hASCT1 receptor, resulting in a diminished myelin protein production (Antony et al. 2007). What mechanisms reactivate Syncytin-1 in the brain in MS is still not clear. This could be the result of viral infection of the brain, such as herpes simplex virus, which has previously been shown to transactivate Syncytin-1 expression (Nellaker et al. 2006), or cytokine deregulation (Perron et al. 2001). Indeed it has been shown in astrocyte cultures that MS detrimental cytokines, IFN-γ and TNF-α are able to induce Syncytin-1 expression through NF-κB activation, while MS protective IFN-β inhibits its expression (Mameli et al. 2007). In addition, Syncytin-1 induction by exogenous TNF-α into the corpus callosum, a region of the brain frequently exhibiting demyelination in MS, leads to neuroinflammation, reduction of myelin proteins level and neurobehavioural deficits in Syncytin-1-transgenic mice, as observed in MS (Antony et al. 2007). Interestingly as a parallel between MS and cancers, NO production in tumor vessels correlates with an increase of the over-all survival as well as the decrease of metastatic potency in experimental systems (Mortensen et al. 2004). On line with this, the level of Syncytin-1 expression represented a positive prognostic indicator for recurrence-free survival of breast cancer patients (Larsson et al. 2007). Conversely, increased Syncytin-1 expression was associated with decreased overall survival in rectal but not in colonic cancer patients (Larsen et al. 2009). The situation appears unclear in endometrial carcinoma where the increase of Syncytin-1 expression in normal endometrium of patients may possibly influence the development of endometriosis (Oppelt et al. 2009). Thus, the prognostic impact of Syncytin-1 expression appears to vary with the tumor type potentially, due to different functions associated with different pathways of reactivation. In a more general way, Syncytin-1 expression was observed for about one-third of breast cancer patients, and additionally, neighbouring endothelial cells were shown to express hASCT2 receptor (Bjerregaard et al. 2006). *In vitro* studies confirmed the involvement of Syncytin-1 in the fusion process between breast cancer cell lines and endothelial cells (Bjerregaard et al. 2006). Syncytin-1 was also found to be expressed in leukemia and lymphoma cells while no expression was identified in blood samples of normal individuals (Sun et al. 2010). Syncytin-1 associated cell-cell fusion was identified in EnCa tumors *in vivo*, but interestingly, *in vitro* studies showed the implication of Syncytin-1 in both the fusion and the proliferation of EnCa cells (Strick et al. 2007). Syncytin-1 up-regulation *via* the cAMP pathway leads to cell-cell fusion while induction by steroid hormones (estradiol) leads to proliferation. This molecular switch is apparently controlled by TGF-β1 and TGF-β3 which are induced by steroid hormones and may override Syncytin-1 mediated cell-cell fusions (Strick et al. 2007).

I.1.4. Conclusion

Our life starts with the sperm-egg fusion, yet this princeps phenomenon remains partly to be elucidated (Kawano et al. 2011). As the embryo develops, skeletal muscle differentiation depends on the fusion of mononucleated myoblasts to form multinucleated muscle fibers. Recent i*n vitro* findings showed that Syncytin-1 and its receptors hASCT1 and hASCT2 are expressed in human myoblasts and involved in myoblast fusion (Bjerregaard et al.

2011). In the adult body, macrophages can fuse to form either multinucleated osteoclasts that control the maintenance of the bones or multinucleated giant cells that are important for the immune response. We begin to know that Syncytin-1 interacts with hASCT2 in differentiating osteoclasts and is expressed in human iliac crest biopsies (Søe et al. 2011). All these findings continue to feed both mechanistic and biological knowledge on gene domestication. More generally, genetic exchanges and their impacts on living structures remain today the crucial evolutionary force it was in ancient time, and in our point of view, retrovirology may significantly support comparative genomics. Definitely, endogenous retroviruses and more broadly retrotransposons represent an impressive mass of insufficiently solicited witnesses of such forces in action.

Acknowledgements

We thank Danièle Evain-Brion, Thierry Heidmann, Thomas E. Spencer, and François-Loïc Cosset for providing pictures and photographs. We are grateful to Laurent Duret for his support in bioinformatics, and we want to pay a tribute to Jean de La Fontaine for the contribution that his fable on the domestication *The Wolf and the Dog* brought to our scientific reflection.

Funding

Advanced Diagnostics for New Therapeutic Approaches (ADNA), a program dedicated to personalized Medicine, coordinated by Mérieux Alliance and supported by the French public agency, OSEO. PP and FM are employees of bioMérieux SA. PAB was supported by a grant from the Ministère français du Travail, de l'Emploi et de la Santé.

I.2. Expression et régulation de l'expression des rétrovirus endogènes humains (hors placenta)

Le succès évolutif de l'intégration d'un provirus HERV-W il y a 25 millions d'années dans le génome des hominoïdes et, plus largement, de la domestication, par différentes espèces, de séquences d'enveloppes rétrovirales dont les fonctions acquises indépendamment convergent pourtant vers une contribution au développement placentaire, n'est qu'un aspect du devenir de la masse des séquences rétrovirales endogènes. Le séquençage du génome humain en 2001 a évalué à un peu plus de 200 000 le nombre de copies HERV essaimées sur les différents chromosomes (International Human Genome Sequencing Consortium 2001). Si la caractérisation de l'ensemble de ces séquences reste fragmentaire, une quarantaine de familles multicopies, correspondant à autant d'infections uniques passées, sont aujourd'hui définies et étudiées (Mager and Medstrand 2003) (Tableau I-1). Ce réservoir génétique de vestiges rétroviraux, à remettre en perspective avec les quelques 25 000 gènes codants chez l'homme, pose la question de l'implication des HERV dans les processus biologiques.

Peu d'observations associent l'expression des rétrovirus endogènes humains à des contextes physiologiques en dehors du placenta, et ces descriptions s'arrêtent généralement au niveau transcriptionnel. Toutefois, l'abondance de séquences LTR, qui sont les promoteurs naturels des rétrovirus, au voisinage des gènes cellulaires, participe dans de nombreux cas à la régulation de l'expression génétique. La profusion des descriptions d'activité des HERV en contexte pathologique témoigne pourtant bien du potentiel latent de notre patrimoine rétroviral, qui se révèle lorsque les niveaux de contrôle mis en place par l'hôte sont levés, par exemple sous la contrainte de stimuli extérieurs. Ainsi, les cancers, les maladies auto-immunes et certaines affections du système nerveux sont les principaux contextes de réexpression des HERV, bien que la participation des rétrovirus endogènes à l'étiologie de ces pathologies reste incertaine. En impliquant des phénomènes de transcription, mais aussi de synthèse protéique et, dans certains cas, la production de particules virales, les situations pathologiques peuvent constituer des modèles d'études pour renforcer la compréhension de la biologie des rétrovirus endogènes. L'identification d'éléments HERV particuliers, en association avec un état de différenciation cellulaire peut, quant à elle, constituer une source de marqueurs de pathologies. Ces différents niveaux de connaissances sont l'objet de ce développement bibliographique.

Nom usuel	Nom dans Repbase		Autres noms	PBS	Nombre de copies
	Région interne	LTR			
Classe I					
HERV-I	HERVI	LTR10	RTVL-I, HERV-FTD	I	250 (1000)
HERV-ADP	HERVP71A_I	LTR71A		P	40 (300)
HERV-HS49C23	MER57I	MER57		L	200 (1000)
MER110	MER110I	MER110,110A		?	20 (60)
HERV-Rb	PABL_BI	PABL_A, PABL_B		R/L	8 (1000)
HERV-E	HERVE	LTR2,2B,2C	4-1	E	250 (1000)
HERV-R	HERV3	LTR4	ERV-3	R/L	100 (125)
RRHERV-I	HERV15I	LTR15		I	40 (250)
S71	HERVS71I	LTR6A_B	HERV-T	T	80 (400)
PRIMA41	PRIMA41I	MER41A,41B,41C,41D		?	40 (3000)
HERV-S71-7	MER66I	MER66A,66B		R	50 (1500)
HERV-P72D	MER50I	MER50		?	50 (2000)
MER84	MER84I	MER84		I	25 (300)
HERV-W	HERVI7	LTR17	MSRV	W/R	40 (1000)
HERV30	HERV30I	LTR30		R/L	35 (120)
ERV9	HERV9	LTR12,PTR5		R	300 (5000)
MER52A	MER52AI	MER52A,52B,52C		P	200 (1000)
HERV-P	HUERS-P3	LTR9	HuERS-P3	R/G	200 (1000)
HERV-XA34	HERVFH21I	LTR21A	HERV-F(XA-34)	F	30 (60)
HERV-H	HERVH	LTR7	RTVL-H, RGH	H	1000 (1000)
HERV46	HERV46I	LTR46		F	40 (200)
HERV-Fb	HERVH48I	MER48		F	60 (100)
HERV-F	HERVFH19I	LTR19		F	45 (550)
Classe II					
HML-1	HERVK14I	LTR14A,14B	NMWV6	K	70 (350)
HERV-K(HML-2)	HERVK	LTR5	HERV-K,HERV-K10, HTDV,NMWV1	K	60 (2500)
HML-3	HERVK9I	MER9	HERV-K70A,HERV76, HERV50,NMWV5	K/Y	150 (700)
HML-4	HERVK13I	LTR13	ERV-3/L3,HERV-K(T47D)	K	10 (300)
HML-5	HERVK22I	LTR22,22A,22B	NMWV2	K/M	100 (450)
HERV-K(HML-6)	HERVK3I	LTR3,3B	NMWV4	K	50 (400)
HML-7	HERVK11D1	MER11D	NMWV7	K	20 (140)
HML-8	HERVK11I	MER11A,11B,11C	NMWV3	?	60 (600)
HML-9	-	-	NMWV9	W	10 (40)
HERV-K(C4)	HERVK.C4,	LTR14	HML-10	K	10 (100)
HERVK14C	HERVK14CI	LTR14C		K	15 (120)
Classe III					
HERV-L	HERVL,ERVL	MLT2A1,2A2,MLT2B1,2B2		L	200 (6000)
HERV16	HERV16	LTR16,16A,16A1,16C,16D		L	15 (25)
HERV-S	HERV18	LTR18,18B		S	50 (150)
HERV37	HERV37	LTR37		S	30 (200)
MER70	MER70I	MER70A,70B		S	40 (200)
HERV32	HERV32	LTR32		S	10 (50)
HERVL74	HERVL74	MER74,74A,74B,74C		S	25 (200)
HERVL40	HERVL40	LTR40A,40B,40C		S	10 (300)

Tableau I-1 Nomenclature et nombre de copies HERV du génome. L'estimation du nombre de copies fait une distinction entre les structures provirales internes et, entre parenthèses, les LTRs. D'après (Mager and Medstrand 2003).

I.2.1. Contextes physiologiques

I.2.1.1. Activité des HERV

I.2.1.1.1. Cellules de la lignée germinale et tissus de la reproduction

Les cellules sexuelles représentent, avec le placenta, des lieux privilégiés d'expression des HERV. Le testicule, à titre d'exemple, a été le seul contexte d'activité de la famille HERV-W à avoir été identifié par Northern Blot en dehors du placenta (Mi et al. 2000). En utilisant des méthodes de détection plus sensibles, des transcrits appartenant à de nombreuses familles HERV ont été détectés dans le testicule (Ahn and Kim 2009; Buzdin et al. 2006a; Crowell and Kiessling 2007; de Parseval et al. 2003; Flockerzi et al. 2008; Larsson et al. 1994; Larsson et al. 1997; Muradrasoli et al. 2006; Seifarth et al. 2005). Parmi ces expressions, certaines sont d'ailleurs communes avec le placenta, comme les familles HERV-W et HERV-R, et d'autres apparaissent davantage spécifiques de l'organe. En particulier, l'expression d'un provirus de la famille HERV-K situé sur le chromosome 16, et nommé HERV-K-MEL, n'est observée de manière significative en condition physiologique que dans le testicule (Schiavetti et al. 2002) et certains éléments, parmi lesquels l'enveloppe de HERV-F(c)1 ou de HERV-H3, semblent avoir un tropisme fortement réduit à cet organe (de Parseval et al. 2003). Quoi qu'avec une variété de transcrits observés plus réduite, l'ovaire est également un lieu d'activité important des HERV. Les familles les plus exprimées, détectées au niveau de leur gène d'enveloppe (de Parseval et al. 2003) ou de leur séquence *pol* (Seifarth et al. 2005), sont HERV-R, HERV-W, HERV-FRD, HERV-H, ERV-9 et HML-6. Des expressions associées à la famille HERV-E (Hu et al. 2006) et au provirus HML-2 K10 (Georgiou et al. 2009) ont aussi été décrites dans l'endomètre et l'oocyte, respectivement. Notons enfin que des protéines d'enveloppe de la famille HERV-3 ont été mises en évidence à la surface des oocytes (Nilsson et al. 1999).

I.2.1.1.2. Autres tissus adultes

L'expression des HERV est beaucoup plus réduite dans les tissus adultes non impliqués dans la fonction de reproduction. Il a été montré, par exemple, que la famille ERV-9 est exprimée à des niveaux très faibles dans presque tous les tissus, quoi qu'à un niveau un peu plus fort dans les glandes surrénales (Svensson et al. 2001). Plusieurs études indiquent que les glandes surrénales sont également un lieu d'expression prononcé pour la famille HERV-R {Larsson, 1997; Katsumata, 1998; Andersson, 1998; Andersson, 2002; de Parseval, 2003; Andersson, 2005}, tout comme les glandes sébacées (Andersson et al. 1996; Andersson et al. 2005; Larsson et al. 1997), ce qui peut suggérer un mode de régulation par les hormones stéroïdiennes. Des transcrits HERV-E et HERV-T ont été décrits dans la thyroïde (de Parseval et al. 2003; Shiroma et al. 2001) et le pancréas (Shiroma et al.

2001). Les familles HERV-FRD, HML-4 et HML-6 semblent s'exprimer plus fortement dans le foie que dans les autres tissus (Seifarth et al. 2005) ; dans le thymus, il s'agirait plutôt de la famille HERV-P (Ahn and Kim 2009). L'expression de plusieurs groupes de séquences HERV a été rapportée dans le rein, incluant les familles HERV-W (Forsman et al. 2005) et HML-2 (Ahn and Kim 2009), avant qu'un profil transcriptomique HERV complet du rein ne soit finalement établi (Haupt et al. 2011). Sur la famille HERV-W, notons que l'expression de la Syncytine-1 au niveau transcriptionnel et protéique a récemment été associée à la fusion des ostéoclastes (Søe et al. 2011). Enfin, dans le cerveau, des approches expérimentales (Flockerzi et al. 2008) et l'exploitation de banques d'ESTs (Stauffer et al. 2004) concluent à l'expression de certains éléments HML-2, en particulier du locus K108, bien que des transcrits reliés aux familles HERV-I et HERV-H ait également été détectés (Forsman et al. 2005). Il convient de souligner que la part de variation inter individuelle, dans ces études, ne fait jamais l'objet d'évaluation.

I.2.1.2. Les LTRs, régulateurs de l'expression génique

A l'instar des séquences provirales *gag*, *pol* et *env*, l'expression des LTRs en contexte physiologique se présente comme une anomalie dans le cadre théorique de la répression des éléments transposable par l'hôte. S'il est plus délicat de mettre au point des systèmes de détection qui les ciblent spécifiquement, il n'en demeure pas moins que l'insertion et l'activité des LTRs constituent une source d'expression et de régulation d'expression, en conférant, par exemple, des tropismes particuliers aux gènes cellulaires environnants. Plusieurs types de contributions des LTRs à l'activité génique ont été proposés depuis une vingtaine d'années, d'abord par l'étude au cas par cas de transcrits chimériques sur la base de l'exploration de banques d'ADNc (Feuchter et al. 1992), puis, avec le développement des techniques de séquençage et de bioinformatique, en appliquant des méthodes algorithmiques de recherche de transcrits de fusion (Buzdin et al. 2006b; Conley et al. 2008; Conley and Jordan 2010; Huda et al. 2011). La littérature décrivant les interactions supposées entre LTRs et gènes et aujourd'hui très fournie, mais les niveaux de démonstrations expérimentales varient énormément d'un cas à l'autre. Aussi, nous proposons ici de focaliser sur quelques illustrations bien documentées de LTRs tenant lieu soit de promoteur alternatif, soit de modulateur de la transcription (enhancer), soit de signal de polyadénylation (Figure I-6).

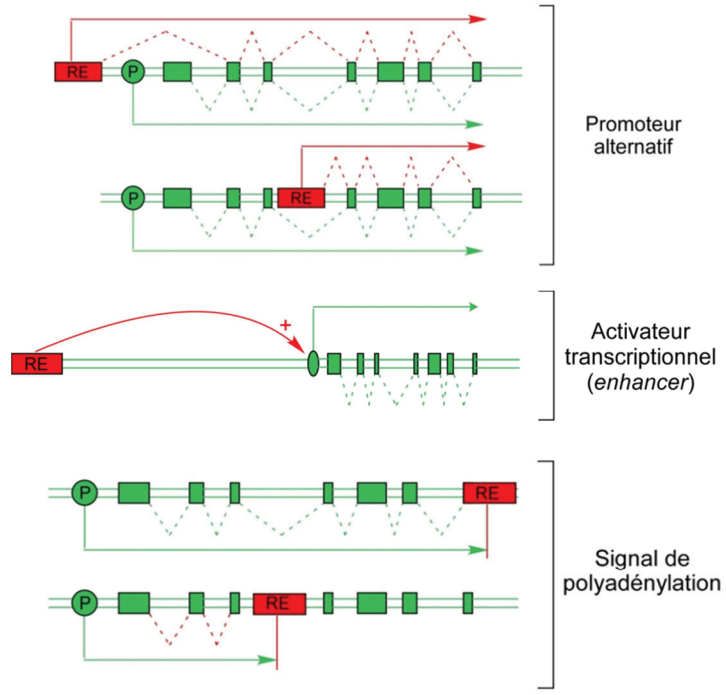


Figure I-6 Cas théoriques de contrôle de l'expression des gènes cellulaires par les LTRs. RE : repeat element, une LTR ici. P : promoteur naturel du gène. Adapté de (Buzdin 2004).

I.2.1.2.1. LTRs promotrices

I.2.1.2.1.1. Acquisition d'un nouveau tropisme d'expression

L'impact le plus significatif des LTRs sur la biologie de l'hôte est certainement l'acquisition d'un tropisme d'expression nouveau par l'intégration d'un élément promoteur rétroviral en 5' d'un gène cellulaire. L'étude des gènes orthologues pour lesquels une telle insertion est absente permet d'identifier ces cas de figures, et des modèles évolutifs basiques ont été proposés à l'acquisition de sites de fixation de facteurs de transcription tissu-spécifiques (Figure I-7).

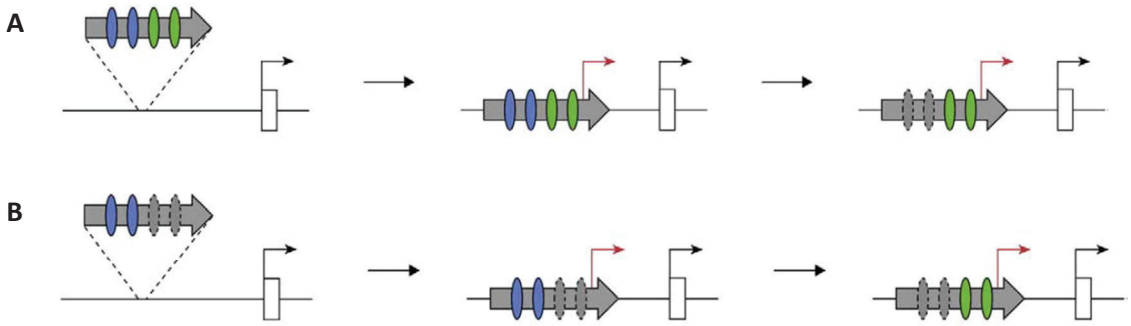


Figure I-7 Scénarios évolutifs d'acquisition d'un nouveau tropisme d'expression par l'intégration d'une LTR. La LTR qui s'intègre en amont du gène possède des sites de fixation de facteurs de transcription génériques (bleu) ou tissu-spécifiques (vert). L'activité du gène cellulaire dans un nouveau contexte tissulaire découle (A) du maintien ou (B) de l'acquisition de sites de fixation de facteurs de transcription tissu-spécifiques. D'après (Cohen et al. 2009).

La majorité des rares exemples d'une expression tissu-spécifique induite par l'insertion de séquences LTRs concerne des gènes à tropisme placentaire. Les cas de PTN, CYPA19A1, NOS3, et IL2RB ont été discutés dans la revue introductive (partie I.1.2.3.3.1). En dehors du placenta, le gène B3GAL-T5, impliqué dans la synthèse des chaines carbohydrates dans le pancréas et le conduit gastro-intestinal, est un exemple de ce que l'équipe de Dixie Mager propose d'appeler un phénomène d'exaptation[2] de LTR. Il a en effet été montré qu'une LTR solitaire de la famille HERV-L est l'un des promoteurs alternatifs de B3GAL-T5 et que, si cette insertion confère une activité dans des tissus variés, elle assure un rôle de promoteur dominant dans le côlon (Dunn et al. 2003). De plus, l'activité dans le côlon est largement dépendante du facteur de transcription HNF-1, qui trouve des sites de fixation sur la LTR et sur le promoteur naturel du gène ; cependant, les sites du promoteur naturel sont en orientation antisens, ce qui a pu contribuer au succès évolutif de l'intégration de la LTR HERV-L. Il est intéressant de constater que l'intégration d'une LTR peut, à l'inverse d'instaurer un nouveau tropisme restreint, étendre celui-ci à différents tissus. La présence d'une LTR ERV-9 en amont du gène HEP27 (Heinz et al. 2002) ou encore d'une LTR HERV-H en 5' du gène GSDMB (Sin et al. 2006a) en sont deux illustrations. Plus particulièrement dans le cas du gène de la gasdermin B (GSDMB), des transcrits chimériques ont été caractérisés dans des tissus variés alors que l'expression du gène par son promoteur naturel n'a été associée qu'à l'estomac et certaines lignées cellulaires. Une particularité de cette LTR HERV-H est d'être intégrée en orientation antisens par rapport au gène GSDMB. Des tests d'activité menés sur des systèmes *in vitro* ont montré que l'activité promotrice de la LTR HERV-H diminue lorsqu'elle est clonée en orientation sens par rapport au gène (Sin et al. 2006a) et des expériences de mutations ont caractérisé les régions de la LTR essentielles à son activité (Huh et al. 2008).

I.2.1.2.1.2. Substitution totale ou partielle du promoteur naturel

Un autre schéma d'association de séquences LTRs avec des gènes cellulaires consiste au maintien du tropisme d'expression génique après l'intégration du promoteur viral. Dans ce cas, deux scénarios évolutifs peuvent être rencontrés (Figure I-8) : ou bien la LTR finit par prendre exclusivement le relais de l'activité promotrice du gène, ou bien deux (parfois plusieurs) promoteurs coexistent. Cette seconde situation peut d'ailleurs très bien être pensée comme une étape transitoire vers la première.

[2] En référence au concept évolutionniste formulé par Stephen J. Gould qui postule que des fonctions peuvent être exercées à un temps *t* (par un organe, par exemple), indépendamment de ce qu'elles furent au moment de leur sélection positive naturelle.

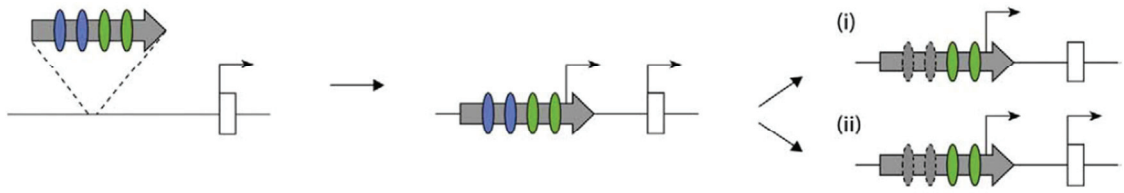


Figure I-8 Scénarios évolutifs de substitution ou de renfort d'un promoteur naturel. La LTR qui s'intègre en amont du gène possède des sites de fixation de facteurs de transcription génériques (bleu) et tissu-spécifiques (vert) et contribue à une l'activité promotrice alternative. L'évolution peut soit (i) aboutir à la perte du promoteur natif soit (ii) maintenir l'activité de deux promoteurs tissu-spécifiques. D'après (Cohen et al. 2009).

Deux situations de substitutions complètes du promoteur naturel peuvent être citées. Le seul promoteur apparent du gène à tropisme hépatique BAAT humain (*bile acid coA; Amino acid N-acyltransferase*) est une LTR de la famille HERV-K HML-8, insérée en amont du premier exon. Aucun élément transposable n'existe à proximité du gène orthologue chez la souris bien que des niveaux d'expression similaires dans le foie soient constatés (van de Lagemaat et al. 2003a). Le second exemple touche un membre de la famille des alcool déshydrogénases de classe 1. Chez l'homme, trois gènes ADH1 issus de duplications interviennent dans la dégradation de l'éthanol mais ne présentent pas la même spécificité d'expression tissulaire hors du foie. Seul le gène ADH1C se distingue par une insertion d'une LTR de la famille ERV-9 dans son voisinage amont. L'analyse des sites de fixation des facteurs de transcription de la séquence de la LTR a fait apparaitre des éléments de réponse génériques et tissu-spécifiques qui expliquent à eux-seuls la spécificité d'expression de ADH1C (Chen et al. 2002).

Des contributions mineures de l'activité promotrice d'une LTR à l'expression de son gène voisin ont également été décrites. Dans le foie, toujours, le gène de l'apolipoprotéine C1 est sous la dépendance de deux promoteurs, dont une LTR de la famille HERV-E (Medstrand et al. 2001). Il est estimé que jusqu'à 15 % de l'activité transcriptionnelle du gène APOC1 serait apportée par son promoteur rétroviral. Une étude plus approfondie de la fraction de transcrits dépendant de l'activité d'un promoteur rétroviral alternatif a été réalisée sur le gène de la sémaphorine 4D (Cohen et al. 2009) (Figure I-9). SEMA4D partage en effet l'initiation naturelle de sa transcription avec l'utilisation de la fonction promotrice d'une LTR de la famille ERV-9, et son expression est associée à de multiples contextes cellulaires. Ainsi, sur un panel de tissus, il a été montré que l'initiation de transcription dans la LTR peut, dans certains cas, dépasser 20 % de la transcription globale, et notamment dans le cerveau et les lymphocytes T, contextes pour lesquels l'activité fonctionnelle de la protéine est connue (Kumanogoh and Kikutani 2004). Ce double système d'expression peut alors être vu soit comme un avantage évolutif consolidant un tropisme tissulaire, soit comme une phase transitoire de substitution d'un promoteur cellulaire au profil d'un promoteur rétroviral, ayant éventuellement pour conséquence à terme de limiter le tropisme d'expression du gène.

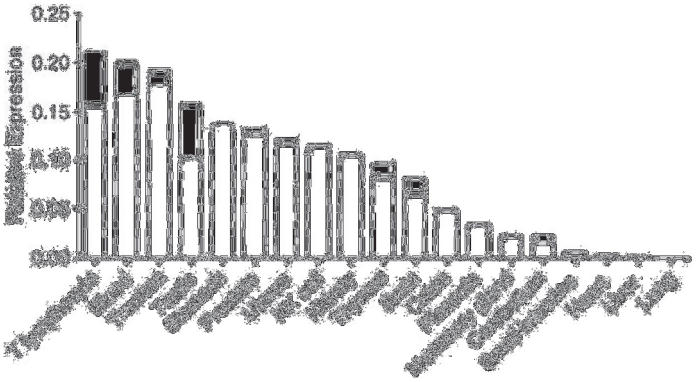


Figure I-9 Contribution de l'activité promotrice de la LTR ERV-9 à l'expression de SEMA4D. La fraction noire des barres correspond à l'activité apportée par la LTR. D'après (Cohen et al. 2009).

Des situations plus complexes d'initiation de la transcription par des promoteurs peuvent également intervenir, par exemple dans le cadre de la coexistence de variants d'épissage. Des travaux récents ont par exemple révélé l'implication d'une LTR de la famille ERV-9 dans les stratégies d'épissage du locus p63. L'intégration de cette LTR est récente d'un point de vue évolutif puisqu'elle n'est présente que chez les hommes et les grands singes. Différents transcrits p63 sont impliqués dans le maintien des cellules de la lignée germinale femelle, mais il a été montré que l'isoforme transcriptionnel initié dans la LTR, et sa protéine correspondante baptisée GTAp63, sont spécifiquement exprimés dans les stades précurseurs des cellules sexuelles mâles (Beyer et al. 2011; Liu and Eiden 2011). De manière assez analogue à la fonction de la protéine native exercée chez la femme, GTAp63 est stimulée par la voie des caspases en réponse à des dommages génétiques et participe aux mécanismes de l'apoptose.

I.2.1.2.2. LTRs enhancer

La mise en évidence d'éléments LTRs enhancer est compliquée par la distance (parfois plusieurs centaines de kilobases) qui peut séparer une séquence régulatrice du ou des gènes qu'elle régule. L'analyse des marques histones contribue à identifier des groupes d'éléments co-régulés, mais les relations de cause à effet entre ces éléments restent souvent difficiles à clarifier. En particulier, le sens d'insertion d'une LTR ne permet pas de privilégier certaines hypothèses d'association dans la mesure où la fonction d'enhancer est par définition indépendante de l'orientation de la séquence. La première démonstration d'une fonction enhancer LTR, hors placenta, a été achevée en étudiant le tropisme d'expression des isoformes de l'amylase salivaire, AMY1 et AMY2. Ces enzymes, qui dégradent l'amidon en maltose, s'expriment de manière assez spécifique dans la parotide (AMY1) et dans le pancréas (AMY2). Il a été montré que l'expression parotidienne de AMY1 est conférée par la présence d'une LTR HERV-E située environ 1 kb en amont du gène, dont est privé AMY2, et que cette séquence est suffisante pour transférer une spécificité d'expression salivaire au promoteur du gène de la thymidine kinase (Samuelson et al. 1988; Ting et al. 1992). Plus

tard, l'étude du locus contrôle de la beta globine (β-LCR) a révélé que l'insertion d'une LTR ERV-9, environ 2 kb en amont, amplifie la transcription du locus dans les tissus embryonnaires et différentes cellules de la lignée hématopoïétique (Ling et al. 2002). Des transcrits initiés dans la région U5 de la LTR et qui s'étendent dans le locus contrôle ont également été caractérisés, indiquant vraisemblablement une double activité d'enhancer et de promoteur. En utilisant comme modèle le poisson zèbre, il a par la suite été suggéré que la LTR ERV-9 du β-LCR pourrait, chez l'humain, contribuer à l'activation de loci maternels requis dans les stades précoces de l'embryogénèse ou, alternativement, participer à la potentialisation de la structure chromatinienne dans l'oocyte et les cellules souches progénitrices de l'adulte (Pi et al. 2004). Enfin, très récemment, les modes de régulation de l'expression de la kallikréine 3 (KLK3) par la voie des androgènes ont été complétés par la démonstration de l'effet enhancer d'une LTR de la famille HERV-L40 (Lawrence et al. 2012). Le recrutement du récepteur aux androgènes sur cette séquence LTR distale se traduit par une augmentation de l'expression de certaines kallicréines dans des modèles cellulaires et, au niveau moléculaire, il a été montré que la duplication de la LTR est requise pour obtenir une réponse maximale aux androgènes.

I.2.1.2.1. LTRs polyA

La caractérisation de signaux polyA naturels ou alternatifs amenés par une LTR a fait l'objet d'un nombre plus restreint d'études, basées principalement sur l'analyse de banques d'ADNc et d'ESTs. Une des premières observations de transcrits chimériques dont le signal AATAAA est apporté par une séquence rétrovirale endogène a été faite en criblant une banque d'ADNc de cellules du sang (Mager 1989). Dans quelques-unes des occurrences retenues, une LTR de la famille HERV-H a été identifiée comme source du signal de polyadénylation, terminant des transcrits initiés dans des régions non HERV. Dix ans plus tard, par l'exploitation des ESTs, la même équipe a caractérisé deux gènes humains dont le promoteur naturel est une LTR HERV-H (Mager et al. 1999). Le premier de ces gènes, nommé HHLA2 pour HERV-H LTR-associating 2, possède un cadre de lecture ouvert de 414 acides aminés organisés en trois domaines homologues aux immunoglobulines, et s'exprime dans l'intestin, les reins et les poumons. Le second gène, HHLA3, a un tropisme d'expression moins restreint et code différentes protéines aux fonctions mal connues, selon des stratégies d'épissages alternatifs. L'analyse des gènes orthologues à HHLA2 et HHLA3 chez le babouin a montré que, en l'absence des LTRs, un signal polyA est apporté par des séquences cellulaires (Mager et al. 1999). Différents niveaux de preuves expérimentales ont également été apportés par d'autres équipes à partir de quelques modèles d'étude. Il a ainsi été proposé que la LTR 5' du provirus HERV-Fb puisse remplir une fonction de site polyA alternatif pour le gène ZNF195 (Kjellman et al. 1999), ou encore que des séquences LTRs de la famille HERV-K apporteraient les signaux de polyadénylation à des transcrits cellulaires, enregistrés dans des banques d'ADNc mais d'origine mal caractérisée (Baust et

al. 2000). Malgré le faible niveau de connaissances associées, à ce jour, à l'activité polyA des séquences HERV, il n'est pas déraisonnable de penser que le potentiel de terminaison des LTRs du génome est d'une ampleur comparable à celui, mieux connu, d'initiation de la transcription. Des approches bioinformatiques récentes, que ce soit chez la souris ou chez l'homme, contribuent actuellement au développement de méthodes d'analyses complexes pour la recherche d'éléments fonctionnels de type polyA (Kim and Hahn 2011; Li et al. 2012), mais n'ont pas encore fait l'objet de validations expérimentales.

I.2.2. Contextes pathologiques

Comme nous l'avons vu dans le chapitre introductif sur la domestication rétrovirale, la dérégulation d'un élément HERV domestiqué peut conduire à la désorganisation d'un tissu, à l'augmentation des phénomènes apoptotiques et *in fine* à un échec de développement de l'organisme. Le cas de la syncytine-1 est toutefois particulier puisque l'évolution pathologique dans le placenta (pré-éclampsie) est associée à une baisse du niveau d'expression du gène. La grande majorité des observations concerne plutôt une réactivation transcriptionnelle des HERV en contextes pathologiques, en lien avec des hypothèses de contrôle épigénétiques, de recombinaison génique ou d'initiation de la transcription par des éléments promoteurs alternatifs. A ce titre, une réactivation de la Syncytine-1 hors de son contexte d'expression naturel est associée au cancer du sein (Larsson et al. 2007) de l'endomètre (Strick et al. 2007) et de certaines formes de cancers colorectaux (Larsen et al. 2009).

I.2.2.1. Activité des HERV

I.2.2.1.1. Pathologies cancéreuses

De nombreuses études ont associé l'activation de séquences HERV avec des états cancéreux (Tableau I-2). Ces descriptions se font essentiellement au niveau transcriptionnel de familles de rétrovirus et vont parfois jusqu'à la caractérisation de protéines ou de particules virales dans des modèles d'étude précis. L'implication des HERV en tant qu'effecteurs dans le développement de maladies cancéreuses reste cependant discutée dans la plupart des cas. Nous présenterons ici les références principales en développant les travaux les plus récents.

Tumeur	Famille HERV	Détection	Région	Références
Cerveau	H,K,W	ARN	*gag, env*	(Balaj et al. 2011; Flockerzi et al. 2008; Skog et al. 2008)
Foie	H	ARN	*env*	(Ahn and Kim 2009)
Gastrique Intestinale	H,K,W	ARN, prot.	*gag, pol, env*	(Alves et al. 2008; Larsen et al. 2009; Liang et al. 2009a; Liang et al. 2012; Liang et al. 2007; Mullins and Linnebacher 2012; Pichon et al. 2006; Stauffer et al. 2004; Wentzensen et al. 2007; Wentzensen et al. 2004; Willer et al. 1997)
Leucémie Lymphome	K,H,E	ARN, prot, part.	*gag, pol, env*	(Contreras-Galindo et al. 2008; Depil et al. 2002; Patzke et al. 2002; Prusty et al. 2008)
H9, H562, Jurkat HL60 et autres	K,H,E (MaLR)	ARN	*LTR, gag, pol, env*	(Brodsky et al. 1993; Iwabuchi et al. 2004; Lamprecht et al. 2010; Lindeskog and Blomberg 1997; Patzke et al. 2002; Prusty et al. 2008; Willer et al. 1997)
Mélanome	K	ARN, prot part.	*gag, env, rec, np9*	(Buscher et al. 2005; Buscher et al. 2006; Hahn et al. 1998; Muster et al. 2003; Schiavetti et al. 2002)
HNEM HEK293 Et autres	K	ARN, prot, part.	*pol, env, np9*	(Katoh et al. 2011; Reiche et al. 2010; Schanab et al. 2011; Serafino et al. 2009)
Ovaire-endomètre	K,E,R,W	ARN, prot.	*gag, pol, env*	(Hu et al. 2006; Menendez et al. 2004; Pichon et al. 2006; Strick et al. 2007; Wang-Johanning et al. 2007)
Pancréas	K, H	ARN	*gag, env*	(Schmitz-Winnenthal et al. 2007; Wentzensen et al. 2007)
Poumon	K,E,R,H	ARN	*LTR, pol, env*	(Ahn and Kim 2009; Pichon et al. 2006; Tomita et al. 1990)
Prostate	K,E,L	ARN, prot.	*gag, pol, env*	(Bai et al. 2007; Goering et al. 2011; Ishida et al. 2008; Molinaro et al. 2006; Pichon et al. 2006; Stauffer et al. 2004; Tomlins et al. 2007a; Wang-Johanning et al. 2003a)
Sein	K,E	ARN, prot.	*gag, env*	(Frank et al. 2008; Golan et al. 2008; Pichon et al. 2006; Wang-Johanning et al. 2003b; Wang-Johanning et al. 2001; Wang-Johanning et al. 2008; Zhao et al. 2011)
T47D, MCF7, MB-231 et autres	K,E,F,W, T, FRD,I	ARN, prot.	*gag, pol, env*	(Bjerregaard et al. 2006; Contreras-Galindo et al. 2008; Ejthadi et al. 2005; Larsson et al. 2007; Patience et al. 1996; Seifarth et al. 1995; Wang et al. 2001; Wang-Johanning et al. 2012; Willer et al. 1997; Yin et al. 1997)
Testicules	K,H,2	ARN, prot, part.	*LTR, gag, pol, env, rec*	(Ahn and Kim 2009; Boller et al. 1993; Flockerzi et al. 2008; Gimenez et al. 2010; Herbst et al. 1996; Kleiman et al. 2004; Lower et al. 1993; Pichon et al. 2006; Sauter et al. 1995; Trejbalova et al. 2011; Vinogradova et al. 2002)

Tableau I-2 Activité des HERV en contextes cancéreux. Détection de l'activité HERV au niveau transcriptionnel (ARN), traductionnel (prot.) ou particulaire (part.). La présentation s'attache autant que possible à faire une distinction entre ce qui a été trouvé à partir de prélèvements biologiques humains et ce qui est observé exclusivement sur modèles cellulaires.

I.2.2.1.1.1. Mise en évidence à partir de cellules tumorales

I.2.2.1.1.1.1. Activité transcriptionnelle

I.2.2.1.1.1.1.1. HERV-K HML-2

La super-famille HERV-K, et plus particulièrement le sous-groupe HERV-K HML-2, ont été abondamment étudiés dans les cancers. La famille HML-2 est constituée d'environ 60 provirus, dont la plupart ont conservé des cadres de lectures ouverts pour tout ou partie des gènes viraux (Subramanian et al. 2011). Les provirus HML-2 sont classés en fonction de la présence (type 1) ou de l'absence (type 2) d'une délétion de 292 bases dans le gène de l'enveloppe, conduisant à la fusion des gènes *pol* et *env* (Figure I-10). Le type 2, qui est le prototype proviral HML-2, produit des transcrits sous-génomiques mono-épissés *env* et des transcrits de petites tailles résultant d'épissages supplémentaires ou alternatifs. Parmi ceux-là, un transcrit de 1,8 kb code une petite protéine de régulation appelée cORF (aussi désignée par Rec) et ayant des similarités avec la protéine Rev du VIH. En l'absence des 292 bases, les provirus HML-2 de type 1 présentent un codant stop dans l'enveloppe. Une stratégie d'épissage des provirus de type 1 peut cependant produire des transcrits courts qui codent une protéine appelée Np9, et dont les 15 premiers acides aminés sont communs avec cORF.

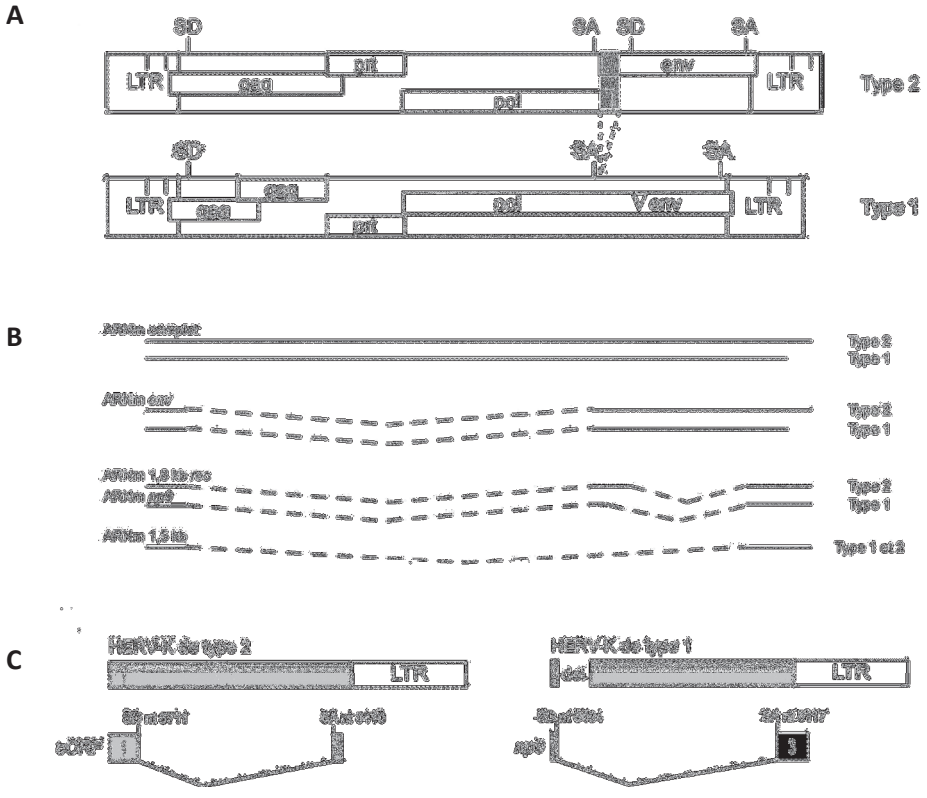


Figure I-10 Provirus, transcrits et protéines HERV-K HML-2 (A) Provirus de type 1 et 2. Une délétion de 292 bases dans le gène de l'enveloppe caractérise les provirus HML-2 de type 1 et conduit à la fusion de *pol* et *env*. Le codon stop ainsi créé est symbolisé par un triangle. (B) Variants d'épissages possibles pour les transcrits des provirus de type 1 et 2. (C) Régions codantes (numérotées de 1 à 3) des épissages cORF et np9. D'après (Lower et al. 1995) et (Armbruester et al. 2002).

L'activité transcriptionnelle de la famille HML-2 a été associée à des contextes cancéreux variés : (i) dans les mélanomes, l'application d'UV ou de restrictions nutritives à des cellules en culture conduit à une activation transcriptionnelle des gènes *pol* et *env* (Schanab et al. 2011; Serafino et al. 2009) ainsi que des transcrits d'épissage d'enveloppe Rec et Np9 (Buscher et al. 2006; Reiche et al. 2010). *Ex vivo*, la comparaison de mélanocytes néonataux et de biopsies de mélanomes métastatiques a montré que les transcrits d'épissages de l'enveloppe sont absents des cellules saines (Buscher et al. 2005). L'observation de transcrits HML-2 de type 1 et 2 a également été associée (ii) au cancer du sein (Ejthadi et al. 2005; Pichon et al. 2006; Wang-Johanning et al. 2001; Wang-Johanning et al. 2003b) où le niveau d'expression de l'enveloppe a en particulier été relié à un pronostique de survie sur une cohorte chinoise (Zhao et al. 2011), (iii) au cancer de l'ovaire où la détection de transcrits d'enveloppe est plus importante dans les tissus cancéreux épithéliaux que dans les tissus normaux (Wang-Johanning et al. 2007), (iv) au cancer de la prostate où l'expression différentielle d'une séquence HML-2 a été mise en évidence chez certains patients (Bai et al. 2007), (v) dans les lymphomes, où des titrages importants d'ARNm de séquences HML-2 ont été réalisés (Contreras-Galindo et al. 2008) puis confirmés pour une séquence *gag*, exprimée 5 à 10 fois plus dans le sang de malades que dans les PBMC d'individus sains (Depil et al. 2002) et (vi) pour des échantillons des cellules de la lignée germinale (Herbst et al. 1996; Herbst et al. 1998; Lower et al. 1995) et en particulier de séminomes pour lesquels une expression différentielle de la famille a été observée entre des échantillons sains et cancéreux (Ahn and Kim 2009; Pichon et al. 2006). Toutes ces observations portent généralement sur le comportement global de la famille HML-2 ou sur l'étude de loci candidats. Il est intéressant de noter qu'une étude de l'expression de séquences HML-2 dans les cancers de la lignée germinale a permis d'identifier la surexpression de 23 provirus individuels, bien que la contribution de chaque locus reste à préciser (Flockerzi et al. 2008). Si la famille HML-2 a été largement étudiée en raison de son insertion récente dans le génome et de la conservation des structures provirales, d'autres familles présentent une activité transcriptionnelle en contextes cancéreux, ce qui tend à montrer que la surexpression de séquences HERV peut toucher un large spectre d'éléments.

I.2.2.1.1.1.1.2. HERV-H

L'expression de la famille HERV-H a été fortement associée au cancer du côlon. Après une première séries observations d'expressions différentielles dans des tumeurs du côlon (Pichon et al. 2006; Wentzensen et al. 2004), différents loci HERV-H ont été évalués sur des cohortes de patients (Alves et al. 2008; Liang et al. 2009a; Liang et al. 2009b; Liang et al. 2007; Wentzensen et al. 2007). Un locus en particulier, désigné par Xp22.3, présente une forte réactivation transcriptionnelle en contexte cancéreux pour au moins 50 % des patients inclus dans les études. Il faut noter que cette réexpression semble assez spécifique du cancer du côlon, ce qui peut renforcer des hypothèses

étiologiques associées à cette séquence HERV et ouvrir la voie à l'utilisation de ce locus comme un marqueur spécifique de pathologie. La recherche de nouvelles séquences HERV-H associées au cancer du côlon a récemment conduit à l'identification d'une vingtaine de provirus exprimés dans des tissus cancéreux et des lignées cellulaires, sans toutefois fournir d'information sur les niveaux d'expression respectifs ni sur la variabilité inter individus (Liang et al. 2012). Malgré un tropisme d'expression manifeste dans le côlon, des ARNm de séquences HERV-H ont aussi été détectés dans les cancers du foie, du poumon, et du testicule (Ahn and Kim 2009), du pancréas (Wentzensen et al. 2007) et de manière plus discutable dans des cellules de leucémies myéloïdes et lymphoïdes où une séquence apparentée à HERV-H mais ayant un PBS complémentaire à la phénylalanine (F) a été isolée (Patzke et al. 2002).

I.2.2.1.1.1.1.3. HERV-E

La première association de transcrits de la famille HERV-E en situation cancéreuse est réalisée en 2003 avec la détection de plusieurs transcrits d'enveloppe dans environ 40 % des adénomes prostatiques testés, et l'absence d'expression dans les tissus sains contrôles (Wang-Johanning et al. 2003a). Plus tard, la réactivation de la famille HERV-E dans le cancer de la prostate a été montrée au laboratoire (Pichon et al. 2006). Bien que peu de travaux aient porté spécifiquement sur l'étude de la famille HERV-E dans le cancer de la prostate, il est intéressant de noter qu'une équipe américaine, qui exploitait des modèles cellulaires de métastases prostatiques pour étudier les défenses virales *via* la réponse interféron, a mis en évidence que des transcrits d'enveloppe de la famille HERV-E peuvent se fixer à l'oligoadenylate synthetase (OAS) et déclencher l'activation du système de défense virale RNAse L (Molinaro et al. 2006). Les auteurs de ce travail suggèrent que l'activation du système OAS/RNAse L par des séquences ARN HERV réactivées participerait d'un processus plus protecteur que nuisible pour l'organisme, en initiant notamment les mécanismes de défense antiprolifératifs nécessaires à l'élimination des cellules tumorales. En marge du cancer de la prostate, une réexpression de séquences d'enveloppes a été décrite dans des tumeurs ovariennes (Wang-Johanning et al. 2007), des hausses d'expression de séquences *pol* HERV-E ont été observées dans certains échantillons de carcinomes mammaires (Frank et al. 2008), et différents transcrits, dont un de 7,1 kb incluant les séquences *gag*, *pol* et *env* d'un locus en position 8p23, ont été détectés dans des cellules hématopoïétiques malignes (Prusty et al. 2008).

I.2.2.1.1.1.1.4. HERV-W

La famille HERV-W, initialement découverte par le criblage d'une banque d'ARN placentaire (Blond et al. 1999), est composée de plusieurs séquences dont la transcription a été associée à des contextes tumoraux. Le chapitre introductif sur la domestication rétrovirale a décrit l'expression du locus ERVWE1/Syncytin-1 en contextes cancéreux, dont on rappelle brièvement ici que les cancers testiculaire (Gimenez et al. 2010; Trejbalova et al. 2011), du sein (Bjerregaard et al. 2006), de l'ovaire

(Strick et al. 2007), du côlon (Larsen et al. 2009) et les lymphomes (Sun et al. 2010) sont les principaux terrains de connaissances aujourd'hui. L'étude des séquences HERV-W du génome, d'abord au niveau global de la famille (Pichon et al. 2006), puis au niveau de loci individuels (Gimenez et al. 2010) a cependant progressivement permis, au laboratoire, de mettre en évidence la réexpression de nouvelles séquences HERV-W dans le cancer du testicule. Ainsi, l'activation de 5 loci, dont une LTR solitaire, a constitué la première caractérisation de séquences HERV-W hors Syncytine-1 dans un cancer de la lignée germinale (Gimenez et al. 2010).

I.2.2.1.1.1.2. Détections protéiques et contribution à l'oncogenèse

La détection de protéines HERV dans des cellules tumorales a été réalisée dans plusieurs contextes (Buscher et al. 2006; Contreras-Galindo et al. 2008; Golan et al. 2008; Lower et al. 1995; Sauter et al. 1995), mais un des meilleures éléments de preuve de leur implication dans les pathologies cancéreuses découle de l'étude des petites protéines de régulation Rec et Np9, appartenant à la famille HML-2. Chez la souris, l'induction de l'expression de Rec altère le développement des cellules de la lignée germinale, ce qui conduit à des lésions du tissu testiculaire proches de celles observées dans un cas de séminome humain (Galli et al. 2005). La détection de transcrits Rec dans les tumeurs humaines de la lignée germinale (Lower et al. 1995), mais aussi la présence de séquences Rec et Np9 dans les mélanomes (Buscher et al. 2006; Reiche et al. 2010) a posé la question des mécanismes d'action de ces protéines dans le développement cancéreux. Il a pu être montré que Rec et Np9 interagissent avec la protéine à doigts de zinc promyelocytique PLZF (Denne et al. 2007), dont la fonction naturelle est d'inhiber la transcription du proto-oncogène *c-myc*. La co-expression de Rec et Np9 avec PLZF lève l'inhibition exercée sur *c-myc* et se traduit par l'expression des gènes régulés par *c-myc*, parmi lesquels des gènes qui déclenchent la prolifération cellulaire et diminuent l'apoptose. Il a également été établi que la protéine Np9 peut interagir avec le ligand de la protéine Numb X, LNX (Armbruester et al. 2004). Cette interaction, qui a pour effet une modification de la localisation cellulaire de LNX, peut se traduire par un déséquilibre de la voie de signalisation Numb/Notch et contribuer ainsi à la dérégulation des équilibres cellulaires naturels. Au-delà du cadre de la famille HML-2, il faut ici préciser que la Syncytine-1 peut, par sa contribution à des phénomènes de prolifération (Strick et al. 2007), de différenciation cellulaire (Frendo et al. 2003) ou par son action anti-apoptotique (Knerr et al. 2007), définir le portrait d'un oncogène. Sur ce point spécifiquement, nous renvoyons à la lecture de la revue 'A comparative portrait of retroviral fusogens and Syncytins', jointe en annexe I du manuscrit.

I.2.2.1.1.3. Observations particulaires

La première observation de particules rétrovirales en contexte cancéreux a été réalisée en 1983 sur une lignée cellulaire issue de tératocarcinomes et qui produit spontanément des particules de 1,16 g/mL ayant une activité reverse transcriptase (Boller et al. 1983). Des caractéristiques immunologiques ont permis de distinguer ces particules des rétrovirus animaux connus, et le nom d'HTDV (Human Teratocarcinoma-Derived Virus) a été proposé pour désigner ce nouveau rétrovirus humain d'origine endogène. L'utilisation d'antisérum spécifique de la matrice de HERV-K et la construction de virus recombinants ont par la suite permis de démontrer que HTDV est codé par la famille HML-2 (Boller et al. 1993; Boller et al. 1997; Lower et al. 1993). Dans le but d'identifier quelles séquences sont à l'origine des particules observées, différentes études ont précisé que des provirus HML-2 de type 1 et de type 2 sont transcrits dans les cellules de tératocarcinomes (Bieda et al. 2001; Lower et al. 1995). Plus récemment, il a été montré que le plus jeune locus de la famille, dénommé HERV-K133, une fois cloné dans un vecteur d'expression de type baculovirus, peut produire des particules rétrovirales dont la morphologie correspond à celle observée dans les cellules de tumeurs de la lignée germinale (Boller et al. 2008). Bien que cette démonstration suggère que HERV-K133 puisse contribuer à la formation de particules endorétrovirales, il n'est pas exclu que des phénomènes plus complexes, par exemple de recombinaison, soient nécessaires pour la production de particules *in vivo*, comme cela a été montré pour la famille HML-2 (Dewannieux et al. 2006; Lee and Bieniasz 2007). Il faut noter également qu'en plus des observations faites sur les cellules de tératocarcinomes, des particules HML-2 ont été détectées dans les surnageants de cellules primaires (Muster et al. 2003) et de lignées cellulaires (Buscher et al. 2006) de mélanomes, ainsi que dans le plasma de patients atteints de lymphomes (Contreras-Galindo et al. 2008). De telles observations permettent d'envisager une mobilité des séquences HERV de la tumeur d'origine, pour permettre des détections directes ou indirectes à distance des cellules cancéreuses.

I.2.2.1.1.2. Détection de l'activité HERV à distance des cellules tumorales

I.2.2.1.1.2.1. Détections directes

Les plus récents développements des travaux de caractérisation des particules rétrovirales HML-2 dans les lignées Tera-1 ont permis de montrer que certains transcrits d'ARNm HERV peuvent être embarqués de manière préférentielle dans les particules (Ruprecht et al. 2008), dans un processus qui pourrait être assez indépendant de l'identité génomique des séquences à l'origine des protéines de structure Gag. Un saut important a été réalisé dans les démonstrations expérimentales lorsque qu'une équipe de Harvard a identifié pour la première fois la présence de séquences d'ARN HERV dans des microvésicules extracellulaires de glioblastomes (Skog et al. 2008). Si l'identification précise des séquences HERV contenues dans les vésicules n'est pas réalisée dans ce travail, il est

clairement défendu par ses auteurs qu'un phénomène d'embarquement préférentiel d'ARNm HERV, ainsi que de micro ARN, est réalisé par la machinerie cellulaire. Ces microvésicules sont également détectables dans le sérum de patients atteints de glioblastome et leur contenu a permis d'identifier un quart des patients malades. Une caractérisation plus précise du contenu des microvésicules sériques de patients atteints de glioblastome et de médulloblastome a révélé une forte présence d'éléments répétés, parmi lesquels des séquences LINE-1 et Alu, mais aussi des transcrits des familles HERV-K, HERV-H et HERV-W (Balaj et al. 2011). Il a par ailleurs été montré que le contenu nucléique de ces microvésicules peut être délivré dans des cellules réceptrices, et participer ainsi du transfert horizontal d'informations génétiques. Si différents éléments répétés ont donc été identifiés dans les microvésicules sériques, il est intéressant de constater que tous ne semblent pas spécifiques d'un état cancéreux. Ainsi, l'abondance de transcrits Alu-Y s'est révélée être comparable dans des microvésicules de patients sains ou atteints de glioblastome (Noerholm et al. 2012). Ce constat renforce l'association d'ARNm HERV dans des vésicules extracellulaires avec un état cancéreux, et ouvre un segment de recherches biomédicales sur lequel se positionnent des sociétés de diagnostic[3], en mettant au point des méthodes d'extraction et de purification souvent innovantes (Liang et al. 2010). Si l'association avec des microvésicules ou des particules virales reste plus floue, d'autres groupes de recherche ont également réussi à détecter la présence de transcrits ou de protéines HERV dans des fluides biologiques d'individus atteints de cancer : la surexpression d'un transcrit de la famille HERV-K HML-2 a été observée dans les urines et le sang de patients atteints d'un cancer de la prostate (Bai et al. 2007), et des transcrits *gag* et *env* HML-2 ont été détectés dans le plasma d'individus atteints de mélanome et de lymphome (Contreras-Galindo et al. 2008). Cette dernière étude apporte d'ailleurs un des rares cas de détection directe de protéines Gag et Env dans le sérum de patients. Ajoutons quand-même ici qu'une expérience récemment conduite par Hervé Perron semblerait être parvenue à la détection d'antigène Env dans le sérum de patients atteints de sclérose en plaques (Perron et al. 2012).

I.2.2.1.1.2.2. Détection indirecte via la réponse immunitaire

La grande majorité des détections d'antigènes HERV passent en fait par la détection d'une réponse immunitaire. Ainsi, des 1995, il est établi que des anticorps anti Gag et Env HML-2 sont produits par des patients atteints de séminomes et que les titrages d'anticorps sont beaucoup plus faibles chez les individus sains (Sauter et al. 1995; Sauter et al. 1996). Puis, une étude rétrospective sur 2 000 échantillons de sérum a montré qu'une forte réponse immunitaire dirigée contre la famille HML-2 était détectable pour 60 % des patients atteints de séminomes, et que cette réponse décline

[3] Les travaux cités sur le glioblastome sont actuellement exploités par la société Exosomes Diagnostics.

après l'ablation de la tumeur (Boller et al. 1997). Une étude indépendante portant sur une cohorte plus modeste a confirmé la détection dans environ deux tiers des sérum de séminomes testés et l'absence de détection dans les échantillons contrôles, ajoutant que l'apparition d'anticorps anti HML-2 pouvait être observée jusqu'à 6 mois avant le diagnostic de la maladie (Goedert et al. 1999). Enfin, une étude clinique prospective menée de 1996 à 2002 a établi une corrélation entre la détection d'anticorps anti Gag et anti Env HML-2 et l'évolution de la maladie, soulignant en particulier la valeur indicative dans le cadre du succès d'un traitement par chimiothérapie (Kleiman et al. 2004). Le séminome n'est toutefois pas le seul contexte cancéreux pour lequel une réponse immunitaire a été rapportée. Des anticorps anti HERV-K ont été détectés chez des patients atteints de cancer du sein (Wang-Johanning et al. 2008), de la vessie, du foie, du poumon, de l'ovaire, de la prostate et du mélanome (Ishida et al. 2008). Dans le cas du mélanome, la détection d'anticorps anti Gag HML-2 dans le sérum de patients a par ailleurs été reliée au pronostic de survie (Hahn et al. 2008). Bien qu'aucune étude ne fasse mention de titrages d'anticorps HERV dans le sang de patients atteints du cancer du côlon, il est à noter pour finir qu'une équipe allemande vient récemment de démontrer que la séquence d'enveloppe du locus HERV-H Xp22.3 peut activer la prolifération des cellules T CD8 (Mullins and Linnebacher 2012).

I.2.2.1.2. Pathologies du système immunitaire

I.2.2.1.2.1. Maladies auto-immunes

I.2.2.1.2.1.1. Polyarthrite rhumatoïde

La polyarthrite rhumatoïde est une maladie auto-immune d'origine inconnue. Elle se manifeste par des inflammations chroniques des articulations et peut conduire à la déformation puis à la destruction des articulations atteintes. L'hypothèse d'une infection par un rétrovirus exogène humain a été envisagée très tôt (Hart et al. 1979) et parfois suivie à tort dans le cas des contaminations HRV-5 de lapin (Brand et al. 1999), avant que des séquences d'origine endorétrovirales ne soient associées à cette maladie. Des transcrits ERV-3, ERV-9, HERV-K et HERV-L ont été détectés dans le liquide et les cellules de la synovie de patients malades (Nakagawa et al. 1997; Takeuchi et al. 1995), puis la présence d'ARN apparentés à MSRV/HERV-W a été avérée sur un petit nombre d'échantillons de liquide synovial (Gaudin et al. 2000). En raison des hypothèses de super-antigènes associées à ses capacités codantes, la famille HERV-K HML-2 a été la plus étudiée depuis. Ainsi, la détection du provirus K18 s'est avérée être significativement plus importante dans le sang de patients atteints de forme juvénile de polyarthrite rhumatoïde que dans le sang d'individus sains (Sicat et al. 2005) et l'ARN messager de la région *gag* du provirus K10 a été trouvé à un niveau plus important dans le sang de patients malades que dans les prélèvements de donneurs en bonne santé (Ejtehadi et al. 2006). Les transcrits d'enveloppe HML-2 rec et np9, ainsi que la protéine Rec,

ont été caractérisés dans des cellules de synovie en conditions normale et pathologique, et il a été suggéré que des déséquilibres dans les formes d'épissages puissent être liés au développement de la maladie (Ehlhardt et al. 2006). Notons aussi que des charges virales importantes, correspondant aux provirus HML-2 de type 1 et 2, ont été mises en évidence dans le plasma et le liquide synovial de la plupart des patients atteints de polyarthrite rhumatoïde, et que le type-1 en particulier semble associé aux phases actives de la maladie (Reynier et al. 2009). Face à ces différents constats de réactivation de séquences HML-2, il a été récemment proposé que le terme de polyarthrite rhumatoïde puisse en réalité englober des pathologies distinctes, auxquelles des groupes de séquences HERV participeraient du développement (Freimanis et al. 2010).

I.2.2.1.2.1.2. Lupus érythémateux

Le lupus érythémateux est une maladie inflammatoire chronique qui affecte les tissus conjonctifs. Des hypothèses d'infections par des rétrovirus exogènes ont été proposées à l'étiologie de la maladie (Blomberg et al. 1994; Griffiths et al. 1999; Lipka et al. 1996) mais, finalement, furent peu suivies face à l'accumulation d'évidences que la réaction immunitaire provient d'antigènes du soi (Bengtsson et al. 1996; Krieg et al. 1989). Les premières identifications de peptides immunogènes HERV chez des patients atteints de lupus ont concerné des séquences d'enveloppe de rétrovirus de type C (cf. définition des groupes dans la revue introductive, au paragraphe I.1.2.1.2), parmi lesquels ERV-9 et HERV-H, puis la région *gag* d'une séquence endogène apparentée à HTLV-1 et nommée HRES-1 (Bengtsson et al. 1996). HRES-1 est une copie HERV unique localisée en 1q42 et qui présente des variants alléliques pour un site de restriction HindIII. Des travaux ont suggéré que le génotype de HRES-1 serait lié à la susceptibilité de développer la maladie (Magistrelli et al. 1999), puis il a été établi que ce locus représente un site privilégié de recombinaisons qui pourraient influencer le développement du lupus (Pullmann et al. 2008). Par ailleurs, l'antigène HRES-1/Rab4 régule l'expression de surface de CD4 en participant à la voie de recyclage des endosomes. Il a été montré que HRES-1/Rab4 est surexprimé dans les cellules T de lupus et que cette expression corrèle avec le recyclage de CD4 et CD3, contribuant ainsi à la régulation négative de CD3/TCRzeta *via* la dégradation lysosomale (Perl et al. 2010). Une réponse immunitaire dirigée contre d'autres antigènes HERV a également été détectée dans le sérum de patients atteints de lupus érythémateux. C'est par exemple le cas pour un peptide endogène apparenté à la séquence immunomodulatrice CKS-17, qui aurait le potentiel d'activer les lymphocytes T et de favoriser la production de l'interleukine 6 (Naito et al. 2003). Des ARNm correspondant au peptide CKS-17 ont aussi été identifiés chez des individus atteints de la maladie (Ogasawara et al. 2000; Ogasawara et al. 2001) en lien avec des hypothèses de levée de méthylation (Ogasawara et al. 2003). De manière intéressante, les cellules CD4+ de patients SLE, mais pas les CD8+, semblent avoir une capacité plus faible à méthyler leur ADN (Richardson et al.

1990), et semblent également être associées à des marques histones particulières (acétylation des résidus H3 ou H4 par exemple) (Hu et al. 2008), ces degrés de modification étant possiblement corrélés à l'activité de la maladie (Renaudineau and Youinou 2011). Enfin, la présence de transcrits np9 HML-2 s'est révélée être plus fréquente chez des individus atteints de lupus érythémateux que chez les sujets sains (Ren et al. 2005), sans qu'aucun mécanisme ne soit pour l'heure proposé au développement de la pathologie.

I.2.2.1.2.1.3. Psoriasis

Le psoriasis est une maladie auto-immune de la peau qui affecterait entre 1 % et 3 % de la population mondiale, et dont les causes sont mal connues. En 1997, il a été établi qu'un locus HERV situé dans le complexe majeur d'histocompatibilité et proche du HLA-1 représente un déterminant génétique majeur de la maladie (Trembath et al. 1997). Plus tard, un provirus HERV-K a été caractérisé au sein de ce locus, et un polymorphisme allélique portant sur deux nucléotides de la région dUTPase a été relié aux facteurs de risques de développement de la pathologie (Foerster et al. 2005). Des expériences utilisant des protéines recombinantes ont montré que les dUTPases sauvages et mutées pouvaient toutes les deux induire l'activation de NF-κB par le récepteur Toll-like 2, indépendamment de l'activité enzymatique. Lorsque des cellules primaires sont traitées avec ces protéines, une sécrétion des cytokines impliquées dans la formation des plaques psoriasiques est observée, supportant l'hypothèse d'une contribution de la séquence HERV-K dUTPase (Ariza and Williams 2011). Une étude clinique récente impliquant plusieurs centaines de patients a caractérisé de nouveaux variants alléliques en lien avec la détection d'anticorps anti-dUTPase (Lawrence et al. 2012). La famille HERV-K n'est cependant probablement pas la seule contributrice à la réaction d'auto-immunité. Des anticorps dirigés contre les produits des gènes *gag* et *env* de MLV ont par exemple été détectés à un niveau beaucoup plus élevé dans le sérum d'individus atteints de psoriasis que dans celui des contrôles sains, ce qui suggère l'implication de séquences HERV apparentées au groupe des gammarétrovirus (Moles et al. 2003). Un marquage positif de l'enveloppe de séquences HERV-E a ainsi été observé sur la plupart des échantillons de peau affectés par un psoriasis ou une atopie, bien qu'une expression basale plus faible ait aussi été constatée chez les sujets sains (Bessis et al. 2004), en accord avec l'hypothèse qu'une activité HERV basale existe dans les cellules de la peau et que la prolifération et l'inflammation des kératinocytes accroît cette activité (Moles et al. 2007). Notons enfin que des ARNm d'une séquence HERV-W contenant des cadres de lectures ouverts pour la protéine Gag et la protéase virale sont rarement détectés dans les échantillons normaux alors qu'ils le sont dans près des deux tiers des lésions psoriasiques étudiées (Moles et al. 2005).

I.2.2.1.2.1.4. Diabète de type 1

Le diabète de type 1, dit insulino-dépendant (acronyme anglais IDDM) se caractérise par la destruction des îlots de Langherans du pancréas par les lymphocytes T. Plusieurs groupes de gènes contribuent à la prédisposition individuelle, dont le complexe majeur d'histocompatibilité est le plus important (Vyse and Todd 1996). Les rétrovirus endogènes ont été suspectés de jouer un rôle dans le développement de l'auto-immunité du diabète de type 1 après l'observation que des anticorps anti-insuline possédaient également une activité contre une protéine Gag IAP chez la souris (Hao et al. 1993). Une recherche de superantigènes endorétroviraux a alors été conduite chez l'homme (Conrad et al. 1994) pour finalement aboutir à l'identification d'un provirus, nommé IDDMK$_{1,2}$22, exprimé dans le surnageant de cellules d'îlots de Langherans en culture (Conrad et al. 1997), et dont la partie N-terminale de l'enveloppe affiche un potentiel superantigènique. La caractérisation du provirus a établi qu'il s'agit d'une séquence HML-2 nommée K18, localisée en 1q21, et située dans l'intron d'un gène codant une protéine membranaire impliquée dans la régulation de l'activité lymphocytaire (Hasuike et al. 1999). Par la suite, différentes études ont réfuté l'association de IDDMK$_{1,2}$22/K18 avec le diabète de type 1 (Badenhoop et al. 1999; Jaeckel et al. 1999; Muir et al. 1999), allant jusqu'à la remise en question des propriétés superantigéniques du locus (Azar and Thibodeau 2002). Des travaux aux résultats contradictoires ont alors tenté de relier le polymorphisme de IDDMK$_{1,2}$22/K18 avec une prédisposition à la maladie (Kinjo et al. 2001; Marguerat et al. 2004; Ramos-Lopez et al. 2006). Notons que, parallèlement à l'étude du provirus IDDMK$_{1,2}$22/K18, des hypothèses d'implication de LTRs localisées dans le complexe HLA de classe II ont été suivies (Donner et al. 1999; Krach et al. 2003; Pani et al. 2002). En particulier, l'activité d'une LTR solitaire de la famille HERV-K serait reliée à un facteur de risque, possiblement par la fonction promotrice qu'elle exercerait sur le gène DQB1 voisin. Comme nous le constatons, la connaissance de l'implication des HERV dans le diabète de type 1 reste à ce jour largement imparfaite.

I.2.2.1.2.2. Infection par HIV (et HTLV)

Les deux rétrovirus infectieux exogènes qui représentent actuellement une menace pour l'homme sont le virus T-lymphotropique humain (HTLV) et le virus de l'immunodéficience humaine (HIV). Ils appartiennent respectivement aux genres des delta- et lentivirus. De manière remarquable, ces genres sont absents des familles de rétrovirus endogènes humains, composées de beta- gamma- et spumaviruses. Des formes endogènes de lentivirus ont cependant été caractérisées chez certains primates (Gifford et al. 2008; Gilbert et al. 2009). L'infection par HIV a été le modèle le plus étudié d'interactions entres rétrovirus exogènes et endogènes chez l'homme. Au niveau transcriptionnel, une surexpression de la famille HERV-K est observée *in vitro* dans les cellules T CD4+ et dans les PBMC infectés par HIV, par-rapport aux mêmes cellules non-infectées (Contreras-Galindo et al.

2007). L'analyse de plasma de patients porteurs du VIH a aussi montré une surexpression de transcrits HERV-K (Contreras-Galindo et al. 2006; Contreras-Galindo et al. 2012). La majorité de ces transcrits proviennent d'éléments de la famille HML-2 et, dans une moindre mesure, de la famille HML-3. Différents loci semblent à l'origine de cette expression, mais un locus HML-2 de type 2 en particulier, situé en 4q35, jouerait un rôle prédominant. D'autres éléments de la famille HML-2 détectés dans le plasma, dont la séquence *pol* du provirus K102, ont été associés à l'infection virale par HIV, mais des résultats similaires résultent de l'infection par le virus de l'hépathite C (Laderoute et al. 2007). Ceci suggère l'existence de mécanismes plus généraux de réactivation des séquences HERV-K, dus aux infections virales et à une stimulation du système immunitaire. Bien que ces mécanismes restent encore largement incompris, des travaux récents suggèrent que les protéines de régulation Tat de HIV et Tax de HTLV pourraient activer des LTRs de provirus HERV-K (Gonzalez-Hernandez et al. 2012; Toufaily et al. 2011). Il a par ailleurs été montré que les virions HIV incorporent peu de transcrits HERV, en comparaison avec des systèmes de vecteurs rétroviraux de type MLV (Zeilfelder et al. 2007). Ceci s'explique probablement par le fait que la nucléocapside d'HIV ne reconnaît pas les signaux d'embarquement des ARN des séquences HERV non-lentivirales. Sur l'aspect protéique, bien que des travaux contradictoires discutent les réponses immunitaires anti-HERV dans le cadre de l'infection HIV (Boller et al. 1997; Lefebvre et al. 2011; Vogetseder et al. 1993), une détection directe de la protéine Gag HML-2 a été réalisée dans les lymphocytes T CD4+ de patients infectés, par rapport aux individus non porteurs du virus (Contreras-Galindo et al. 2007). Dans cette étude, les lymphocytes T CD8+, qui ne sont normalement pas infectés par HIV, ont également affiché une sur expression de la protéine Gag HML-2, ce qui peut renforcer l'idée que les mécanismes de réactivation des HERV par les infections virales passent par des étapes intermédiaires. Si des précisions qualitatives et quantitatives sur l'expression des protéines HERV dans les cellules infectées par HIV restent à définir, il a toutefois été montré que la lyse, par les lymphocytes T CD8+, des cellules infectées par HIV, pouvait être activée par la reconnaissance d'épitopes HERV-K, HERV-H, HERV-L et HERV-W (SenGupta et al. 2011), ce qui ouvre la voie à des stratégies de vaccinations. Enfin, la transcomplémentation de protéines HIV et HERV a également fait l'objet de travaux expérimentaux. Il a par exemple été montré que l'intégrase du provirus HERV-K10 peut transcomplémenter un virus HIV défectif pour cette enzyme, bien que le pouvoir infectieux du recombinant s'en trouve fortement réduit (Pestana et al. 1999). De manière plus étonnante mais de façon très inefficace, la Syncytine-1 semble pouvoir complémenter des virions HIV défectifs pour le gène de l'enveloppe, et aboutir ainsi à des particules infectieuses ciblant des cellules CD4 négatives (An et al. 2001). Un tel mécanisme pourrait permettre une expansion du tropisme HIV en contexte d'infection naturelle.

I.2.2.1.3. Pathologies du système nerveux

La cas controversé de l'implication de MSRV et de la Syncytine-1 dans la sclérose en plaque, maladie auto-immune qui affecte le système nerveux, a été discuté dans la revue introductive et ne sera pas abordé de nouveau ici. Le lecteur est invité à se référer aux paragraphes I.1.2.2.3 et I.1.3.3.2 pour se faire une opinion personnelle sur la question. Tout comme dans la SEP, l'expression de séquences HERV, et notamment HERV-W, a été rapportée dans des affections du cerveau, au rang desquelles la maladie d'Alzheimer (Johnston et al. 2001), la schizophrénie (Frank et al. 2005; Karlsson et al. 2001; Perron et al. 2008; Weis et al. 2007; Yao et al. 2008) et des troubles bipolaires et neuropsychiatriques divers (Diem et al. 2012; Flockerzi et al. 2008; Frank et al. 2005; Weis et al. 2007). Plus précisément s'agissant de la schizophrénie, qui est le modèle d'étude ayant apporté le plus d'éléments de compréhension sur l'implication des HERV, des transcrits *pol* de la famille HERV-W ont été détectés dans le liquide céphalo-rachidien pour environ 30 % des malades alors que ces transcrits n'ont pu être mis en évidence chez les individus sains. Une augmentation de l'expression de transcrits HERV-W a également été montrée, *post mortem*, dans la région du cortex préfrontal de patients schizophrènes (Karlsson et al. 2001). Une détection plus importante de transcrits *gag* HERV-W a été obtenue dans les cellules mononucléaires de malades (Yao et al. 2008), bien qu'une diminution des niveaux protéiques Gag HERV-W soit observée dans le cerveau des patients (Weis et al. 2007). Une réponse immunitaire dirigée contre les protéines Env et Gag HERV-W a toutefois été mise en évidence dans le sang de près de la moitié des patients schizophrènes, contre moins de 5 % des individus sains de l'étude (Perron et al. 2008). Un travail récent a proposé que l'augmentation d'expression des séquences d'enveloppe HERV-W entrainerait une dérégulation positive du gène BDNF, impliqué dans la schizophrénie, mais aussi des gènes NTRK2 et DRD3, et augmenterait la phosphorylation de la protéine CREB qui est par ailleurs nécessaire à l'expression de BDNF (Huang et al. 2011), ce qui constitue la première tentative visant à élucider les voies métaboliques affectées par la réactivation de séquences HERV dans la schizophrénie. Certains éléments de la famille HERV-K HML-2 ont également été associés à des cas de schizophrénie, et plus largement à des troubles psychiatriques (Flockerzi et al. 2008; Frank et al. 2005). Il est enfin intéressant de noter que la prise de médicaments antipsychotiques dans la gestion des maladies neuropsychiatriques pourrait contribuer à la réactivation de groupes de séquences HERV, comme l'a proposé une étude allemande à partir d'expériences menées sur un large spectre de lignées cellulaires cérébrales (Diem et al. 2012).

I.2.3. Niveaux de (dé)régulation en contextes (physio)pathologiques

La réactivation des rétrovirus endogènes humains en contextes pathologiques est certainement multifactorielle. Nous développerons ici trois niveaux qui sont parmi les mieux documentés à l'heure actuelle : la génétique (recombinaisons, réarrangements chromosomiques et polymorphismes), l'épigénétique (code histone et méthylation) et le contrôle au niveau transcriptionnel (facteurs de transcription, stratégies d'épissage et promoteurs alternatifs). D'autres influences, comme le stress cellulaire, la régulation hormonale et la transactivation virale seront passées brièvement en revue.

I.2.3.1. Génétique

I.2.3.1.1. Recombinaisons et réarrangements chromosomiques

Les anomalies chromosomiques telles que les translocations, les délétions, les réarrangements et les amplifications sont des caractéristiques communes à beaucoup de pathologies, notamment les cancers (Bayani et al. 2007). Le processus de recombinaison est au cœur même de l'histoire évolutive des HERV (Belshaw et al. 2007; Katzourakis et al. 2007). Toutefois, peu de démonstrations convaincantes ont définitivement fait un lien entre l'observation de l'activité HERV en contexte pathologique et des phénomènes de recombinaisons. Comme nous l'avons vu dans la revue consacrée à la domestication de la Syncytine-1, une des hypothèses qui a été proposée à l'observation de particules virales dans la sclérose en plaques (Perron et al. 1991; Perron et al. 1989) est d'envisager une recombinaison complexe de plusieurs loci HERV-W du génome (Laufer et al. 2009; Roebke et al. 2010). Si de tels mécanismes restent à l'heure actuelle difficiles à décrire, il n'est en revanche pas formellement exclu (compte tenu du polymorphisme humain) que le génome humain contienne les éléments nécessaires à la renaissance de rétrovirus infectieux. En suivant cette logique, la construction d'une séquence provirale consensus à partir des loci HML-2 les mieux conservés du génome a permis de restaurer artificiellement un rétrovirus infectieux, nommé Phoenix (Dewannieux et al. 2006). D'autres équipes ont également tenté la « résurrection » avec succès (Lee and Bieniasz 2007), mais toujours sur des modèles cellulaires d'étude. Notons cependant qu'une lignée de neuroblastome isolée chez la souris produit des rétrovirus infectieux qui résultent très probablement d'évènements de recombinaisons entre deux rétrovirus endogènes murins (Pothlichet et al. 2006).

Des réarrangements chromosomiques de différentes natures et impliquant des séquences HERV commencent à être décrits dans certaines pathologies. En 2006, l'exploitation de données issues d'approches de biologie intégrative révèle des fusions de gènes impliquant, majoritairement, la partie 5'UTR du gène androgéno-dépendant TMPRSS2 et des gènes membres de la famille des ETS

(E26 transformation specific) tels que ERG, ETV1 ou ETV4, dans plus de 60 % des cancers de la prostate (Tomlins et al. 2005; Tomlins et al. 2007a). De manière intéressante, pour une fraction minoritaire des fusions qui impliquent ETV1, des séquences HERV-K, localisées sur les chromosomes 22 (Tomlins et al. 2007a) ou 17 (Hermans et al. 2008), deviennent les partenaires 5'. L'expression de ces deux loci HERV est également inductible par un traitement aux androgènes dans la lignée cellulaire de prostate LNCaP. Il n'est pas possible de conclure en l'état des connaissances sur un caractère strictement contingent d'une telle observation ou sur une implication fonctionnelle de séquences chimériques HERV-gènes à l'étiologie de la maladie. Il est revanche confortant de constater que les réarrangements chromosomiques initiés par des séquences HERV ont été proposés comme mécanismes à l'origine de certains cas d'infertilité masculine (Arruda et al. 2008; Sin et al. 2011).

I.2.3.1.2. Polymorphismes

Différents niveaux de polymorphismes ont été décrits pour les séquences HERV, principalement pour la famille HML-2. Plus particulièrement, les deux provirus les plus jeunes du génome humain, HERV-K133 et HERV-K115, seraient fixés dans environ 30 % et 15 % de la population humaine, respectivement (Turner et al. 2001), et jusqu'à 8 loci HML-2 présenteraient un polymorphisme d'insertion (Belshaw et al. 2005). Des SNPs ont également été montrés sur certains éléments HML-2, par exemple pour le locus tandem HOM (Mayer and Meese 2005). Enfin, des polymorphismes d'insertion et de séquences au sein de la famille HML-2 ont été reliés à des groupes populationnels (Kinjo et al. 2001; Macfarlane and Simmonds 2004; Mamedov et al. 2004), mettant en évidence des différences parfois significatives entre des individus africains, européens et asiatiques. Dans ce contexte, quelques équipes ont tenté d'établir des liens entre polymorphisme et pathologies. Des résultats contrastés découlent d'études sur le locus HERV-K18/IDDMK$_{1,2}$22, dont les propriétés super-antigéniques de l'enveloppe ont été discutées dans l'étiologie du diabète de type 1 (Conrad et al. 1997; Jaeckel et al. 1999). Si aucun polymorphisme du locus K18 n'a pu être mis en évidence dans le cadre de cette maladie (Kinjo et al. 2001; Ramos-Lopez et al. 2006), des variations alléliques de sa région génomique flanquante, et en particulier du gène adjacent CD48, pourraient refléter une susceptibilité génétique au développement du diabète de type 1 (Marguerat et al. 2004). En revanche, des haplotypes de l'enveloppe K18 sont fortement associés au diabète de type 2 mais chez des patients atteints de schizophrénie (Dickerson et al. 2008). L'hypothèse d'un polymorphisme de K18 dans la schizophrénie a, par ailleurs, été réfutée (Nyegaard et al. 2012) ; il a en revanche été fortement suggéré, s'agissant de l'insertion d'un autre élément de la famille HML-2, le locus K115 (Otowa et al. 2006). La prévalence d'insertion du jeune provirus K113, pour sa part, a été reliée à un groupe de patients polonais atteints de sclérose en plaques et de polyarthrite rhumatoïde (Krzysztalowska-Wawrzyniak et al. 2011). Le couple K113 K115, évalué dans le cadre du cancer du

sein, pour finir ce panorama, n'a révélé aucune différence de fréquence d'insertion entre les patients malades et sains (Burmeister et al. 2004). Par là on voit que beaucoup de clarifications restent encore à apporter sur ce sujet, en sortant des modèles d'études opportunistes qui ont pu être utilisés jusqu'à présent, et en adoptant une démarche plus générale d'identification des polymorphismes humains. L'étude en cours des 1 000 génomes promet des avancées sur ces questions (1000 Genomes Project Consortium 2010).

I.2.3.2. Epigénétique

I.2.3.2.1. Code histone

L'insertion, puis les phénomènes de recombinaison, de trans-complémentation, d'expression et de réinfection des nombreuses séquences HERV représentent, pour le génome, un fort potentiel délétère, qui peut se traduire par des mutations germinales ou des transformations cancéreuses. Les mécanismes épigénétiques de répression de l'activité des HERV peuvent alors être vus comme une réponse adaptative de l'hôte. Les signatures associées aux histones constituent un ensemble complexe de modifications pouvant moduler l'expression de l'ADN. Différentes marques peuvent affecter le nucléosome, dont les plus étudiées sont l'acétylation, l'ubiquitination, la phosphorylation et la méthylation de certains résidus. Le contrôle de l'expression des rétrovirus endogènes par le code histone n'a été que très peu étudié chez l'homme. Chez la souris, les connaissances acquises sur les cellules souches embryonnaires peuvent cependant fournir des éléments d'appréciation de ce qui peut se passer en situation pathologique. Notamment, la triméthylation du résidu lysine 9 des histones H3 (H3K9me3) et du résidu lysine 20 des histones H4 (H4K20me3) semble jouer un rôle important dans la répression de l'expression des ERV durant l'embryogénèse (Maksakova et al. 2008; Maksakova et al. 2011; Martens et al. 2005). Ces mécanismes impliqueraient la methyltransferase SETDB1 (Karimi et al. 2011) et une voie indépendante de la méthylation de l'ADN dont fait partie le facteur de transcription intermédiaire KAP1 (Matsui et al. 2010; Rowe et al. 2010). Chez l'homme, la liaison covalente de la biotine au résidu lysine 12 des histones H4 (H4K12bio) et lysine 9 des histones H2A (H2AK9bio) par l'holocarboxylase synthetase (HCS) s'est avéré être un mécanisme de répression pour une LTR de famille HML-5 et une LTR 5' d'un provirus RRHERV-I (Chew et al. 2008). La suppression de l'enzyme HCS chez la drosophile stimule la rétrotransposition dans les cellules de la lignée germinale, et la déplétion en biotine corrèle avec l'augmentation de production de particules virales. Les auteurs de ce travail suggèrent enfin que la biotinylation des histones est un prérequis à la méthylation des LTRs.

I.2.3.2.2. Méthylation

La méthylation des ilots CpG de d'ADN est un mécanisme central de la régulation de l'expression de gènes et des éléments transposables. Au cours des phases de développement embryonnaire, des vagues de déméthylations globales réactivent la transcription de séquences ERV chez la souris lors de la mise en place des motifs d'empreinte parentale (Maksakova et al. 2008; Walsh et al. 1998). Des levées de méthylation générales ou ciblées, plus tard dans la vie de l'organisme, ont pour effet de déréguler l'expression de séquences, tant pour les gènes cellulaires (Hassler and Egger 2012), que pour les éléments répétés (Rodriguez et al. 2008; Szpakowski et al. 2009) et les rétrovirus endogènes (Howard et al. 2008), contribuant ainsi à des désordres d'expression et au développement de cancers. *In vitro*, l'exposition de cellules Tera-1 à l'agent déméthylant 5-azacytidine se traduit par l'augmentation du niveau d'expression des protéines Gag HML-2 (Gotzinger et al. 1996). L'étude de plusieurs LTR 5' HML-2 de la lignée Tera-1 a fait apparaitre différents niveaux de méthylation des sites CpG, en lien avec l'activité transcriptionnelle (Lavie et al. 2005). Plus récemment, l'hypométhylation de 5 LTRs de la famille HERV-W a été associée à la réactivation de l'expression de ces loci HERV dans le séminome (Gimenez et al. 2010), et des observations similaires ont porté sur la LTR 5' du locus ERVWE1/Syncytine-1 (Gimenez et al. 2010; Trejbalova et al. 2011). Le travail sur des échantillons de testicules sains en plus d'échantillons tumoraux et péri-tumoraux semble montrer une bascule progressive d'un état méthylé à déméthylé, induit par un échappement permissif de l'élément de contrôle MaLR en amont du locus de la Syncytine (Gimenez et al. 2010). Des descriptions du niveau de méthylation de séquences HERV existent pour d'autres contextes cancéreux que ceux touchant aux cellules de la lignée germinale, qu'il s'agisse d'évaluation d'un comportement global ou que les démonstrations portent sur un locus en particulier. Il a ainsi été discuté, dès 1999, qu'un phénomène d'hypométhylation global de la famille HML-2 puisse être un évènement précoce dans le développement du cancer de la vessie (Florl et al. 1999), puis, plus tard, qu'une levée de méthylation générale des éléments promoteurs de la famille HERV-W contribuerait au développement du cancer ovarien (Menendez et al. 2004) et que l'augmentation des niveaux d'expression de séquences HML-2 dans le mélanome viendrait de l'activité promotrice des LTR 5' HML-2 déméthylées (Stengel et al. 2010). Dans des lignées cellulaires de cancer du côlon, il a été montré récemment qu'un traitement déméthylant s'accompagne de la réexpression de séquences HERV-H (Liang et al. 2012). Enfin, l'activation aberrante d'une LTR de la famille THE1B, sous l'effet d'une déméthylation, est à l'origine de la transcription du proto-oncogène CSF1R dans le lymphome de Hodgkin (Lamprecht et al. 2010).

I.2.3.3. Transcription

I.2.3.3.1. Facteurs de transcription

La balance des facteurs de transcription présents dans un type cellulaire à un moment donné influence certainement l'expression des HERV par la mise en jeu d'éléments de réponses ubiquitaires ou tissu-spécifiques. Encore une fois, la meilleure caractérisation des sites de fixation de facteurs de transcription concerne le locus ERVWE1/Syncytine-1 (Cheng et al. 2004a; Prudhomme et al. 2004; Yu et al. 2002) comme cela a été développé plus haut pour Ap2, Sp1, Oct1 et GCM (I.1.3.1.1 et I.1.3.1.2), mais différents niveaux de connaissances existent sur les LTRs d'autres familles HERV. Ainsi, pour la famille HERV-K HML-2, il a été montré que la protéine à doigt de zinc YY1, qui active la transcription, forme un complexe avec un site consensus de la région U3 (Knossl et al. 1999). D'autres facteurs ubiquitaires peuvent réguler l'activité des LTRs HML-2, comme cela a été montré expérimentalement pour Sp1 (Sjottem et al. 1996) ou plus récemment pour Sp3 (Fuchs et al. 2011). Des activateurs plus spécifiques d'un type tissulaire viennent également d'être décrits pour la séquence HML-2.HOM dans des modèles cellulaires de mélanomes (Katoh et al. 2011) : la transfection de l'isoforme M de la protéine MITF (*microphthalmia-associated transcription factor*) augmente fortement l'expression de la séquence HERV par sa LTR 5' dans la lignée HEK293. Il peut également être mentionné que trois protéines nucléaires, EFR-1, EFR-2 et EFR-3 sont capables d'interagir avec la partie 5' de la région U3 des LTRs HERV-K, indépendamment de la vague d'intégration dans le génome auxquelles ces séquences se rapportent (Akopov et al. 1998). S'agissant d'autres familles de rétrovirus endogènes, la caractérisation de sites de fixation de facteurs de transcription reste souvent théorique, soit qu'elle se fonde sur des prédictions bioinformatiques, soit que les études fonctionnelles *in vitro* n'apportent que des informations assez générales. Les facteurs Sp1 et Sp3 semblent ainsi pouvoir influencer l'expression des familles ERV-9 (La Mantia et al. 1992) et HERV-H (Anderssen et al. 1997; Nelson et al. 1996; Sjottem et al. 1996). Pour cette dernière famille, des motifs répétés, agissant comme régulateurs négatifs, sont exclusivement associés au groupe phylogénétique des éléments inactifs (Nelson et al. 1996; Schon et al. 2001; Sjottem et al. 1996). Parmi les LTRs HERV-H qui présentent une activité promotrice connue, la copie H6 a été étudiée plus en détails. Sa région U3 possède 4 sites de fixation pour des facteurs de transcription Myb, et des essais de transfection utilisant un système reporter ont montré que l'activité de H6 est fortement augmentée par la fixation des protéines Myb sur sa séquence LTR 5' (de Parseval et al. 1999).

I.2.3.3.2. Variants d'épissage

L'expression de transcrits génomiques et sous-génomiques, auxquels peuvent s'ajouter des épissages alternatifs codant des petites protéines de régulation, constituent, pour les rétrovirus

infectieux, des étapes du cycle de réplication et de propagation. Dans leur version endogène, l'épissage devient pour les rétrovirus un niveau de contrôle transcriptionnel supplémentaire. Une bonne connaissance des différents transcrits générés par le locus domestiqué ERVWE1/Syncytine-1 a été acquise ces 10 dernières années, tant en contextes physiologique que pathologiques, comme cela a été développé dans la revue introductive. D'autres familles HERV ont été étudiées sous l'angle des variations d'épissages, et notamment la famille HML-2, apparentée aux bétarétrovirus, dont certaines séquences provirales du génome peuvent générer des transcrits d'épissage multiple aboutissant aux protéines Rec ou Np9. L'observation de ces transcrits est spécifiquement associée au cancers : l'expression de Rec et Np9 est observée dans les mélanomes (Buscher et al. 2006; Reiche et al. 2010) et Rec est décrit dans les cancers de la lignée germinale (Lower et al. 1995). Bien que moins documentées, des formes particulières de transcrits d'autres familles HERV ont été détectées dans des pathologies cancéreuses. S'agissant du cancer du côlon, une analyse transcriptionnelle des provirus HERV-H présentant ou non une délétion du gène *env* a montré que la transcription n'est pas liée à la présence de la séquence d'enveloppe (Liang et al. 2009a), puis différents profils d'épissages pour quelques loci HERV-H actifs dans des lignées cellulaires ont pu être établis (Liang et al. 2012). Des transcrits génomiques et des épissages multiples de séquences HERV-H ont également été rapportés dans des lignées cellulaires de leucémies, faisant intervenir des sites accepteurs jusque-là inconnus, localisés dans la région *pol* (Lindeskog and Blomberg 1997). Des épissages alternatifs protéase-enveloppe de loci HERV-H ont par ailleurs été trouvés, accompagnés de formes sous-génomiques plus conventionnelles, dans environ 40 % des patients atteints de sclérose en plaque, contre 10 % des individus contrôles (Christensen et al. 2003). D'une manière générale, la signification biologique de ces ARNm HERV-H non codants reste à préciser.

I.2.3.3.3. Promoteurs et promoteurs alternatifs

Un dysfonctionnement des différents niveaux de régulation de l'activité des HERV, et en particulier du contrôle épigénétique, est l'hypothèse mécanistique généralement admise de réactivation de séquences en contexte pathologique. Cette réactivation peut conduire à l'embarquement de séquences HERV de manière passive ou, dans le cas de l'activation d'un élément promoteur ou enhancer, provoquer une expression autonome. Cette dernière situation constitue probablement l'implication principale des HERV à l'étiologie d'une pathologie, par exemple en activant un proto-oncogène voisin, ou en entrant en compétition pour la fixation de facteurs de transcription avec des séquences proximales, ou encore en devenant perturbateur pour un transcrit cellulaire naturel dont la séquence HERV constituerait un intron. Les cancers ont été le modèle d'étude privilégié pour l'identification de LTRs actrices de tels processus. *In vitro*, il a été suggéré que deux LTRs de la famille HERV-H pourraient jouer le rôle de promoteurs alternatifs pour les gènes GSDML et DNAJC15 dans des lignées cellulaires cancéreuses (Sin et al. 2006b; Sin et al. 2006a), sans

qu'une interprétation claire de la fonction de ces transcrits ne soit établie. *In vivo*, un transcrit hybride de 6 kb impliquant une LTR solitaire de la famille HERV-K a été détecté dans le sang de patients atteints de leucémie alors qu'il n'a pas été possible de le mettre en évidence chez les individus sains testés (Cassens et al. 1994). La démonstration la plus aboutie de l'implication d'une LTR promotrice dans le développement des leucémies a été rapportée récemment par une équipe allemande : la réactivation ectopique du proto-oncogène CSF1R dans des cellules B dérivées d'un lymphome de Hodgkin est sous la dépendance d'une LTR THE1B à 6,2 kb en amont du gène, et qui court-circuite son promoteur naturel situé à 25 kb (Lamprecht et al. 2010). Concernant les tumeurs solides, la réactivation du locus Xp22.3 dans le cancer du côlon a été associée à une fonction promotrice de sa LTR 5' (Liang et al. 2009a). Au laboratoire, l'autonomie de transcription de 5 loci HERV-W ré-exprimés dans le cancer testiculaire a été reliée à l'activité promotrice de leur LTRs (Gimenez et al. 2010). Enfin, un procédé d'identification des LTRs HML-2 promotrices du génome a été appliqué aux tumeurs de la lignée germinale et a permis à ses auteurs d'identifier quelques dizaines d'évènements d'initiation de transcription (Buzdin et al. 2006b). Les limites du procédé utilisé dans ce dernier travail posent cependant la question des méthodes d'analyses à mettre en œuvre pour développer une approche fonctionnelle à l'étude du transcriptome HERV.

I.2.3.1. Stress cellulaire

L'expression des rétrovirus endogènes peut être affectée, tout comme les équilibres cellulaires en général, par différents facteurs environnementaux. Plusieurs études ont montré que l'irradiation de cellules cutanées par des UV-B ou des UV-C augmente la transcription de la famille HERV-K HML-2 (Hohenadl et al. 1999; Reiche et al. 2010), pouvant aller jusqu'à l'induction d'une production de particules virales (Schanab et al. 2011). En culture, des chocs thermiques (Vinogradova et al. 2001), des privations nutritives (Nellaker et al. 2006) ou encore les variations de la teneur en oxygène du milieu (Knerr et al. 2003) induisent des modifications d'expression des HERV. L'effet de différents produits chimiques, dont le butyrate de sodium (NaBut), la 5'-iodo-2'-deoxyuridine (IUdR) ou la 5-azacytidine (AzaC) a été évalué sur des cellules de singe, montrant quelques modifications d'expressions des éléments endorétroviraux (Ma et al. 2011). Il a également été avancé, à partir des résultats obtenus sur un petit nombre de patients, que la fumée de cigarette entrainerait une augmentation de l'activité transcriptionnelle de certaines familles HERV (Gabriel et al. 2010).

I.2.3.2. Transactivations virales

Des phénomènes de transactivation de séquences rétrovirales endogènes ont été mis en évidence suite, essentiellement, à l'infection de cellules par différents virus à herpès. Une production de particules virales et l'augmentation de l'activité RT ont été initialement observées dans les

surnageants de cellules de leptoméninges infectées par HSV-1 (*herpes simplex virus type 1*), et un mécanisme de transactivation *via* les protéines ICP0 et ICP4 du virus a été proposé (Perron et al. 1993). Par la suite, il a été montré que l'expression de protéines Gag et Env de la famille HERV-W est induite par HSV-1 dans les neurones et les cellules endothéliales cérébrales (Ruprecht et al. 2006). Par ailleurs, l'infection par le virus d'Epstein-Barr (HHV-4) peut induire l'expression du gène d'enveloppe du provirus K18, aux propriétés superantigéniques (Sutkowski et al. 2001). Des mécanismes associant la protéine LMP-2A du virus et le récepteur lymphocytaire CD21 ont été proposés pour l'activation du superantigène endogène (Hsiao et al. 2006; Hsiao et al. 2009; Sutkowski et al. 2004). La transactivation de K18 a également été démontrée dans le cas d'infection HHV-6A (Tai et al. 2009) et HHV6-B (Turcanova et al. 2009). En dehors des virus à herpès, une équipe a démontré l'induction de séquences HERV-W par l'infection du virus de la grippe A/WNS/33 (Nellaker et al. 2006). Dans ce travail, des transcrits *gag* et *env* sont surexprimés après infection dans les lignées de l'endothélium cérébral, dans des cellules neuronales et astrocytaires ainsi que dans une lignée de monocytes. Une transactivation de la Syncytine-1 a également été observée par cette équipe dans les lignées astrocytaires et monocytaires, sans qu'aucun mécanisme de réactivation des séquences de la famille HERV-W n'ait été établi.

I.2.3.3. Hormones et cytokines

L'activité privilégiée de certaines séquences HERV dans le placenta ou encore dans les glandes surrénales et sébacées indique la possibilité d'une composante de régulation hormonale. Nous ne reviendrons pas sur la description des éléments de contrôle (URE/MaLR) du locus ERVWE1/Syncytine-1, composés en particulier de sites de fixation pour les récepteurs aux glucocorticoïdes et à la progestérone (cf. I.1.3.1.1). Les glandes surrénales sont productrices, entre autres, de cortisol et d'aldostérone, et la régulation du fonctionnement des glandes sébacées fait principalement intervenir des hormones stéroïdiennes androgènes. Il a été montré que la LTR 5' de la famille HERV-R, qui s'exprime dans ces glandes (Andersson et al. 2005), contient des sites récepteurs aux androgènes (Andersson et al. 2002). Plus largement, l'étude de la famille HERV-K a montré que (i) le traitement par l'œstradiol et la progestérone conduit à une surexpression de séquences HML-2, principalement de type 1, dans des cellules de lignée mammaire (Ejthadi et al. 2005; Ono et al. 1987; Wang-Johanning et al. 2001; Wang-Johanning et al. 2003b), (ii) les LTRs de la famille HML-4 contiennent un élément de réponse aux glucocorticoïdes et leur expression est inductible par les hormones stéroïdiennes (Seifarth et al. 1998) et (iii) l'acide rétinoïde pourrait induire l'expression de certains éléments HERV-K dans les cellules de carcinome embryonnaire (Seifarth et al. 1998). Les cytokines, en tant que médiateur de la communication cellulaire, semblent également jouer un rôle dans le contrôle de l'activité des HERV. Il a été montré, sur des cellules primaires vasculaires endothéliales, que le traitement par le TNF-α, l'IL-1α et l'IL-1β augmente l'expression de la famille

HERV-R, et que l'interféron gamma diminue au contraire cette activité (Katsumata et al. 1999). L'interféron alpha permet d'induire l'expression du superantigène K18 à des niveaux comparables à celui d'une transfection du même superantigène constitutivement actif (Stauffer et al. 2001). Enfin, une baisse de réactivité d'anticorps anti-enveloppe HERV-H et HERV-W a été observée comme conséquence d'un traitement à l'IFN-β chez des patients atteints de sclérose en plaques (Petersen et al. 2009).

I.3. Méthodes d'analyse du transcriptome

De nombreuses techniques peuvent être appliquées à l'étude de l'expression d'un gène. Les approches électrophorétiques, de type Northern blot ou RPA (*ribonuclease protection assay*), ont été abondamment utilisées au cours des quarante dernières années ; le procédé général d'hybridation cible/sonde, appliqué *in situ*, permet la détection de transcrits spécifiques dans leur contexte biologique par la méthode FISH, et des constructions de promoteurs cellulaires couplés à des systèmes reporter permettent de travailler sur des modèles de transcription. Pour autant, la notion de transcriptome apparait lorsque l'étude de l'expression passe d'un gène à un groupe de gènes constituant un répertoire génomique. Dans cette perspective, nous orienterons cette présentation technologique sur deux grandes méthodes expérimentales utilisées pour l'étude de répertoires transcriptomiques et qui ont été déclinées en applications spécifiques à l'étude du transcriptome HERV par différentes équipes dont la nôtre : la RT-PCR quantitative et les puces à ADN. Conscients de l'importance qu'occupent les techniques de séquençage de type SAGE ou MPSS, et plus récemment de la place prise par les nouvelles technologies à haut débit (RNA-Seq parmi d'autres), nous réserverons une partie de la discussion générale de la thèse à une remise en perspective du notre périmètre d'étude avec les enjeux technologiques actuels.

I.3.1. Méthodes génériques

I.3.1.1. La RT-PCR quantitative

I.3.1.1.1. Principe

Le principe d'amplification d'ADN par polymérisation en chaine, inventé, si l'on en croit son auteur, sur une route sinueuse de montagne par une nuit de pleine lune (Mullis 1990), constitue l'ossature des méthodes de référence d'étude de l'expression des gènes. La RT-PCR quantitative abaisse la sensibilité de détection théorique jusqu'à 1 transcrit d'ARNm dans un échantillon (en pratique plutôt 10 à 100), par l'amplification d'ADNc sous l'action enzymatique d'une ADN

polymérase thermostable. La quantification est alors basée sur l'interprétation de la cinétique d'amplification (Figure I-11). Les premiers cycles génèrent un signal infraliminaire qui suit une croissance exponentielle, et qui devient détectable lorsque le seuil de sensibilité du lecteur optique est franchi. La valeur expérimentale du nombre de cycles à cet instant (C_T) permet, en fonction de l'efficacité d'amplification, de faire une estimation du nombre initial de copies. La cinétique de réaction s'infléchit dans les cycles qui suivent jusqu'à atteindre une phase plateau en raison de l'apparition de facteurs limitants (maximum du turnover enzymatique, dégradation progressive de l'enzyme, épuisement des dNTP, des amorces ou du système de florescence, etc.).

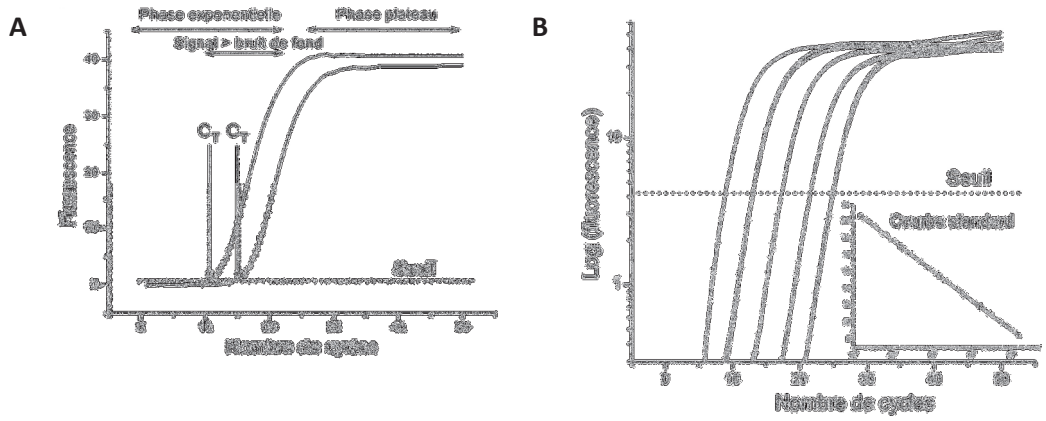


Figure I-11 Détection par RT-PCR quantitative. (A) Courbes de cinétique d'une réaction de PCR quantitative. Le seuil de détection de la phase exponentielle permet de lire la valeur de C_T associée à une réaction. (B) Détermination de la valeur d'efficacité d'un système d'amplification. Représentation logarithmique des cinétiques d'une gamme de dilution pour un système d'amplification donné. La pente de la droite de régression des valeurs de C_T en fonction de la quantité initiale d'ADN donne l'efficacité du système d'amplification.

I.3.1.1.2. Les différents systèmes de flurescence

Deux grandes familles de systèmes de fluorescence peuvent être utilisées pour suivre une réaction de PCR quantitative : des molécules qui s'intercalent dans la double hélice, ou des fluorochromes incorporés à des sondes de détection. La première est le système le plus simple à mettre en œuvre. Il repose sur l'utilisation de composés chimiques de la famille des cyanines asymétriques qui ont la propriété d'émettre une fluorescence une fois complexés à l'ADN bi caténaire. Le SybrGreen est la molécule intercalante de ce type la plus utilisée, et se fixe dans le petit sillon de la structure double brin en émettant une fluorescence à 525 nm. D'autres intercalants, comme le Resolight, le LC Green ou le Boxto, existent, et par des propriétés physico-chimiques légèrement différentes (densité d'insertion, longueurs d'ondes d'émission) peuvent présenter un intérêt particulier dans certaines applications. Toutes ces molécules ont un pouvoir de fixation non spécifique, c'est-à-dire qu'elles sont à l'origine de la détection de tout produit d'ADN amplifié, indépendamment de la cible recherchée. Pour cette raison, des systèmes de détection basés sur

l'utilisation de sondes se sont développés (Figure I-12). Le principe général de cette méthode repose sur l'association d'une séquence spécifique et d'un système de fluorescence inductible. Le système d'induction est basé sur le couplage d'un fluorochrome (dye) qui émet à une certaine longueur d'onde, avec, à proximité immédiate, une molécule qui absorbe ses émissions (quencher). L'éloignement physique du dye et du quencher, conséquence directe ou indirecte de la fixation de la sonde de détection à sa séquence cible, libère la fluorescence dans le milieu réactionnel. Les sondes à structure d'épingle à cheveux Molecular Beacons (MB), longues de 18 à 30 nucléotides, ont été les premières à être développées à cette fin. Elles sont utilisées dans plusieurs variantes, par exemple dans le cadre de la RT-PCR NASBA, ou au travers de sondes de détection Scorpion. Indépendamment de ce système de sondes tige/boucle, la compagnie Roche Molecular Systems a développé des sondes de détection hydrolysables sensibles à l'activité 5' exonucléase de la Taq polymérase. Ces sondes TaqMan, moins contraignantes à concevoir que des sondes Scorpions, connaissent aujourd'hui un large succès d'utilisation.

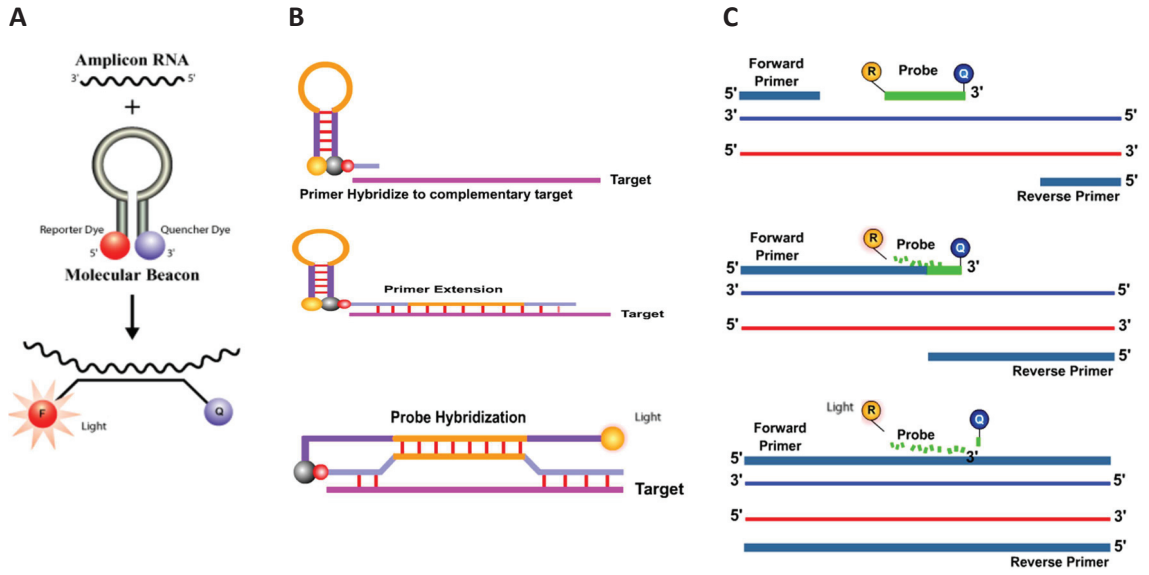


Figure I-12 Principes d'émission de lumière des systèmes de sondes PCR. (A) Sondes Molecular Beacons. En présence de la séquence cible, la sonde se déroule, s'hybride, et la florescence est émise. (B) Sondes Scorpions. L'amorce d'amplification est liée à une sonde de détection de type Molecular Beacons. L'élongation du brin complémentaire entraine un déroulage de la structure en épingle à cheveux de la sonde et l'émission de fluorescence par le dye. (C) Sondes TaqMan. L'activité exonucléase 5' de la Taq polymérase hydrolyse la sonde TaqMan lors de l'étape d'élongation. Le dye (R) est éloigné du quencher (Q) et la fluorescence est libérée dans le milieu.

Par la diversité des molécules fluorescentes existantes, la RT-PCR peut être utilisée pour détecter et quantifier simultanément plusieurs gènes (mais rarement plus de 6) dans la même réaction. Néanmoins, une utilisation fiable de la RT-PCR quantitative repose toujours sur la mise au point et l'utilisation d'amorces spécifiques d'un transcrit donné et nécessite souvent un traitement de l'échantillon qui puisse garantir des conditions d'amplification non biaisées. La contamination par de l'ADN est par exemple un problème fréquent et impose en général de traiter les échantillons à la

DNAse, ce qui peut s'avérer dommageable pour des échantillons précieux obtenus en quantité très faible. Il faut rappeler également que la quantification de l'expression d'un gène doit s'appuyer sur une normalisation qui, en général, dépend du niveau d'expression de gènes cellulaires constitutifs.

I.3.1.2. Les puces à ADN

I.3.1.2.1. Historique introductif aux puces à ADN

Le terme de puces à ADN est relativement ancien et englobe une famille de technologies qui repose sur le principe d'hybridation de séquences d'acides nucléiques complémentaires antiparallèles. La formation de liaisons hydrogènes entre les paires A-T/A-U et G-C permet la capture spécifique d'une cible présente dans un milieu complexe d'acides nucléiques par la sonde qui la complémente. Ce principe d'hybridation est utilisé depuis les techniques de *blots* d'acides nucléiques développées dans les années 1970, mais il aura fallu attendre la transposition sur un support solide et des progrès de miniaturisation relevant de la biologie moléculaire, de l'électronique, de l'informatique et de la chimie des matériaux dans les années 1990 pour voir se développer ces nouveaux outils de détection haute densité, désignés par le terme *microarray* (Figure I-13).

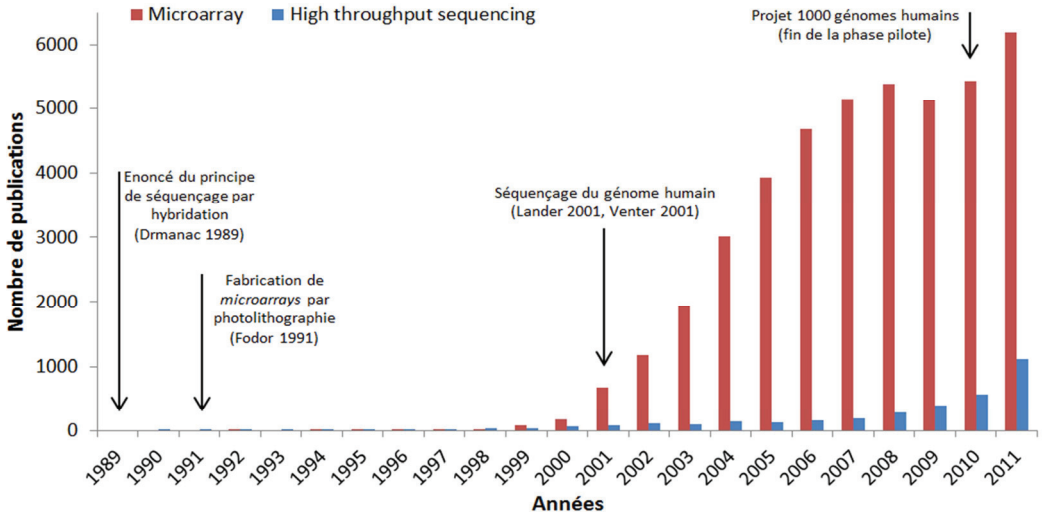


Figure I-13 Evolution du nombre de publications relatives aux puces à ADN et au séquençage à haut débit. Résultats des recherches associant les termes « microarray » ou « high troughput sequencing » à chacune des années de 1989 à 2011 (PubMed).

Les perspectives d'utilisation des puces à ADN ont été radicalement bouleversées par le séquençage du génome humain en 2001 et la mise au point d'outils bioinformatiques nouveaux. Le saut qualitatif qui s'est opéré dans les approches génomiques (détection de mutations, description des polymorphismes, etc.) et transcriptomiques (expression et différentiels d'expression) a rendu possible l'analyse simultanée de plusieurs milliers de gènes, ouvrant la voie à des études des processus cellulaires encore utopiques quelques années auparavant. Parallèlement, la

démocratisation (lente mais régulière) de l'accès aux plateformes d'analyse a permis la diffusion de la technologie dans un grand nombre de laboratoires à travers le monde, ce qui a conduit au développement d'un champ disciplinaire dont il devient impossible de rendre compte d'une manière exhaustive.

I.3.1.2.2. Modes d'étude du transcriptome

De nombreux répertoires nucléiques ont été exploités au travers de générations de puces à ADN dédiées à l'étude du transcriptome, tant chez les espèces animales que chez l'homme. L'évolution du contenu des puces GeneChip Affymetrix humaines est à ce titre une assez bonne illustration des effets de retour sur expérience qui se sont associés à des questionnements biologiques de plus ou plus aboutis. Le transcriptome humain n'était en effet accessible, dans un premier temps, que par la couverture complémentaire de deux puces GeneChip : la puce HG-U133A, basée sur la détection des transcrits de référence (Refseq), et la puce HG-U133B composée principalement de séquences dérivées des banques d'ESTs. L'unification de ses deux répertoires a abouti à la puce HG-U133-PLUS2, dite pan génomique, et qui est aujourd'hui largement utilisée comme système de référence. Parallèlement à ce gain de densité, des puces plus spécifiquement dédiées à l'étude de processus biologiques ont vu le jour. C'est le cas de la puce GeneChip Human Genome Focus qui réduit sa composition à une sélection de gènes clairement documentés et impliqués dans des voix métaboliques connues, facilitant ainsi l'interprétation fonctionnelles des résultats. De nouvelles lectures fonctionnelles se sont également surajoutées progressivement avec le besoin d'étudier les variantes d'épissages d'un même gène, ce qui a conduit au développement de la puce GeneChip Human Exon. Récemment, l'université de Stanford a développé une puce au format GeneChip intégrant à la fois la détection des exons, des épissages alternatifs, de certains SNPs et de 730 transcrits non-codants fonctionnels (Jia et al. 2011) (Figure I-14). Cette nouvelle génération de puce, au nom de projet Glue Grant Array et rebaptisée HTASJ, est actuellement évaluée par un réseau de laboratoires pilotes en vue d'une commercialisation plus large dans un futur proche.

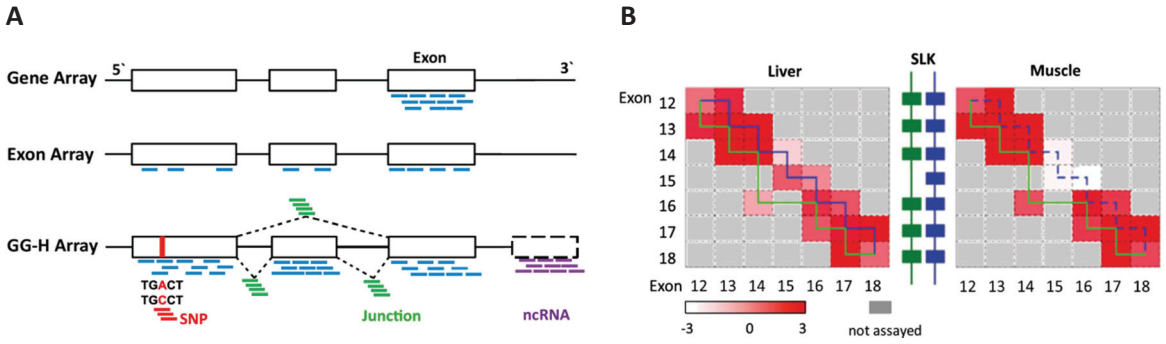


Figure I-14 Analyse qualitative du transcriptome à l'aide de puces à ADN. (A) Trois niveaux d'analyse de l'expression adressés par les puces. (B) Exemple de détection de transcrits et de caractérisation d'un phénomène d'épissage alternatif (exon 15) du gène SLK dans le foie et le muscle. D'après (Jia et al. 2011).

De nombreux travaux de recherche ont confirmé l'intérêt et la pertinence biologique des résultats issus des approches de puces à ADN (Glinsky et al. 2004; Lapointe et al. 2004). La masse d'informations générées par ces approches et la mise en œuvre d'outils et de démarches d'analyse dédiées ont probablement fait basculer la biologie moléculaire dans une aire bioinformatique et systémique qu'elle ne quittera plus[4]. Malgré tout, les générations successives de puces à ADN restent largement confinées aujourd'hui à l'étude des séquences codantes du génome et sont donc loin de couvrir la détection du transcriptome dans son ensemble. La possibilité de développer des puces à ADN à façon pour explorer de nouveaux répertoires nucléiques existe, et passe par la connaissance des procédés de synthèse et une maîtrise de la conception des sondes sur un format donné. Nous abordons dans les paragraphes suivants quelques-uns de ces aspects technologiques en orientant préférentiellement la synthèse d'informations sur la puce GeneChip HG-U133-PLUS2.

I.3.1.2.3. Adressage des sondes

Il existe deux catégories de puces définies par le mode d'adressage de leurs sondes. La technologie *off chip* consiste à greffer ou injecter des sondes pré-synthétisées (par méthode enzymatique ou chimique) sur le support. Différentes mises en œuvre, mécanique (TeleChem, GenemachinesGenetic Microsystems, Genomic Solutions, BioRobotics), piézoélectriques (GeSiM, Packard), ou des dispositifs d'injection d'encre (Lemmo et al. 1998) peuvent être utilisés dans ce but, mais ce mode d'adressage ne permet d'atteindre une densité que de quelque milliers de sondes. La technologie *on chip*, qui utilise un procédé de polymérisation base par base *in situ*, réalise un saut de densité en permettant la synthèse de quelques dizaines de milliers à plusieurs millions de sondes sur un même support. Les dispositifs d'injection d'encre y occupent encore une place (Agilent par exemple), mais c'est véritablement le procédé de synthèse photolithographique qui a ouvert la voie à des applications de très haute densité. La photolithographie *in situ* appliquée aux puces à ADN (Figure I-15) est fondée sur une déprotection spécifique des groupements photolabiles de type MeNPOC (α-méthyl-6-nitropipéronyloxycarbonyle), rendant actives les fonctions hydroxyles à l'extrémité de la chaine, ce qui permet l'ajout successif de nucléotides. L'utilisation de masques, qu'ils soient physiques (Affymetrix) ou virtuels définis *in silico* (NimbleGen et Febit) oriente spécifiquement la synthèse par ces alternances de masquage/démasquage des sondes en construction. La synthèse d'une puce de N mers nécessite ainsi 4xN cycles, quelque soit la taille de la puce, et la résolution du procédé permet d'atteindre une densité supérieure à un million d'oligonucléotides par cellule, pour des cellules de quelques micromètres.

[4] En cela on peut faire le parallèle de l'évolution la biologie moléculaire avec ce qu'ont été les nouveaux besoins de calculs de la physique des particules à la fin du XX^{ième} siècle.

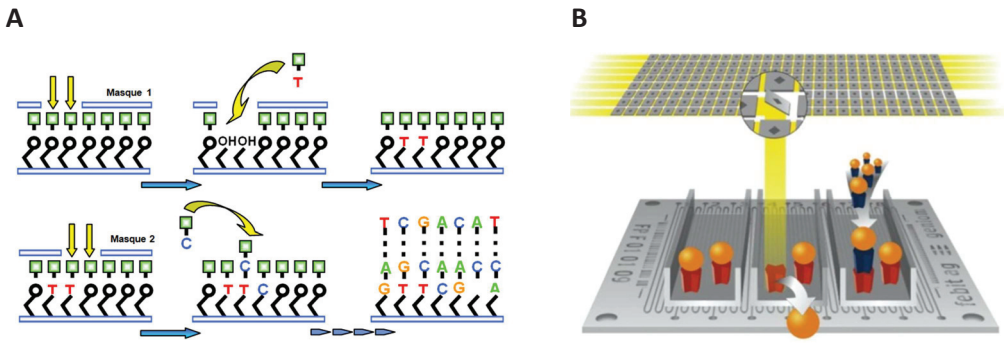


Figure I-15 Principe et mises en œuvre de la photolithographie *in situ*. (A) Principe de la synthèse photolithographique et utilisation de masques physiques (procédé Affymetrix). Un groupement photolabile (carré vert) empêche l'ajout d'oligonucléotides à la chaine. Des masques de chrome, appliqués successivement, laissent passer la lumière UV (flèches jaunes) qui déprotègent alors les extrémités 5' des chaines oligonucléotidiques. La synthèse se poursuit jusqu'aux sondes de 25 mer. (B) Mise en œuvre du procédé photolithographique par l'utilisation d'un masque virtuel (procédé Febit). De gauche à droite : la déprotection par UV et la synthèse qui s'en suit sont sous la dépendance du positionnement de micro miroirs digitaux (partie supérieure) contrôlés par microprocesseur.

I.3.1.2.4. Structure de la puce GeneChip HG-U133-PLUS2

La densité de synthèse gagnée par la maitrise de supports solides et du procédé photolithographique permet aux puces GeneChip (Affymetrix) d'analyser l'expression d'un nombre très élevé de gènes. Une puce GeneChip contient en effet de plusieurs centaines de milliers à quelques millions de sondes de 25 bases, réparties en cellules dont la surface varie de 5 µm² à 11 µm² (Figure I-16). Cela se traduit par l'agencement de « forêts » de sondes identiques au sein d'une cellule (on parle aussi de *feature*, ou de *probe cell*) destinées à la capture des cibles complémentaires qui seront mises en présence *via* le cocktail d'hybridation d'un échantillon.

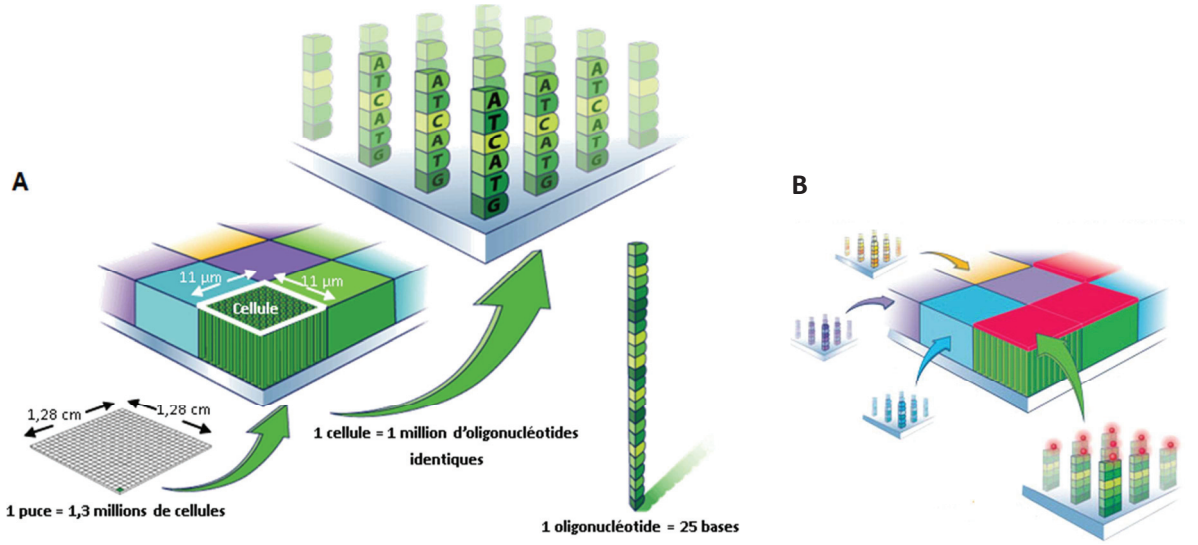


Figure I-16 Structure schématique d'une puce HG-U133-PLUS2 et principe de détection. (A) Vision grossissante depuis la lame de verre visible à l'œil nu jusqu'aux sondes d'ADN simple brin individuelles de 25 bases. (B) Détection de fluorescence associée à une hybridation cible/sonde. Les cellules dans lesquelles l'ADN cible marqué s'est hybridé par complémentarité à une sonde s'allument sous l'excitation d'un laser (en rouge sur le schéma). Images adaptées d'Affymetrix.

Il est à noter que la taille de 25 mer des sondes des puces GeneChip est un compromis entre la sensibilité de détection et la spécificité de capture. En effet, la sensibilité augmente avec la taille de la sonde mais une augmentation de taille entraine avec elle une augmentation du risque de réactions croisées. D'une manière générale, les sondes oligonucléotidiques des puces à ADN ont des tailles comprises entre 25 mer et 70 mer, déterminées par la nature de la matrice d'accroche et de l'environnement réactionnel. Par exemple, dans le cas des puces GeneChip, l'utilisation d'un espasseur de nature aliphatique sur lequel sont greffées les sondes s'avère être une adaptation cruciale qui, en plaçant la réaction d'hybridation en phase semi liquide, serait à l'origine d'un gain de sensibilité d'un facteur 150 (Shchepinov et al. 1997). Il est à noter également que les oligonucléotides ne peuvent pas être greffés directement sur la surface d'un support solide, et que cette étape requiert une fonctionnalisation chimique et parfois l'ajout de groupements, silanes par exemple, pour initialiser la croissance de la chaîne oligonucléotidique (Pease et al. 1994).

I.3.1.2.5. Critères de définition des sondes et aspects de nomenclature

Les puces GeneChip ont pour caractéristique un double niveau d'analyse de l'expression d'un transcrit donné (Figure I-17). D'une part, la définition d'un ensemble de sondes (ou probeset) réparties sur le transcrit permet de moyenner les expressions associées aux sondes individuelles et ainsi de réduire les effets de variations liés aux affinités de capture. D'une manière générale, les sondes des puces GeneChip sont réparties sur les 600 premières bases de la région 3' du transcrit, ce qui tient à une habitude historique prise pour contrecarrer les biais 3' apportés par les premières méthodes d'amplifications transcriptionnelles. D'autre part, chaque sonde est déclinée en paire, comportant une séquence Perfect Match (PM) et une séquence Mismatch (MM) qui intègre une mutation en position centrale. En théorie, les sondes MM permettent une estimation du bruit du fond et des réactions d'hybridations croisées.

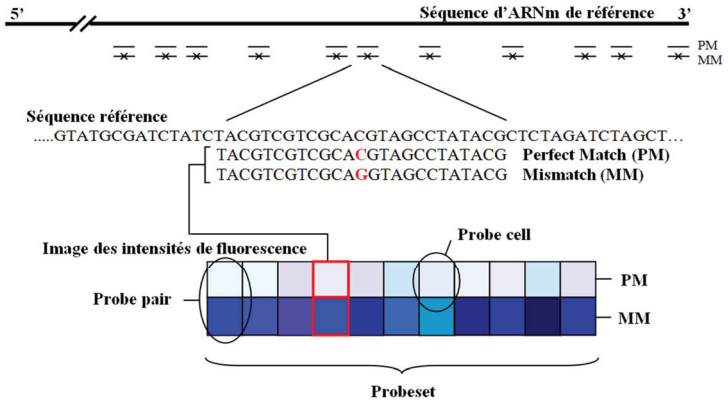


Figure I-17 Définition des sondes PP et MM des puces GeneChip. A partir d'une séquence d'ARNm, une paire de sondes est définie pour analyser l'expression de chaque transcrit. Adapté d'Affymetrix.

Différentes banques de séquences sont utilisées pour alimenter la composition d'une puce. Dans le cas de la puce HG-U133 et de ses améliorations successives, les apports primaires proviennent de GeneBank, dbEST et RefSeq et sont structurés en groupes de séquences à partir des données de la base UniGene. D'autres sources de structuration se sont ajoutées pour le développement de la version HG-U133-PLUS2, comme les assemblages du génome humain de l'université Santa Cruz (Californie) ou des données issues du NCBI (Affymetrix 2003). La méthodologie de définition des sondes, en résumé, utilise plusieurs modèles de régression linéaire qui tiennent compte de facteurs thermodynamiques afin d'évaluer la performance d'hybridation. Des indicateurs de stabilité, associés à des contraintes d'espacement le long du transcrit cible, sont la base d'une sélection de 11 sondes qui entrent dans la composition d'un probeset, destiné à la détection d'une séquence génique donnée. La puce HG-U133-PLUS2 contient un peu plus de 54 000 probesets (les différentes forêts de sondes entrant dans la composition d'un même probeset sont réparties à des endroits isolés de la puce) pouvant mesurer l'expression de 38 500 gènes (associés à environ 47 000 transcrits différents). Cependant tous les probesets de la puce n'ont pas le même niveau de spécificité vis-à-vis de leur cible. Un système de nomenclature basé sur quatre niveaux de qualification renseigne sur la nature des captures (Figure I-18).

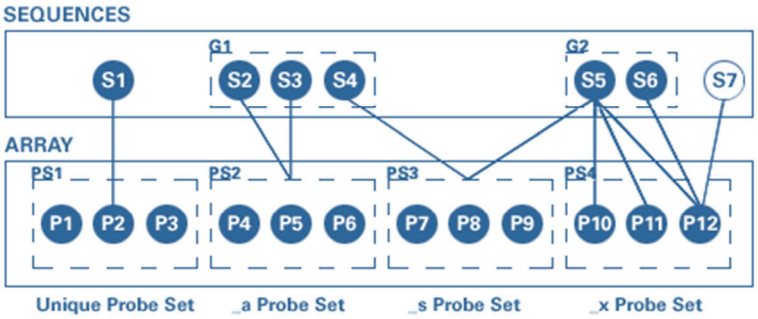


Figure I-18 Système de nomenclature des probesets de la puce HG-U133-PLUS2. Ensemble des relations possibles entre les séquences du génome et leurs probesets correspondants sur la puce HG-U133-PLUS2. G : une série de séquences appartenant à la même famille génique. S : une séquence du génome. P : une sonde de la puce. PS : un probeset de la puce. D'après Affymetrix.

La condition idéale de détection est l'association unique et spécifique d'un probeset avec sa cible (cas PS1). Des suffixes sont ajoutés au nom du probeset dans les cas de figures qui s'écartent de la situation idéale. Ainsi, lorsque que plusieurs séquences d'un ou plusieurs clusters de gènes (dans la plupart des cas des gènes homologues) peuvent être détectés par un même probeset, les suffixes _a ou _s sont utilisés, respectivement. Le suffixe _x est attribué aux probesets dont au moins une sonde croise avec deux séquences identiques appartenant à des gènes différents, rendant en conséquence l'interprétation des signaux plus incertaine[5]. Il est à noter qu'à ce problème de spécificité des sondes

[5] Le lecteur pourra se référer au site GeneAnnot, dont l'objectif est d'établir les relations possibles entre un probeset et un gène pour les puces HG-U95, HG-U133 et HG-U133-PLUS2.

s'ajoute celui de l'annotation des séquences. En effet, une fraction importante des probesets de la puce HG-U133-PLUS2 n'est associée à aucune annotation. A ce sujet et à titre d'exemple, un travail de ré-annotation de la puce HG-U133-PLUS2 a conduit un laboratoire américain à ré-identifier environ 37 % des sondes et en redéfinir ainsi les cibles réelles, encourageant les utilisateurs à travers le monde à bien peser les implications fonctionnelles issues des approches de puces GeneChip (Harbig et al. 2005). Un point particulièrement marquant de ce travail a notamment été de faire le constat qu'une petite fraction des sondes de la puce HG-U133-PLUS2 ne pouvaient être attribuée à aucune séquence d'ARNm humain connue, ce qui soulève des interrogations légitimes sur la pertinence et les limites des critères de définition des puces commerciales.

I.3.2. Méthodes spécifiques à l'analyse du transcriptome HERV

L'étude du transcriptome HERV a essentiellement reposé ces dix dernières années sur l'utilisation de méthodes basées tout ou en partie sur la RT-PCR quantitative. Dans ce paragraphe, nous dresserons un état des méthodologies de référence, en insistant sur le périmètre de chaque technique et les limites qui en découlent.

I.3.2.1. Approches basées exclusivement sur la RT-PCR

On peut distinguer deux courants au sein des études d'expression de séquences HERV par RT-PCR quantitative. L'un focalise sur des loci de références au sein du génome, choisis souvent pour la conservation des trames de lecture de leurs régions fonctionnelles, et adresse la question de l'expression sous l'angle d'une approche locus-candidat. Ainsi, l'équipe de Thierry Heidmann met au point en 2003 des systèmes de détection pour les 16 séquences d'enveloppes HERV qui contiennent un cadre de lecture ouvert (de Parseval et al. 2003). D'autres équipes appliqueront des approches comparables (Ahn and Kim 2009). Une deuxième école met à profit la conservation évolutive des séquences *pol* des rétrovirus (Xiong and Eickbush 1990) pour développer des amorces de PCR dégénérées. Le groupe suédois de Jonas Blomberg a ainsi étudié des comportements d'expression globaux au niveau des familles, pour les séquences HERV apparentées aux gammaretrovirus (HERV-H, HERV-W, HERV-E, HERV-I) (Forsman et al. 2005) et aux bétarétrovirus (HML) (Muradrasoli et al. 2006). Toutes complémentaires qu'elles soient, ces deux approches sont appliquées *in fine* à l'étude du transcriptome HERV sur un panel de tissus dans le but d'identifier des règles de tropisme d'expression. L'approche locus-candidat, en utilisant le meilleur représentant d'une famille, espère tirer une information qui puisse être extrapolable dans une certaine mesure ; l'approche par famille dégrossit des tendances à partir desquelles il restera à identifier les séquences contributrices individuelles.

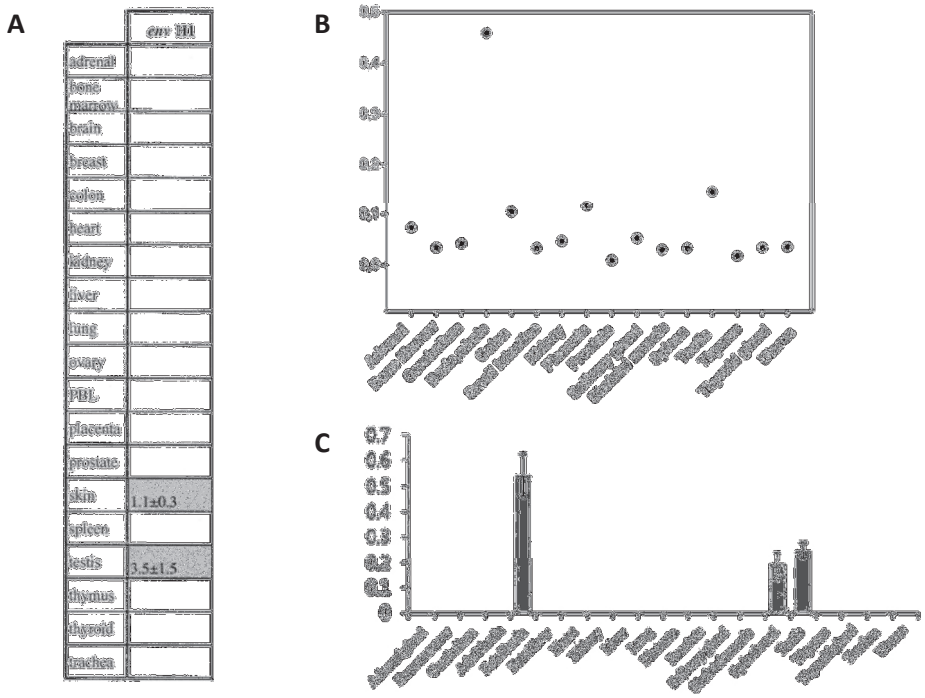


Figure I-19 Tropisme d'expression d'éléments HERV-H évalué par RT-PCR sur des panels de tissus sains. Le même locus de la famille HERV-H (AJ289709 ; 2q24.3) a servi de référence pour la mise au point d'un système de détection par RT-PCR par différentes équipes. (A) Détection du transcrit d'enveloppe du locus, d'après (de Parseval et al. 2003). (B) Utilisation de la région *pol* pour étudier la famille HERV-H dans son ensemble, d'après (Forsman et al. 2005). (C) Détection du transcrit d'enveloppe du locus, d'après (Ahn and Kim 2009).

La figure I-19 illustre les différentes approches de détection basées sur la RT-PCR. Un même locus de la famille HERV-H est à la base de la conception (i) d'amorces dans la séquence d'enveloppe pour les approches locus-candidat (A et C), et (ii) d'amorces dans la séquence *pol* pour l'étude de groupes de séquences apparentées (B). Les auteurs en tirent des règles d'expression sur des panels de tissus sains, mettant de manière complémentaire en évidence un léger tropisme d'expression, d'au moins un locus précis, dans le testis. Alternativement, alors qu'une expression de la famille HERV-H dans le cerveau adulte a été montrée par l'approche par famille, les deux expériences locus-spécifique tendent à confirmer que la source d'expression ne provient pas du locus HERV étudié.

I.3.2.2. RT-PCR couplée à une puce à ADN rétrovirale

En 2000, le journal Aids Reasearch and Human Retroviruses rapporte une technique de détection destinée à l'identification d'un large spectre de transcrits rétroviraux dans des échantillons biologiques (Seifarth et al. 2000). Le procédé utilise un couplage d'une étape d'amplification non quantitative par RT-PCR qui reconnait les motifs de RT les plus conservés au sein des génomes rétroviraux, et d'une d'hybridation en dot-blot inversé sur membrane mettant en jeu des sondes de capture spécifiques d'une famille de rétrovirus. Notons qu'une étape de clonage/séquençage subséquente est nécessaire pour quantifier les expressions observées. Des familles (H)ERV (e.g. : HML-1 à 6, HERV-W, HERV-H pour l'homme, mais aussi MMTV, PERV, BaEV, etc.) accompagnées d'un

ensemble de rétrovirus exogènes humains (HTLV-1/2, HIV-1/2, Foamy virus) sont ainsi greffés à la membrane de détection. Avec une sensibilité de détection permettant de voir 10 transcrits PERV dans un échantillon biologique, cette méthode ouvre notamment la voie à des applications de recherche de contaminants ERV post xénogreffes. Les contraintes expérimentales (liées à la radioactivité des sondes essentiellement) font rapidement évoluer la technique de détection vers l'utilisation d'un support de verre et l'usage de molécules de marquage chimiques (Seifarth et al. 2003) (Figure I-20). Dans le même temps, des projets d'étude du transcriptome HERV sur des panels de tissus se développent.

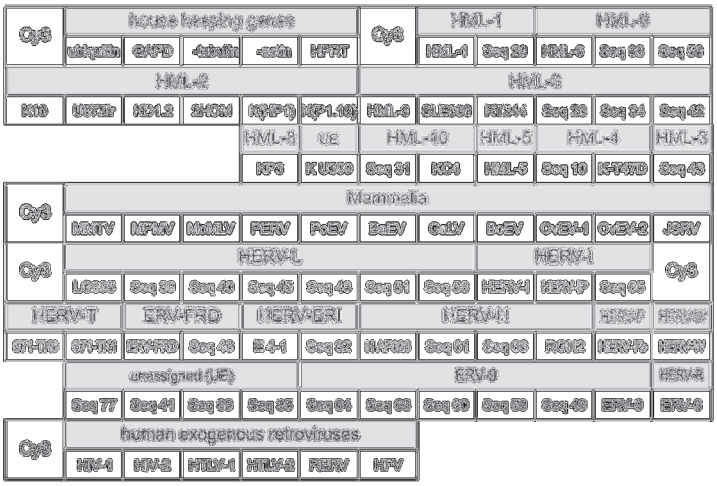


Figure I-20 Représentation schématique de la puce à ADN dédiée aux séquences rétrovirales développée par l'équipe de C. Leib-Mösch. Des oligonucléotides de capture spécifiques de familles (H)ERV et de rétrovirus exogènes humains sont déposés suivant une grille sur un support de lame de verre. L'hybridation de produits d'amplification marqués par Cy3, obtenus à l'aide d'amorces *pol* universelles, révèle le contenu en transcrits rétroviraux de l'échantillon testé. D'après (Seifarth et al. 2003).

C'est ainsi qu'une étude sur 19 tissus sains humains a décrit des règles de tropisme pour 20 familles HERV (Seifarth et al. 2005), allant d'une expression ubiquitaire à des évènements de transcription plus rares. D'une manière générale, les familles HERV de classe I et II apparaissent plus actives que les familles HERV de classe III. En particulier, la famille ERV9 est exprimée dans tous les tissus testés, les familles HERV-E, HERV-F et HERV-W sont exprimées dans tous les tissus à l'exception du rectum et de l'ovaire et la famille HERV-FDR n'est silencieuse que dans les cellules mononucléaires sanguines, l'estomac et la prostate. La famille HERV-H n'a été détectée que dans le rectum, l'ovaire et l'utérus, ce qui contraste avec les résultats obtenus par les approches de RT-PCR développées plus haut. Ceci souligne une limite inhérente à l'utilisation de séquences consensus *pol* dans les étapes d'amplification pour l'étude de familles d'éléments rétroviraux. La famille HERV-H, par exemple, peut avoir des promoteurs actifs dans plusieurs types cellulaires (Schon et al. 2009; Schön et al. 2001) mais environ 90 % des loci HERV-H du génome ont une séquence *pol* tronquée. Malgré cette restriction, la puce à ADN développée par l'équipe de Christine Leib-Mösch est régulièrement utilisée et a notamment permis chez l'homme (i) d'associer une expression

différentielle de sous-groupes de la famille HML-2 à des désordres neuropsychiatriques (Frank et al. 2005) puis d'établir un lien entre la prise de médicaments antipsychotiques et la réactivation des familles HERV-W et ERV9 (Diem et al. 2012) (ii) d'observer une augmentation générale de transcription des familles HERV dans des cellules neuro épithéliales infectées par l'agent infectieux de la toxoplasmose (Frank et al. 2006) (iii) de décrire une surexpression des familles ERV-T, HERV-E, HERV-F et HML-6 dans certains groupes de patients atteints de cancer du sein (Frank et al. 2008) et (iv) d'établir un profil d'expression HERV type des cellules rénales (Haupt et al. 2011). Chez l'animal, la puce à ADN rétrovirale a trouvé des applications au travers de l'étude des effets d'une infection par la protéine prion sur le répertoire ERV de la souris (Stengel et al. 2006a) et de la description de l'activité physiologique des HERV chez les singes de l'ancien monde (Stengel et al. 2006b), ce dernier travail soulignant de manière intéressante que malgré l'existence d'un répertoire commun avec l'être humain, ce sont des familles HERV différentes qui s'expriment dans le cerveau des hommes et des grands singes.

I.3.2.3. MD-PCR OLISA

Par une méthodologie approchant les travaux de Christine Leib-Mösch, notre laboratoire a développé en 2006 un procédé quantitatif basé sur une amplification RT-PCR multiplexée et dégénérée (MD-PCR) des régions *pol* rétrovirales, couplée à la détection de séquences consensus sur un format Oligo Sorbant Array (OLISA) (Figure I-21). Trois paramètres ont fait l'objet d'une optimisation dans le but d'obtenir des efficacités d'amplification équivalentes d'une cible à l'autre et d'assurer une relation d'effet dose : le niveau de dégénérescence des amorces, la concentration relative de chaque amorce et la concentration totale d'amorces dans le milieu. Ainsi, un mélange réactionnel composé de 1 240 conditions de dégénérescence a permis d'étudier l'expression de neuf familles de rétrovirus endogènes humains. La MD-PCR OLISA a été utilisée dans le cadre d'études comparatives exploitant des modèles *in vitro* de différenciation cellulaire ainsi que des prélèvements de tissus sains et cancéreux. Cela a permis d'établir des profils d'expressions tumorales, en montrant en particulier une surexpression de la famille HML-2 dans les cancers du sein, de la famille HERV-W dans le séminome, de la famille HERV-E dans le cancer de la prostate et de la famille HERV-H dans les tumeurs du côlon (Pichon et al. 2006). Sans pouvoir identifier les loci individuels à l'origine de l'expression, cette approche quantitative en multiplexage fournit des règles de conception d'amorces et permet d'envisager des applications de phénotypage tumoral.

A

HERV	Forward primer	Dg level	Final concentration (µM)	Reverse primer	Dg level	Final concentration (µM)
HML-2	TRGAAAGTGTTRCCTCARGGA	8	0.51	CACAYAAAATATYATCAAYAYA	16	1.02
HML-4	TGGAAAGTCCTACCACAAGGC	1	0.03	TGCAGAGGAGATCATCCATGTA	1	0.03
HML-5	TRGAAAGTRCTTCCTSAARGR	32	1.02	CYAGTARRATATCATCCATAAA	8	0.51
HERV-H	TGGRCTGTRCTGCYRCAAGGY	32	2.03	AAAGWAGAAGSTCRTCAAWATA	16	1.02
ERV-9	GTCTTGCCHCAAGGGTTT	3	0.13	AGTAAATCATCCACATAYTGAAG	2	0.13
HERV-W	TGGACTGTTTTACCCCAAGGG	1	0.13	AAgTAAATYATCCMYRTACYRA	64	1.02
HERV-E4.1	TGGACCsRGCYTCCCCAARGG	16	1.02	CCAGCARRAGGTCATCaAYSTA	16	1.02
HERV-R	TGGACTAGTCTCCCACAAGGG	1	0.06	CCAAAAGAAGGTTGTCTATGTA	1	0.06
HERV-L	TTTACTGTCCTACYTCAGGGR	4	0.25	TCARCATAATGYCATCAATGTA	4	0.51

B

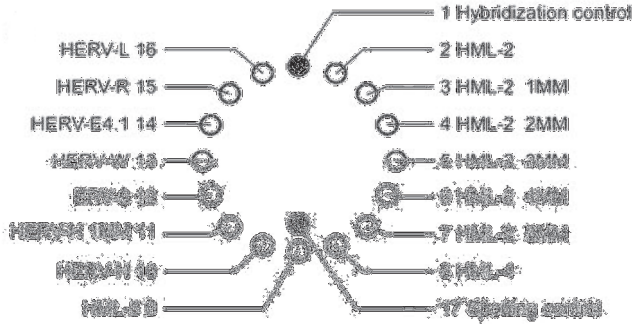


Figure I-21 Eléments de conception du système de MD-PCR OLISA. (A) Amorces d'amplification dégénérées et composition du mélange d'amorces. Dg level : niveau de dégénérescence. 's', 'g' et 'a' représentent respectivement 75%, 78% et 67% des bases observées à cette position sur la population entière. (B) Plan de la plaque de détection. La puce OLISA inclut un dépôt de contrôle (spotting control) et un contrôle d'amplification (hybridization control). Dans le cas des familles HERV-K HML-2 et HERV-H, des déclinaisons de la séquence de capture ont été réalisées (MM). D'après (Pichon et al. 2006).

I.3.2.4. RT-PCR couplée au séquençage

Au lendemain de la phase pilote du projet d'encyclopédie des éléments ADN (ENCODE Project 2007), l'étude du transcriptome prend un tour nouveau. Les résultats présentés, autant fascinants que provoquants, et décrivant une transcription permissive du génome, l'existence de transcrits non-codants dans des territoires chromatiniens supposés silencieux ou encore aboutissant à l'identification de nouveaux sites d'initiation de la transcription, vont conduire à une remise en question progressive de la notion de gène. Dans ce contexte, l'équipe allemande à l'origine de la puce à ADN rétrovirale amorce un virage dans son approche d'étude du transcriptome HERV pour s'intéresser au comportement de l'ensemble des séquences de la famille HML-2 au niveau individuel de chaque locus (Flockerzi et al. 2008). Pour y parvenir, une RT-PCR spécifique des régions *gag* et *env* de la famille HML-2 est réalisée sur un échantillon biologique, puis le produit d'ADNc est cloné et séquencé. La séquence ainsi obtenue est alignée sur le génome, conduisant à l'identification du locus à l'origine de l'expression. En suivant cette démarche, 49 tissus humains, regroupant essentiellement des échantillons tumoraux, ont été inclus dans ce qui constitue la première étude du transcriptome de la famille HML-2 au niveau locus, à grande l'échelle. En substance, plus de 1 500 clones séquencés

ont conduit à l'identification de 23 provirus HML-2 actifs tout en apportant une information quantitative sur les niveaux d'expressions individuels. La transcription du locus 5q33.3 a ainsi été associée aux échantillons de tumeurs cérébrales, et un niveau d'expression important dans les tumeurs testiculaires atrophiques et orchitiques a été mis en évidence pour trois loci, 11q23.2, 21q21.1 et 22q11.21, en accord avec certains résultats obtenus par des approches locus-candidats. Si une telle technique parvient bien à établir des profils d'expression au niveau des séquences HERV individuelles, la mise en œuvre plus que laborieuse de la démarche, compliquée de surcroit par des problèmes de recombinaisons *ex vivo* qui peuvent avoir lieu pendant l'étape de RT-PCR (Flockerzi et al. 2007), a amené les auteurs de ce travail à conclure sur un appel à projets collaboratifs dédiés à d'étude du transcriptome HERV.

I.3.2.5. GREM, méthode d'amplification pour l'identification de LTRs promotrices

Les techniques d'amplifications basées sur les régions *gag*, *pol* ou *env* des provirus ne disent rien de l'activité fonctionnelles des LTRs. Dans le but d'identifier des séquences LTRs promotrices par une approche à grande échelle, un procédé, nommée GREM, a essayé de combiner les avantages des méthodes d'amplification rapides des extrémités d'ADNc (RACE) avec les techniques d'hybridation d'acides nucléiques (Buzdin et al. 2006b) (Figure I-22). La première étape du procédé est la constitution d'une banque d'ADNc par l'utilisation d'amorces oligo-dT, à partir des transcrits ARNm d'un échantillon. Des adaptateurs de clonage sont ajoutés aux extrémités 5' des ADNc double brin ainsi générés avant de procéder à une étape de digestion par Alu I. Le choix de cette enzyme est motivé par l'absence de site de restriction Alu I dans les séquences LTRs, en sorte que les séquences double brin d'ADNc correspondant à des sites d'initiation de transcription dans des LTRs sont préservées de la dégradation. Parallèlement, une étape d'amplification génomique des LTRs et de leur domaine 3' flanquant est réalisée à l'aide d'amorces de PCR spécifiques. Un traitement à l'exonucléase III permet de libérer les extrémités 3' des brins des produits d'amplification. Pour finir, la mise en contact des brins d'ADN génomiques libres à leurs extrémités 3' avec les cibles d'ADNc, après une étape de dissociation des structures bi caténaires par fusion, permet une hybridation par complémentarité. En théorie, seules les structures d'hybridation mettant en jeu des LTRs promotrices (ou des lectures traversantes antisense) seront clonées puis séquencées. L'alignement sur le génome permet alors d'identifier, locus par locus, des LTRs promotrices.

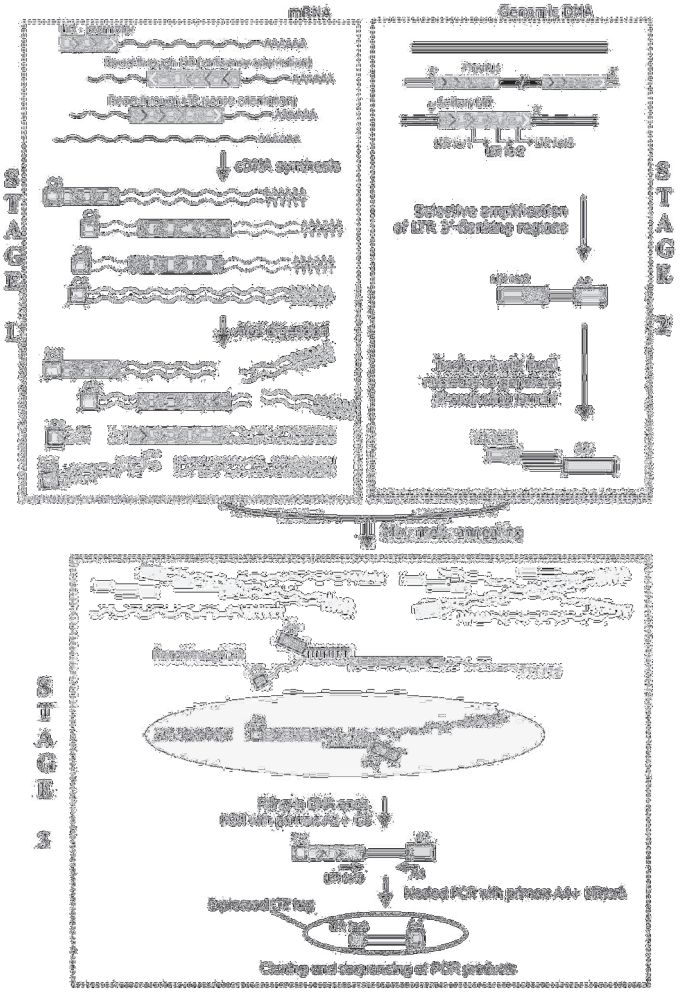


Figure I-22 Représentation schématique de la technique GREM. La première étape (stage 1) consiste en une reverse transcription des transcrits ARNm en ADNc à l'aide d'amorces oligo-dT. Des adaptateurs de clonage sont ajoutés en 5' des séquences double brin (CS) avant une digestion par Alu I (absence de site de restriction dans les LTRs). Parallèlement (stage 2) une amplification par PCR est réalisée sur la région 3' flanquante des LTRs génomiques, à l'aide d'amorces et d'adaptateurs spécifiques. Après une étape additionnelle de PCR nichée, un traitement à l'exonucléase III permet d'obtenir un produit aux extrémités 3' libres. La dernière étape (stage 3) consiste en l'hybridation des cibles d'ADNc avec les séquences génomiques. Les produits d'hybridation double brin sont alors clonés et séquencés, ce qui conduit à l'identification locus spécifique des séquences LTRs actives du génome.

Cette technique a été appliquée aux LTRs de la famille HML-2 et a permis à ses auteurs d'identifier 54 LTRs promotrices dans le parenchyme testiculaire, puis d'étendre les observations à un contexte pathologique pour finalement conclure qu'au moins 50 % des LTRs de la famille HML-2 servent de promoteurs *in vivo* (Buzdin et al. 2006a). GREM n'a cependant pas été appliquée à l'étude des autres familles HERV, en raison sans doute des contraintes expérimentales et probablement à cause des difficultés à définir des amorces de PCR pour les LTRs qui s'inscrivent dans un contexte de provirus. D'autres limites réduisent également le champ d'application de GREM. Premièrement, l'utilisation d'oligo-dT pour l'obtention de banques d'ADNc exclut de l'analyse les transcrits qui n'ont pas de queue polyA, et qui constituent pourtant une proportion importante du transcriptome humain (Bentwich et al. 2005). Deuxièmement, les phénomènes d'épissages qui utilisent des sites donneurs aux extrémités 3' des LTRs mettent en échec la logique d'hybridation basée sur les séquences génomiques 3' flanquantes. Troisièmement, un risque de faux positifs existe dans les cas d'une lecture traversante initiée à une faible distance en amont des LTRs. Enfin, si cette méthode a le mérite de s'attaquer à l'identification de séquence promotrices LTRs, elle ne considère absolument pas le potentiel fonctionnel de terminaison de transcription des LTRs, potentiel *a priori* équivalent à celui de l'initiation de transcription.

I.3.2.6. Apports et limites des banques d'ESTs

Les ESTs, pour Expressed Sequence Tags, sont des petites séquences d'ADN, généralement entre 200 et 500 nucléotides, générées par le séquençage de clones d'ADNc. L'idée générale est d'utiliser ces petits fragments de séquences, qui représentent un évènement de transcription dans un contexte cellulaire donné, pour remonter au locus à l'origine de l'expression par une procédure d'alignement de séquences. La mise en commun des ESTs obtenus à partir des multiples banques d'ADNc mondiales permet notamment de dégager des règles de tropisme d'expression. La base d'ESTs humains du NCBI dbEST, créée en 1992, regroupe aujourd'hui plus de 8,6 millions[6] de tags, ce qui témoigne de l'engouement né autour de cette approche d'étude. Pourtant, des limites contraignantes sont associées au travail à partir de banques d'ESTs, que l'on s'intéresse à l'étude du transcriptome en général ou à l'expression des HERV en particulier. Les évènements d'épissages constituent probablement le premier niveau de difficulté dans l'association d'une séquence EST avec son origine génomique. A cela s'ajoute un système d'annotations parfois défaillant, ce qui complique les étapes d'analyse. Plus particulièrement problématique dans le cadre de l'étude de l'ADN non codant, les erreurs de séquençage des banques d'ADNc, de l'ordre de 3 %, induisent une variabilité purement technique qui dépasse le niveau de polymorphisme de la population humaine, estimé à 1 pour 0,31 kb pour les éléments répétés (Nickerson et al. 1998). Malgré ces handicaps, des équipes ont cherché à identifier des loci HERV individuels à l'origine d'expressions tissu-spécifiques en explorant les banques d'ESTs. Ainsi, l'expression de 32 loci HML-2 a pu être associée à un ensemble de tissus normaux et cancéreux (Stauffer et al. 2004) et l'utilisation de modèles analytiques de type Hidden Markov a mis en évidence quelques centaines d'ESTs pertinents, par une approche purement bioinformatique (Oja et al. 2007). La concordance des résultats, évaluée par Flockerzi dans son étude par couplage RT-PCR et séquençage (Flockerzi et al. 2008), reste pourtant trop faible pour que les ESTs apparaissent comme un matériel pertinent pour l'étude du transcriptome HERV. En particulier, au laboratoire, l'approche par ESTs a échoué à identifier la LTR 3' du locus ERVWE1 comme signal polyA, alors que l'ADNc contenant le transcrit complet et polyadénylé de la Syncytin-1 a été isolé (Blond et al. 1999).

[6] Au 1er juillet 2012. Voir http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html

I.4. Problématiques de la cancérologie clinique

I.4.1. Positionnements et besoins de marqueurs de cancers

I.4.1.1. Notions cliniques usuelles

Une vue simplifiée des problématiques de santé publique liées aux cancers s'articule autour (i) du dépistage, qui est la démarche visant à détecter, au plus tôt et en l'absence de symptômes, des lésions susceptibles d'être cancéreuses ou d'évoluer vers un cancer. Il peut s'adresser à une population ciblée (les individus atteints d'une maladie pulmonaire chronique pour le dépistage du cancer du poumon, par exemple) ou viser un public plus large (inciter toutes les femmes à partir de 55 ans à effectuer une mammographie tous les deux ans), (ii) du diagnostic, qui est le processus selon lequel une maladie est identifiée par ses signes et symptômes et s'adresse donc à une population présentant des manifestations cliniques de la pathologie et (iii) du pronostic qui, une fois le diagnostic posé, évalue les conséquences probables d'une maladie et son évolution dans le temps.

La prédiction individuelle du devenir du patient peut alors reposer sur (iv) la stratification, qui vise à identifier des groupes de patients partageant des mêmes caractéristiques biologiques, organisées souvent d'un point de vue moléculaire ou biochimique. Cela peut orienter vers (v) la surveillance active, qui est le choix thérapeutique, une fois un pronostic personnalisé posé, consistant à ne pas intervenir médicalement sur des tumeurs indolentes à évolution lente mais à suivre régulièrement le patient (l'exemple le mieux documenté sur ce point est le cancer de la prostate). Alternativement, la notion de (vi) réponse au traitement permet, en fonction des prédispositions d'un patient, d'orienter un choix thérapeutique mais également de suivre son efficacité au cours du temps.

Enfin, des programmes visant à identifier la nature du foyer primaire d'une tumeur à partir de ses métastases existent et ont aussi pour ambition de permettre d'adapter les protocoles de traitement (Erlander et al. 2011; Ma et al. 2006). Selon le type de tumeur étudiée (solide, liquide), ces différentes questions peuvent être adressées par l'imagerie, par des prélèvements de tissus (biopsies, chirurgie) ou en cherchant à obtenir une information circulante, par exemple dans la circulation sanguine ou dans les urines.

I.4.1.2. Etat de la recherche de biomarqueurs de cancers

La notion de biomarqueur est définie par les National Institutes of Health comme toute caractéristique pouvant être objectivement mesurée et évaluée comme indicateur d'un processus biologique normal ou pathologique, ou comme réponse pharmacologique à une intervention thérapeutique (Biomarkers Definitions Working Group 2001). L'idéal recherché en matière de

biomarqueurs en oncologie sont des tests non invasifs qui puissent être utilisés pour diagnostiquer un foyer tumoral à un stade précoce, permettre d'établir une classification des tumeurs pour déterminer la meilleure option curative, suivre la réponse au traitement et évaluer la progression ou la récurrence du cancer (Rhea and Molinaro 2011). Si les biomarqueurs peuvent être de toute nature biologique (ADN, ARN, protéines, glucides), les protéines sont les molécules les plus recherchées en raison de la compréhension des voies de signalisations cellulaires que l'on peut y associer (Makawita and Diamandis 2010). Durant les dix dernières années, les technologies de protéomique et de puces à ADN ont contribué à générer plus de 150 000 articles scientifiques revendiquant chacun plusieurs dizaines de marqueurs, mais une centaine de ces molécules, seulement, ont par la suite été validées en clinique (Poste 2011) (Figure I-23). A ce jour, une petite vingtaine de biomarqueurs de cancers sont approuvés par l'agence fédérale américaine des produits alimentaires et médicamenteux (FDA) (Ludwig and Weinstein 2005; Rhea and Molinaro 2011).

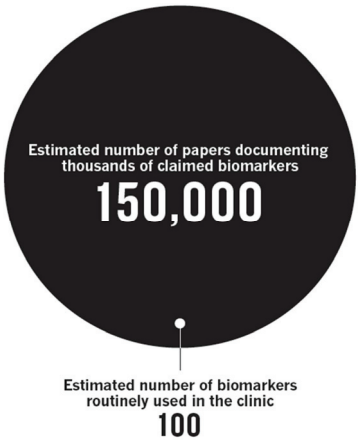


Figure I-23 'A drop in the ocean'. Estimation du nombre de publications scientifiques ayant identifié des biomarqueurs potentiels, remis en perspective avec le nombre de marqueurs qui ont passé les phases d'évaluations cliniques. D'après (Poste 2011)

Pour que la valeur clinique d'un biomarqueur soit reconnue, différentes considérations doivent être prises en compte. Tout d'abord, si l'on vise un diagnostic précoce de cancer, le marqueur potentiel doit être une molécule produite par une petite tumeur asymptomatique et, par exemple, être libéré dans la circulation à une concentration détectable (Diamandis 2010). Jusqu'à maintenant, ce genre de molécules a échappé aux chercheurs. Les marqueurs utilisés permettent, en majorité, une détection de la tumeur à un stade déjà avancé, et sont utiles pour établir des stratifications tumorales ou dans le cadre d'une surveillance active (Tableau I-3). Puis, dans la perspective de différentier un prélèvement sain d'un échantillon tumoral, le marqueur doit autant que possible présenter une forte spécificité tissulaire afin que sa détection n'entre pas en compétition avec des molécules présentes physiologiquement dans d'autres organes. A cela il est bien-sûr souhaitable que l'état de la molécule d'intérêt ne soit pas non plus affecté dans les tissus non-cancéreux.

Cancer	Marqueur	Nature	Prélèvement	Usage
Testicules	α-Fœtoprotéine	Glycoprotéine	Sérum	Stratification
	hCG	Glycoprotéine	Sérum	Stratification
Pancréas	CA19-9	Glucide	Sérum	Surveillance
Ovaire	CA125	Glycoprotéine	Sérum	Surveillance
Col de l'utérus	Pap smear	Frottis cellulaire	Tissu	Dépistage
Colon	CEA	Protéine	Sérum	Surveillance
	EGFR	Protéine	Tissu	Choix thérapeutique
Gastro-intestinal	CD117	Protéine	Tissu	Diagnostic et choix thérapeutique
Thyroïde	Thyroglobuline	Protéine	Sérum	Surveillance
Prostate	PSA	Protéine (totale)	Sérum	Dépistage et surveillance
		Protéine (complexe)	Sérum	Dépistage et surveillance
		Protéine (libre)	Sérum	Diagnostic
	PCA3	ARN non codant	Urine	Diagnostic
Sein	CA15-3	Glycoprotéine	Sérum	Surveillance
	CA27-29	Glycoprotéine	Sérum	Surveillance
	Cytokératines	Protéine	Tissu	Pronostique
	Récepteurs aux œstrogènes et à la progestérone	Protéine	Tissu	Choix thérapeutique
	HER2/NEU	Protéine	Tissu	Pronostic et choix thérapeutique
		Protéine	Sérum	Surveillance
		ADN	Tissu	Pronostic et choix thérapeutique
Vessie	Chr. 3,7,9 et 17	ADN	Urine	Dépistage et surveillance
	NMP22	Protéine	Urine	Dépistage et surveillance
	Fibrine/FDP	Protéine	Urine	Surveillance
	BTA	Protéine	Urine	Surveillance
	CEA and mucine	Protéine	Urine	Surveillance

Tableau I-3 Marqueurs de cancers approuvés par la FDA. BTA : bladder tumour-associated antigen ; CA : cancer antigen ; CEA : carcinoembryonic antigen ; FDP : fibrin degradation protein ; NMP22 : nuclear matrix protein 22 ; hCG : human chorionic gonadotropin-β ; EFGR : epidermal frowth factor receptor ; PSA : prostate-specific antigen ; PCA3 : prostate cancer antigen 3 ; BTA : bladder tumor antigen. Adapté de (Rhea and Molinaro 2011) et (Ludwig and Weinstein 2005).

I.4.1.3. Méthodes pour la recherche et la validation de biomarqueurs

En 2002, le National Cancer Institute et son réseau destiné à la recherche et au développement de nouveaux tests biologiques pour la détection précoce de cancers ont développé une approche en cinq phases dans le but de rationaliser la recherche et l'évaluation de biomarqueurs. Ces recommandations ont par la suite été largement reprises et parfois légèrement adaptées (Bensalah et al. 2007; Ptolemy and Rifai 2010; Rifai et al. 2006), pour finalement aboutir au consensus méthodologique suivant : tout d'abord, une phase préclinique doit explorer des modèles *in vitro* ou animaux dans le but de tester une hypothèse biologique. Cette phase n'est possible que si des modèles adaptés existent en lien avec la problématique d'étude. Suite à cela commence la phase 0 de recherche à proprement parler. Il s'agit d'une étape de mise au point méthodologique sur échantillons cliniques, sans considération sur la valeur potentielle du marqueur. Durant la phase 1

qui suit, le marqueur, ou la stratégie de recherche de nouveaux marqueurs, est évalué sur un petit groupe de patients dans le but de valider l'aptitude à discriminer des classes cliniques. Cette étape inclut une phase de découverte et d'optimisation des protocoles, en définissant notamment le marqueur ou la combinaison de marqueurs candidats avec précision et en mettant au point des règles de prédiction pour délimiter la zone de validité des résultats (cela passe par exemple par l'estimation d'effets populationnels). A ce stade, le test n'utilise pas nécessairement le format technologique final. En phase 2, une validation sur des échantillons indépendants doit confirmer le potentiel du test et établir ses valeurs de référence de reproductibilité et de robustesse. Une fois la fiabilité analytique connue, le niveau de bruit de fond dans la population cible, les variations inter individuelles, ainsi que les fluctuations au sein d'un même individu au cours du temps sont estimés. Vient alors la phase 3, qui consiste à déterminer l'efficacité du marqueur sur une large population, distincte de celle utilisée dans les étapes précédentes (cela consiste, typiquement, à mettre en place une étude multicentrique). L'objectif fondamental est de valider la sensibilité et la spécificité du test. Idéalement, cette étape doit utiliser des essais randomisés pour montrer un bénéfice par rapport aux tests de référence. Enfin, la phase 4 évalue de manière critique les résultats rapportés et interroge en particulier la limite du marqueur vis-à-vis d'autres processus biologiques et d'un nouvel ensemble de pathologies. Ce schéma méthodologique n'est pas simplement une construction intellectuelle mais fournit un cadre commun dans lequel les chercheurs et les patients doivent pouvoir appréhender les avancées dans le développement de biomarqueurs.

I.4.2. Eléments justifiant la recherche de nouveaux marqueurs du cancer de la prostate

I.4.2.1. Généralités et besoins cliniques spécifiques

Le cancer de prostate est la tumeur solide la plus fréquente chez l'homme. A 80 ans, on estime que près d'un homme sur deux est porteur d'une prostate avec des cellules qui ont subi une transformation néoplasique (Sanchez-Chapado et al. 2003). L'évolution du cancer de la prostate est très variable. Près de 90 % des patients développent des tumeurs à évolution lente qui, même après 20 ans de maladie, n'ont pas nécessairement de conséquences nuisibles pour l'individu. Dans certains cas cependant, la tumeur se développe moins tard dans la vie (avant 65 ans) et son diagnostic tardif conduit à la mort du patient. Pour les cancers non-invasifs, il existe d'excellentes options curatives par des mesures opératoires (Huland 2001) ou radiothérapeutiques (Wiegel and Hinkelbein 1998). En revanche, pour les stades avancés, il n'existe que des méthodes thérapeutiques hormonales (ablation des androgènes) avec une efficacité limitée dans le temps. Sous la pression du

milieu oncologique médical, une prise en charge de plus en plus agressive est proposée aux patients, ce qui expose un nombre grandissant d'hommes aux risques d'un sur-diagnostic et d'un sur-traitement. Plusieurs types de problématiques cliniques se posent donc : (i) peut-on trouver un marqueur diagnostique plus performant que ceux dont nous disposons aujourd'hui (PSA, PSA libre/total, PCA-3) ? En effet, 40 % à 50 % des patients sont inutilement biopsiés suite au dosage de leur PSA sanguin (voir le paragraphe suivant). (ii) Une fois le diagnostic établi, l'élément aujourd'hui le plus difficile à évaluer en l'absence de données fiables en imagerie (l'IRM ayant en 2011 une puissance diagnostique encore trop faible) est le potentiel évolutif de la tumeur. Les marqueurs biologiques attendus doivent pouvoir, utilisés seuls ou en combinaison, identifier avec certitude les rares patients qui ont une probabilité d'évolution agressive dans les années qui suivent le diagnostic. (iii) Dans certains cas de tumeurs peu agressives, une attitude de surveillance est proposée de façon expérimentale (protocole PHRC SURACAP). Il fait peu de doute que cette attitude soit bénéfique à un grand nombre de patients. La surveillance pose cependant le problème du suivi, qui repose pour l'instant sur des biopsies à répétition, tous les deux ans environ. La possibilité de disposer de marqueurs pertinents de l'évolution du cancer serait, là encore, particulièrement intéressante. (iv) Pour les individus à l'autre extrémité du spectre qui sont diagnostiqués à un stade tardif, des marqueurs de l'évolution vers le stade d'hormono-indépendance de la maladie seraient probablement bénéfiques. Il serait alors peut-être possible d'anticiper des traitements de seconde ligne plus précocement, dans l'optique de réduire les évènements intercurrents dangereux (fractures pathologiques, compression médullaire, etc.). Par ailleurs, le cancer de la prostate ayant des traitements de seconde et troisième ligne de plus en plus nombreux (Mottet et al. 2011), il est probable que leur hiérarchisation se trouverait aidée par des marqueurs d'efficacité ou de toxicité potentielle du traitement.

I.4.2.2. Dosage du PSA : de la facilité d'utiliser à mauvais escient un biomarqueur

La polémique sur l'utilisation du test PSA sanguin enfle un peu plus chaque année, au point qu'en 2010 son inventeur, le docteur Richard J. Ablin, prend ouvertement position dans les colonnes du New York Times contre l'utilisation de son test pour le dépistage de masse, dans un article intitulé 'The great prostate mistake'. Son estocade s'appuie principalement sur les conclusions des deux grandes études cliniques menées respectivement en Europe (Schroder et al. 2009) et aux Etats-Unis (Andriole et al. 2012). L'étude européenne, qui a réalisé le suivi sur neuf ans de près de 163 000 hommes, conclue qu'un dépistage de masse par le dosage du PSA amène à traiter 48 patients pour en sauver un seul. Les 47 autres, selon toute vraisemblance, ayant troqué dans l'opération leur vie sexuelle contre des problèmes d'incontinence. Le verdict de l'étude américaine est peut-être plus accablant encore en montrant que sur une période de 7 à 10 ans, le dépistage par le PSA ne réduit tout simplement pas le taux de mortalité des hommes de plus de 55 ans. De cela, le US Preventive

Service Task Force (USPSTF) a formulé de nouvelles recommandations en 2011 en se prononçant contre le dépistage du cancer de la prostate par le dosage du PSA dans des populations d'hommes sans symptôme (U.S. Preventive Services Task Force 2012). En France, la Haute Autorité de Santé (HAS) considère pour sa part dans un rapport de 2012 consultable en ligne qu'il n'existe pas de preuve de l'intérêt du dépistage du cancer de la prostate par le dosage du PSA chez les hommes sans symptôme, même chez ceux qui peuvent être considérés comme à plus haut risque. Cela étant dit, les différents auteurs s'accordent sur l'utilité du test dans un contexte de diagnostic et de suivi du traitement, en en soulignant toutefois les limites. Le dosage du PSA n'a par exemple aucune valeur prédictive de l'évolution de la maladie. Or ce qui fait défaut pour une prise de décision médicale est, entre autres, de disposer d'indications permettant de distinguer de manière précoce les formes agressives de cancer des formes non-agressives à évolution lente et sans impact sur la vie des patients. Les positionnements possibles à une recherche de marqueurs pouvant remplacer (ou améliorer) le dosage du PSA vont donc actuellement de la mise au point d'un nouveau procédé de dépistage à l'identification de marqueurs diagnostiques ou pronostiques de la maladie, comme rappelé plus haut.

I.4.2.3. Les précédents en transcriptomique

L'une des études pionnières recherchant des marqueurs transcriptomiques du cancer de la prostate s'inscrit dans un contexte pronostique, les auteurs ayant essentiellement cherché à définir une stratification des patients (Dhanasekaran et al. 2001). Des échantillons provenant de prostatectomies radicales, de tissus normaux adjacents de prostate (NAP), d'hyperplasie bénigne (BPH), de cancers localisés (PCA), de cancers métastatiques réfractaires aux hormones (MET), de lignées cellulaires (DU-145, LnCAP, PC3) et de divers contrôles ont été utilisés sur une puce à façon humaine contenant 10 000 ADNc. Le résultat a été l'identification de 213 gènes présentant une variation d'expression supérieure à 3,5 fois entre deux classes au moins. Les auteurs de ce travail concluent que l'intégration des résultats de puces et des données cliniques et pathologiques est une approche puissante pour le profilage moléculaire du cancer chez l'homme. Plusieurs études ont alors suivi ces résultats novateurs et encourageants, posant des questionnements diagnostiques ou pronostiques aux contours plus ou moins nets, désignés par exemple par 'cancer behavior' (Singh et al. 2002), 'predicting tumor aggression' (Yu et al. 2004), 'clinical outcome' (Glinsky et al. 2004) ou 'subtyping' (Lapointe et al. 2004). Ce mélange de questions est souvent basé sur l'exploitation de critères cliniques comme la classification TNM ou le score de Gleason (Figure I-24, page suivante), mais peut aussi reposer sur une approche sans *a priori* de stratification des patients. En tout cas, il y a souvent un décalage entre le type de recrutement et la question qui est posée, par exemple en abordant des aspects pronostiques sans notion de temps, et, en conséquence, il s'observe fréquemment une dérive dans les résultats présentés au regard du questionnement initial. De plus,

les listes de gènes des différentes études ont un taux de recoupement quasi nul, et ceci même dans le cas du partage d'une même cohorte de patients entre deux équipes.

Classification TNM

T : Tumeur primitive

- T0 : Absence de tumeur
- T1 : Tumeur non palpable ou non visible en imagerie
  - T1a < 5 % du tissu réséqué
  - T1b > 5 % du tissu réséqué
  - T1c : découverte par élévation du PSA et réalisation de BP
- T2 : Tumeur limitée à la prostate (apex et capsule comprise)
  - T2a : Atteinte de la moitié d'un lobe ou moins
  - T2b : Atteinte de plus de la moitié d'un lobe sans atteinte de l'autre lobe
  - T2c : Atteinte des deux lobes
- T3 : Extension au-delà de la capsule
  - T3a : Extension extra-capsulaire
  - T3b : Extension aux vésicules séminales
- T4 : Extension aux organes adjacents (col vésical, sphincter urétral, rectum, paroi pelvienne) ou tumeur fixée à la paroi pelvienne

N : Ganglions régionaux

- N0 : Absence de métastase ganglionnaire
- N1 : Atteinte ganglionnaire(s) régionale(s)

M : Métastases à distance

- M0 : Absence de métastases à distance
- M1 : Métastases à distance
  - M1a : Ganglions non régionaux
  - M1b : Os
  - M1c : Autres sites

Classification de Gleason

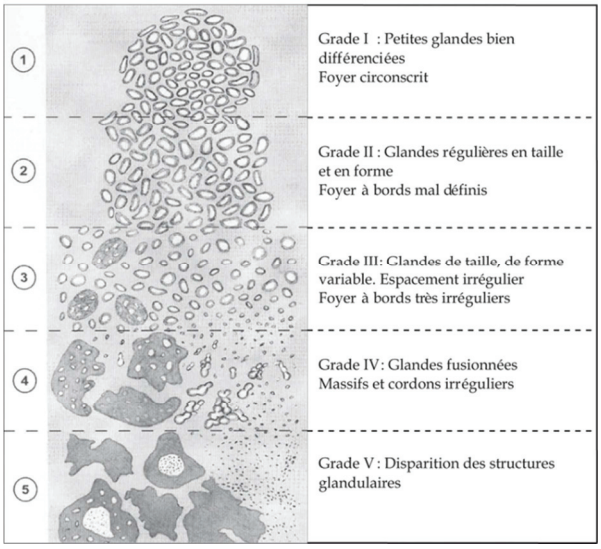


Figure I-24 Classifications TNM et de Gleason. Partie gauche : classification TNM appliquée au cancer de la prostate, d'après la Haute Autorité de Santé (HAS). BP : biopsie prostatique. Partie droite : illustration des grades de Gleason de 1 à 5 dans la version de 2005 (Epstein et al. 2005), adaptée de la classification originale proposée par Donald F. Gleason (Gleason 1966).

La recherche de marqueurs s'organise de fait selon différentes écoles de pensées, en mettant en jeu diverses hypothèses biologiques. Glinsky développe un modèle de souris transgéniques visant à montrer qu'un profil de gènes qui reflètent le statut de cellules souches peut signer des lésions métastatiques (Glinsky et al. 2004; Glinsky et al. 2005). Lapointe développe des approches de stratification de patients indépendamment des critères cliniques en superposant des résultats de transcriptomique et de génomique (nombre de copies). Il parvient ainsi à définir trois sous-groupes de cancer de la prostate, liés à l'évolution de la maladie vers des formes agressives (Lapointe et al. 2004; Lapointe et al. 2007). Enfin, True, Pascal, et d'autres, orientent la recherche de marqueurs en affinant le sous-type cellulaire par un travail de microdissection et établissent le profil transcriptomique des cellules basales, luminales et stromales (Oudes et al. 2006; Pascal et al. 2009b; Pascal et al. 2009a; True et al. 2006). Par l'essor des méthodologies bioinformatiques, un troisième niveau d'analyse entre enfin en jeu, désigné par le terme méta-analyses (Gorlov et al. 2010; Nakagawa et al. 2008; Tomlins et al. 2007b; Varambally et al. 2005). Malgré l'identification de réseaux de gènes altérés ou mis en œuvre dans la pathologie, à nouveau, les meilleurs candidats d'une étude ne sont pas nécessairement ceux d'une autre étude. Une des appréciations partagées est qu'il semble y avoir une différence du contenu relatif stroma-tumeur au cours de l'évolution de la maladie, et que cette différence serait la composante principale de variation entre les études. Une

autre tendance indique que le cancer de la prostate est plus proche de celui du poumon et du sein que du côlon et de l'ovaire, en ce sens qu'il développe des métastases osseuses. Il est à noter, cependant, des différences d'appréciation sur la validité des résultats, selon les auteurs. En outre, la concordance entre le protéome et le transcriptome peut être vue, soit comme mauvaise, soit comme un biais stratégique qu'il pourrait être avantageux d'exploiter.

I.4.2.4. La controverse XMRV

Pour clore ce paragraphe sur les marqueurs du cancer de la prostate, nous rappellerons quelques points marquants autour de la controverse XMRV. Ce gammarétrovirus apparenté à MLV (XMRV pour Xenotropic Murine leukemia-Related Virus) a été initialement détecté dans 40 % des tumeurs de prostate de patients américains homozygotes pour la mutation R462Q du gène de la RNAse L (Urisman et al. 2006). Plus tard, des sujets ne présentant pas cette mutation ont également été testés positifs à XMRV (Schlaberg et al. 2009) et la détection a été étendue au sang d'individus souffrant du syndrome de fatigue chronique (CFS) (Lombardi et al. 2009). L'affirmation qu'un nouveau rétrovirus circule dans la population humaine a cependant été remise en question par plusieurs études, essentiellement européennes, qui ont échoué à détecter XMRV dans des cohortes de patients atteints de cancer de la prostate ou de CFS (D'Arcy et al. 2008; Fischer et al. 2008; Furuta et al. 2011; Hohn et al. 2009; Sakuma et al. 2011). Etant donné que XMRV partage une forte homologie avec les séquences endogènes de la souris, des résultats positifs peuvent possiblement être attribués à une contamination de réactifs de RT-PCR par des séquences murines dans un grand nombre de cas (Sato et al. 2010). Néanmoins, la détection de provirus XMRV intégrés dans des tissus cancéreux de prostate (Dong et al. 2007; Kim et al. 2008) argumente en faveur d'un véritable virus ayant des propriétés de réplication en cellules humaines. La question qui se pose aujourd'hui est donc de savoir comment XMRV peut (ou a pu, ou pourrait) entrer dans la population humaine et s'y propager (Van der Kuyl et al. 2010). Une hypothèse propose une transmission directe du virus de la souris à l'homme (Hue et al. 2010; Weiss 2010), comme cela a pu être vu précédemment pour des hantaviruses et des arenavirus (Charrel and de, X 2010; Hart and Bennett 1999; Klein and Calisher 2007). La transmission de virus xenotropic murins (X-MLV) est en effet possible dans la mesure où les cellules humaines expriment le récepteur XPR1 (Marin et al. 1999). En plus de cela, la réplication de XMRV est très sensible aux protéines humaines APOBEC3 et aux téthérines (Bogerd et al. 2011; Chaipan et al. 2011; Groom et al. 2010; Stieler et al. 2010), ce qui laisse planer peu de doutes sur le fait qu'une réplication virale puisse avoir lieu, par exemple, dans les PBMC de patients CFS (Paprotka et al. 2010). Une hypothèse alternative est défendue par Antoinette Cornelia van der Kuyl (Van der Kuyl et al. 2010) et envisage la possibilité d'une infection par des produits biologiques dérivés de la souris, dont les vaccins produits en cellules murines pourraient être les candidats les plus sérieux. Par ailleurs, si XMRV a été introduit dans la population humaine par l'usage de certains produits

biologiques au cours des dernières décennies, alors un niveau de bruit de fond de détection du virus, qui peut varier avec les zones géographiques et des groupes d'âges, est attendu. Une telle présence à un faible niveau pourrait expliquer les écarts de détection d'une étude à l'autre, ainsi que l'association controversée du virus à certaines maladies.

II Partie Expérimentale

II.1. Contexte et stratégies d'étude

Le séquençage du génome humain en 2001 (International Human Genome Sequencing Consortium 2001) a révélé la composante rétrovirale portée en chacune de nos cellules. Les quelques 200 000 éléments à LTRs, répartis sur l'ensemble de la chromatine, sont la trace d'évènements d'infections et de propagations qui ont eu lieu au cours de l'évolution et qui ont abouti à la constitution du patrimoine rétroviral des hominidés (Mallet and Prudhomme 2004). Malgré la dérive génétique et certains niveaux de contrôles exercés par l'hôte, des éléments rétroviraux ont avantageusement été mis à profit du développement des espèces : les Syncytines 1 et 2 chez l'homme, mais plus largement les domestications d'enveloppes rétrovirales impliquées dans la morphogénèse placentaire des mammifères, illustrent une convergence évolutive qui préfigure un potentiel physiologique. A l'inverse, l'association de transcrits, de protéines ou de particules endorétrovirales avec des états cancéreux ou certaines formes de maladies auto-immunes pose la question de la régulation de l'expression des HERV en contexte pathologique.

Les capacités transcriptionnelles des rétrovirus endogènes humains s'étendent au-delà des seuls gènes rétroviraux. Par leurs caractéristiques fonctionnelles d'initiation et de terminaison de la transcription, les LTRs du génome constituent un réservoir de séquences pouvant moduler la transcription cellulaire, et de nombreux transcrits chimériques HERV-gènes ont été mis en évidence dans des contextes variés (Cohen et al. 2009). Le fait que 85 % des séquences HERV soient des LTRs solitaires et, qu'en conséquence, on évalue statistiquement entre 10 et 20 le nombre moyen de LTRs au voisinage de chaque gène, ouvre une perspective de recherche d'interactions entre les répertoires génomiques et invite à revisiter le dogme de l'ADN poubelle.

L'étude du transcriptome HERV est cependant fortement ralentie par la nature répétée des séquences et les insuffisances d'annotations des régions non-codantes du génome. L'expression des HERV a donc été très majoritairement étudiée sous l'angle de comportements globaux, au niveau des familles d'éléments, par l'utilisation de systèmes consensus d'amplification et de détection (Forsman et al. 2005; Pichon et al. 2006; Seifarth et al. 2003). En lien avec certaines hypothèses biologiques, des approches locus-candidat existent mais échouent à donner une vision d'ensemble du répertoire (Ahn and Kim 2009; de Parseval et al. 2003). Enfin, les travaux les plus ambitieux, qui ont cherché à décrire une activité locus-spécifique à grande échelle ou à identifier des éléments fonctionnels par des approches basées sur la RT-PCR et le clonage/séquençage, se résignent à l'étude laborieuse d'un sous-ensemble du répertoire et concluent finalement à la nécessité de projets collaboratifs destinés à l'étude du transcriptome HERV (Buzdin et al. 2006b; Flockerzi et al. 2008).

Dans ce contexte, le laboratoire développe, depuis 2006, une méthodologie de conception de sondes pouvant détecter spécifiquement un transcrit HERV et lui attribuer une unique origine génomique (Gimenez et al. 2010). L'utilisation du format GeneChip (Affymetrix) comme support de détection haute densité permet, avec la version actuelle de la puce HERV, d'étudier simultanément l'expression de 5 573 loci HERV distincts, dont plus de 4 000 éléments LTRs, et d'accéder à une lecture fonctionnelle pour plus de 1 500 d'entre eux.

En utilisant cet outil, le premier volet du travail expérimental de la thèse a consisté à décrire le transcriptome HERV en exploitant un panel de tissus qui reflète des contextes physiopathologiques variés. La caractérisation d'un tropisme d'expression au niveau locus et l'identification de fonctions promotrices et polyA, en lien avec un environnement génomique, constituent une première esquisse du paysage transcriptomique HERV. Puis, dans une logique d'évaluation de la variabilité inter individuelle des éléments présentant une activité différentielle, l'expression des séquences HERV associées spécifiquement au cancer du côlon a été évaluée sur une cohorte de tissus.

Une fois établi le potentiel d'utilisation de la puce HERV-V2 pour une recherche de marqueurs transcriptomiques en oncologie, les outils fondamentaux du laboratoire ont été replacés dans le contexte clinique du partenariat entre les Hospices Civils de Lyon et BioMérieux. Une des finalités biomédicales du Laboratoire Commun de Recherche est de pouvoir répondre à l'intérêt des cliniciens de diagnostiquer de manière précoce le développement du cancer de la prostate, par une méthode de détection non-invasive. Nous avons donc cherché à mettre au point des protocoles adaptés aux techniques de transcriptomique à partir de prélèvements urinaires afin de pouvoir réaliser une étude clinique pilote impliquant 45 patients. L'utilisation de la puce commerciale HG-U133-PLUS2 comme référent méthodologique a aidé à la définition de critères qualitatifs qui, appliqués au répertoire HERV, ont permis de mettre en évidence (puis de confirmer sur le même jeu d'échantillons en RT-PCR) l'expression différentielle de quelques loci HERV.

II.2. Outils et méthodes d'analyses spécifiques utilisés dans le cadre de la thèse

II.2.1. Puce haute densité HERV

II.2.1.1. Historique introductif aux puces HERV développées au laboratoire

L'étude du transcriptome des rétrovirus endogènes humains, fortement contrainte par la nature répétée des séquences a, on l'a vu, favorisé la mise en œuvre d'approches locus-candidat (Ahn and Kim 2009; de Parseval et al. 2003) ou l'évaluation de comportements à l'échelle globalisante des familles d'éléments (Forsman et al. 2005; Seifarth et al. 2003) sur des panels de tissus. Dans ce contexte, le laboratoire dispose, en 2006, d'un procédé de MD-PCR OLISA basé sur la détection de séquences consensus *pol*, ayant permis la description d'un tropisme d'expression de familles HERV sur un échantillonnage de tissus (Pichon et al. 2006). La transition technologique vers un format de puces GeneChip a été motivée par le besoin d'atteindre un niveau d'analyse spécifique (c'est-à-dire avoir une information au niveau d'un locus et non plus au niveau d'un ensemble de séquences homologues) et de tendre à l'exhaustivité du répertoire (c'est-à-dire avoir l'ambition de couvrir les 200 000 séquences à LTR du génome humain). En particulier, le postulat que, à l'échelle individuelle, les dizaines de milliers de LTRs éparpillées à travers le génome puissent représenter autant de promoteurs naturels ou alternatifs à la transcription cellulaire a été à la base de contraintes de conception ayant pour finalité d'ajouter un niveau de lecture fonctionnelle à la mesure de l'expression. Les difficultés associées au projet de développement d'une puce HERV haute densité sur support Affymetrix tiennent, bien sûr, à la méthodologie de conception des sondes (ou comment obtenir une spécificité de capture sur des éléments répétés) mais également, plus en amont, à la maitrise du répertoire HERV et de son annotation. La faisabilité d'un tel projet a été établie en utilisant une première génération de puce HERV issue du laboratoire, intégrant 4 familles, et conduisant à l'identification et à la validation de 5 nouveaux loci HERV-W réactivés sous la dépendance de leur LTR 5' dans le séminome (Gimenez et al. 2010). Les paragraphes qui suivent donneront des détails sur les critères majeurs de conception et d'amélioration qui ont conduit à la seconde génération de la puce HERV (HERV-V2), qui est celle utilisée dans ce travail de thèse.

II.2.1.2. Création et curation d'une base de données HERV : HERV-gDB

HERV-gDB est une base de données, développée au laboratoire en 2006 par Bertrand Bonnaud, qui regroupe et annote les séquences HERV du génome humain. Pour chaque famille

HERV, la copie la plus conservée du génome a servi de prototype à la recherche des séquences qui lui sont apparentées (i.e. : ayant au minimum 80 % d'homologie) sur l'ensemble du génome. Le programme RepeatMasker, dédié à l'identification de séquences répétées (Smit et al. 1996; Tempel 2012), a été utilisé cette fin. Les séquences identifiées, découpées en éléments fonctionnels (U3/R/U5 gag/pol/env) et annotées, sont finalement structurées au format ACNUC pour se conformer au format standard des bases de séquences développées au pôle bioinformatique lyonnais (PBIL) (Figure II-1).

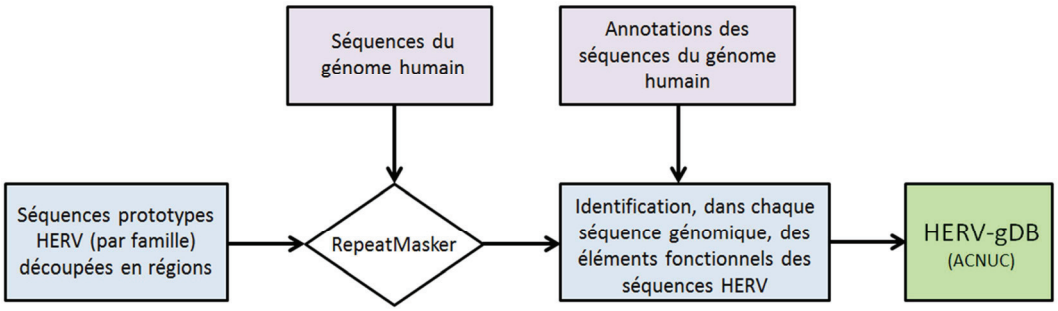


Figure II-1 Schéma du principe de création de la base HERV-gDB. Adapté du manuscrit de thèse de B. Bonnaud.

HERV-gDB est conçue pour être interrogée par familles de séquences HERV ou par type d'éléments fonctionnels. Elle n'offre néanmoins pas la souplesse d'un catalogue de copies et son interface utilisateur est réduite *a minima*. C'est en cela une structure noyau exclusivement. Différentes versions de HERV-gDB ont été développées au cours du temps, consistant essentiellement en l'ajout de nouvelles séquences et au renforcement du système d'annotation. La version utilisée pour la définition de la puce HERV-V2, HERV-gDB3, contient 10 035 copies HERV distinctes, appartenant à 6 familles, et incluant des séquences provirales complètes (environ 6 000) et partielles (environ 4 000).

II.2.1.3. Procédés de conception des sondes de la puce HERV-V2

La conception des sondes de la puce HERV-V2 repose essentiellement sur l'utilisation d'un modèle de stabilité d'hybridation adapté de la bibliographie à la problématique des éléments répétés (modèle EDA), mais la spécificité de certains questionnements biologiques a amené à recourir à deux procédés additionnels : ROSO et MANO.

II.2.1.3.1. Modèle de stabilité d'hybridation EDA

En 2006, deux travaux de recherche décrivent des modèles de stabilité d'hybridation cible/sonde sur puces à ADN pour des oligonucléotides de 18 et 20 mer, respectivement, et analysent l'importance des types de mésappariements et de leurs positions relatives sur la séquence

(Pozhitkov et al. 2006; Wick et al. 2006). Ces modèles s'accordent sur le fait qu'un mésappariement dans le tiers central impacte fortement la stabilité de l'hybridation mais que cette donnée à elle seule est insuffisante à la prédiction. Le modèle de Pozhitkov a servi de base, au laboratoire, à l'établissement de critères qualitatifs permettant d'évaluer les potentiels d'hybridation des sondes reconnaissant des séquences HERV. A ces données extérieures ont alors été ajoutées les retours d'expériences obtenus par la première génération de puce HERV et notamment l'observation que 3 mésappariements sur des séquences LTRs, quelque soient leurs positions respectives, conduisent à une absence d'hybridation. Un modèle empirique, complétant celui de Pozhitkov, a ainsi été développé pour tenir compte de l'impact de deux mésappariements en fonction de leurs positions relatives. Enfin, un troisième et dernier critère d'estimation du potentiel d'hybridation, purement correctif, a compensé une limitation inhérente à l'utilisation de l'algorithme BLAST. En effet, cette méthode d'alignement, de référence mais inadaptée aux séquences de petites tailles, renvoie dans certains cas un résultat sur 21 bases seulement, en comptabilisant à tort le reste de la séquence de 25 bases comme autant d'erreurs (Figure II-2).

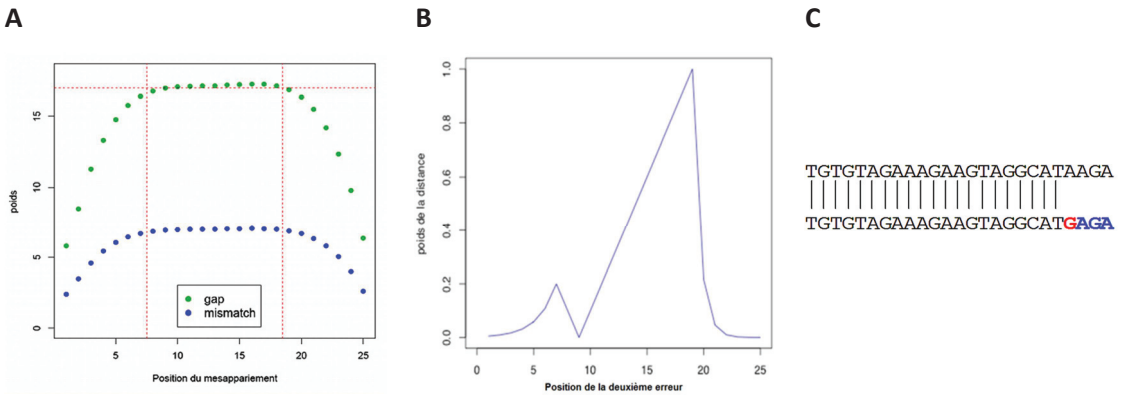


Figure II-2 Les trois paramètres d'évaluation de la stabilité d'hybridation cible/sonde. (A) Impact du type de mésappariement (gap ou mismatch) et de sa position sur une séquence de 25 mer. Modèle issu de Pozhitkov et réévalué au laboratoire. (B) Incidence empirique de la position d'une deuxième erreur ; ici exemple d'un premier mésappariement en position 9. (C) Biais introduit par l'utilisation de l'algorithme d'alignement BLAST (rouge : mésappariement réel, bleu : bases considérées comme mésappariées à tort.

Ces trois paramètres composent, ensemble, le modèle de stabilité EDA, pour : type et position des Erreurs, Distance entre deux erreurs, et longueur de l'Alignement. Un score de pénalité a été défini par la somme de ces trois paramètres, et chiffre les potentiels d'hybridation d'une sonde avec toutes ses cibles potentielles, c'est-à-dire sa spécificité de capture (Equation 1).

$$F_{SCORE} = \sum_{pos}^{toutesPos} Poids_{typePos}[pos] + Poids_{distance} + Poids_{longAlign}$$

Equation 1 Fonction de pénalité EDA. Le score de pénalité EDA est un poids cumulatif du type de mésappariement, de la distance entre deux erreurs d'appariement et de la longueur de alignement par l'utilisation de BLAST. Ce dernier terme disparait par l'utilisation de l'algorithme d'alignement KASH.

Le procédé de création et de sélection des sondes HERV est basiquement assez comparable à celui des sondes des puces commerciales Affymetrix. A partir d'une base de séquences (HERV-gDB3 ici), un découpage de sondes chevauchantes de 25 mer est tout d'abord réalisé. La sélection des sondes qui vont entrer dans la composition de la puce HERV-V2 passe en revanche par le calcul systématique de leurs scores EDA vis-à-vis de l'ensemble du génome humain, puis sont filtrées à un seuil donné avant d'être regroupées en probesets (Figure II-3).

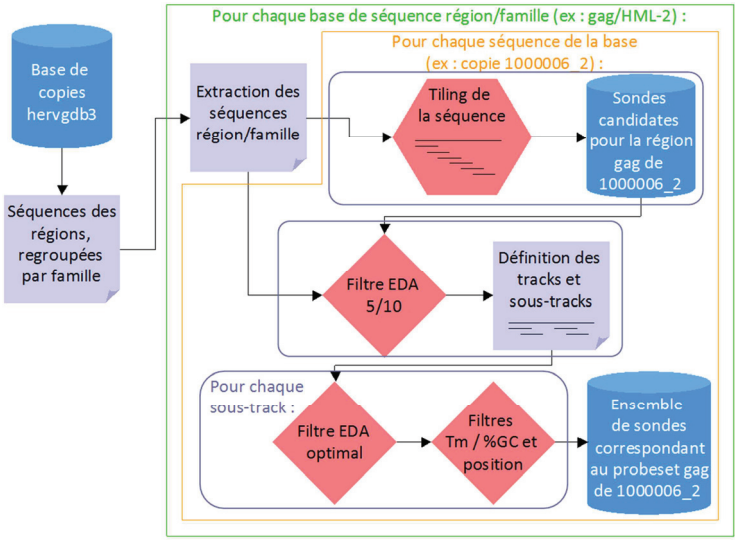


Figure II-3 Processus de sélection des sondes par le filtre EDA. Les séquences de la base HERV-gDB3 sont extraites puis découpées en sondes candidates de 25 mer (étape de tiling). Un score EDA est attribué à chaque sonde et un filtre est appliqué (5/10). Un deuxième niveau de sélection consiste à limiter le nombre total de sondes par probeset en fonction de la configuration des régions (étape de sous-track).

Dans cette démarche, le filtre 5/10 signifie que pour une séquence donnée, un maximum de 5 alignements ayant un score EDA nul est toléré (correspondant donc à l'existence de 5 cibles 100 % homologues sur le génome), ainsi qu'un maximum de 10 alignements avec un score EDA < 15 (seuil empirique au-delà duquel la spécificité d'une sonde devient trop faible). Ce compromis permet d'éviter une chute drastique de nombre de séquences HERV représentées sur la puce en voulant imposant des spécificités parfaites, tout en limitant le nombre de validations RT-PCR potentiellement nécessaires à la confirmation d'un résultat d'expression (en l'occurrence, jusqu'à 5 systèmes à concevoir par séquence dans les plus mauvais cas de figures, étant entendu que parmi les 10 alignements proches, le gradient d'intensité doit permettre de distinguer les hybridations croisées).

II.2.1.3.2. Procédé complémentaire ROSO

Un procédé générique à la conception de sondes de puces à ADN dénommé ROSO[7] (Reymond et al. 2004) a été utilisé pour définir des sondes dans les 16 transcrits d'enveloppes

[7] Le logiciel ROSO a été développé au pôle lyonnais de bioinformatique. Il est librement accessible sur internet : http://pbil.univ-lyon1.fr/roso

rétrovirales comportant des cadres de lectures ouverts complets (de Parseval et al. 2003), ainsi que des sondes pour un ensemble de gènes candidats essentiellement associés à des processus cancéreux (voir plus loin II.2.1.4). Ce répertoire, s'il reste marginal face aux apports du procédé EDA, permet un ancrage sur une méthode de référence dans une perspective d'évaluation et d'amélioration des procédés de conception qui sont propres au laboratoire.

II.2.1.3.3. Ajouts MANO

Enfin, un ensemble de sondes a fait l'objet d'une définition manuelle pour adresser des questionnements biologiques spécifiques. Ainsi, dans le but de révéler des stratégies d'épissages, des sondes aux jonctions théoriques ont été définies individuellement pour certains loci (Figure II-4). C'est le cas pour une trentaine de séquences de la famille HML-2, pour lesquelles la stratégie d'épissage de l'enveloppe est reliée au sous-type 1 ou 2 et peut conduire à la synthèse de petites protéines de régulation (Lower et al. 1995). C'est le cas également pour les loci ERVWE1 et ERVFRDE1, à l'origine des protéines d'enveloppes rétrovirales Syncytine-1 et Syncytine-2.

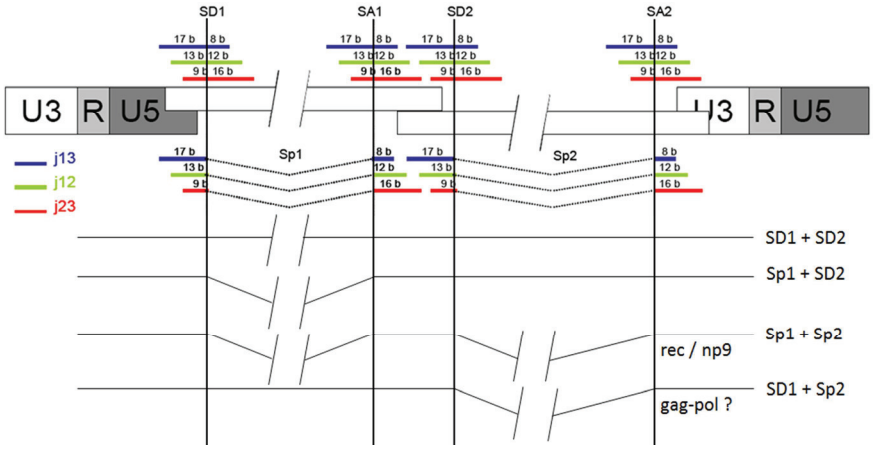


Figure II-4 Stratégie de conception des sondes aux jonctions d'épissages. Exemple d'un locus de la famille HML-2. Sur chaque site donneur (SD) et accepteur (SA), un trio de sondes chevauchantes est défini (positions 1/3, 1/2 et 2/3 pour j13, j12 et j23, respectivement). L'interprétation des transcrits ainsi détectés est indiquée sur la moitié inférieure du schéma.

Parallèlement, une sélection d'éléments présentant un polymorphisme d'insertion dans la population humaine, appartenant aux familles HERV-H (Jern et al. 2004) et HML-2 (Turner et al. 2001), a été ajoutée au répertoire de la puce HERV-V2.

II.2.1.4. Contenu de la puce HERV-V2

Les six familles HERV de la base HERV-gDB3 ont été représentées sur la puce HERV-V2 par des apports différents qui sont fonction du nombre d'éléments d'une famille donnée au sein du génome et des difficultés à concevoir des sondes à un bon niveau de spécificité au sein des familles. Cette dernière contrainte est particulièrement sensible pour les intégrations les plus récentes dans la

mesure où le niveau de divergence entre deux séquences d'une même famille est sous la dépendance du temps d'évolution[8]. Au total, la puce HERV-V2 permet la détection de 2 883 LTRs solitaires et de 2 690 provirus complets ou partiels, soit 5 573 loci HERV distincts (Tableau II-1).

	HERV-W	HERV-H	HERV-E	HERV-FRD	HERV-K HML-2	HERV-K HML-5	Total
LTRs solitaires	432	553	120	1189	512	77	**2883**
Provirus	304	1354	427	218	215	172	**2690**
LTRs 5'	120	444	29	33	29	18	673
LTRs 3'	171	485	29	43	85	19	832
gag	162	787	228	80	85	125	1467
ppol	222	0	0	0	0	0	222
pol	0	1154	307	35	93	135	1724
env	205	513	63	127	66	97	1071

Tableau II-1 Représentation en séquences rétrovirales du contenu de la puce HERV-V2. Le nombre d'éléments HERV distincts représentés sur la puce est donné pour chaque famille. Les séquences provirales incluent des structures complètes (5'LTR-gag-pol-env-3'LTR) et tronquées. La somme des deux valeurs de total en gras donne le nombre de loci HERV totaux représentés sur la puce (soit 5 573).

A ce large répertoire endo-rétroviral, les séquences de 318 gènes cellulaires issues de la puce commerciale HG-U133-PLU2 ont été ajoutées (Tableau II-2). Ces gènes candidats ont été choisis pour leur association à des processus cancéreux (certains sont spécifiques à un type de cancer, d'autres associés aux cancers en général, d'autres encore ont une qualité pronostique), à des mécanismes de régulation épigénétiques (ADN methyltransferases, enzymes de modification du code histone, machinerie des miRNA) ou encore parce que participant aux voies de régulation des HERV (hormones, cytokines). Le choix et le nombre de ces candidats ajoutés à la puce HERV-V2 a été un compromis entre l'opportunité d'obtenir des d'informations complémentaires sur un processus biologique et l'encombrement de l'espace sur la puce induite par l'ajout de séquences non rétrovirales.

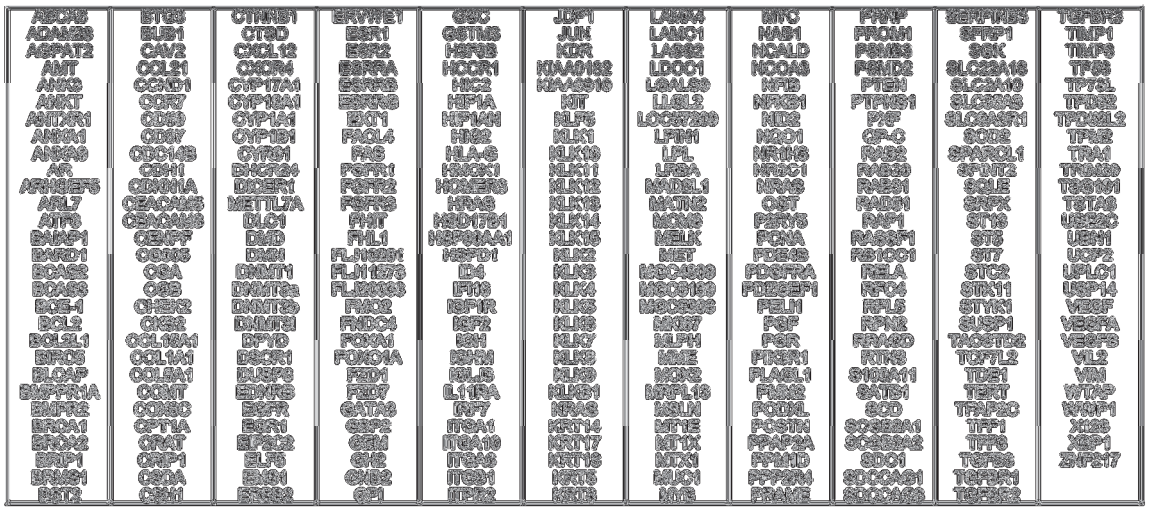


Tableau II-2 Gènes candidats présents sur la puce HERV-V2. Les séquences de ces gènes sont directement transposées de la puce HG-U133-PLUS2.

[8] A l'exclusion notable des éléments domestiqués, voir (Mallet et al. 2004).

Il est à noter, enfin, qu'un volumineux jeu de sondes d'apprentissage a été ajouté au contenu de la puce HERV-V2. Ces sondes, qui sont des déclinaisons de combinaisons d'erreurs à partir des séquences Perfect Match de 19 gènes à tropisme placentaire, servent à affiner la compréhension du phénomène de réactions croisées en général, et sa traduction par l'amélioration des critères du modèle EDA en particulier.

II.2.1.5. Méthodes d'analyses spécifiques aux données de la puce HERV-V2

L'analyse des puces HERV s'inscrit dans le schéma classique d'une analyse de puces à ADN Affymetrix. Un rappel méthodologique couvrant les aspects génériques est en conséquent renvoyé en annexe II. Nous développerons ici les singularités de l'analyse des puces HERV-V2.

II.2.1.5.1.1. Lecture fonctionnelle de l'expression des LTRs

Une caractéristique essentielle de la conception de la puce HERV-V2 tient en la tentative de définition systématique de probesets dans les régions U3 et U5 des LTRs. La dichotomie de signaux U3/U5 offre ainsi une lecture fonctionnelle de l'expression et permet l'identification d'éléments autonomes pour la transcription (i.e. : promoteur ou polyA) (Figure II-5). Il est donc possible d'étudier une LTR indépendamment de son contexte proviral, sans *a priori* sur sa fonction. L'ensemble des LTRs représentées sur la puce HERV-V2 (LTRs solitaires, LTRs 5' et LTRs 3') peut alors être abordé comme une collection d'éléments indépendants, répartis sur le génome, et pouvant être autant d'unités de contrôle de la transcription.

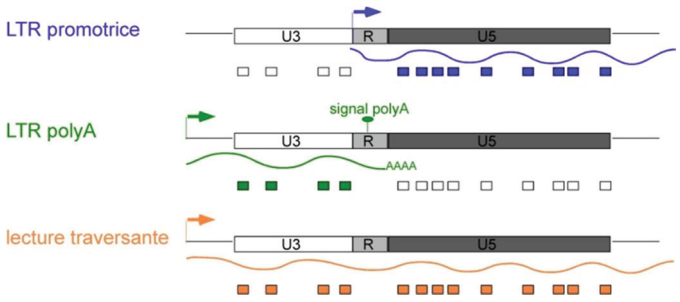


Figure II-5 Lectures fonctionnelles des signaux LTRs de la puce HERV-V2. Schéma représentant les trois situations théoriques de transcription pour une LTR. De haut en bas : la transcription commence à la frontière U3/R et la LTR a un rôle de promoteur ; la transcription s'arrête dans la région R par l'ajout d'une queue polyA ; la LTR est traversée de manière passive. Les boites représentées sous les séquences correspondent à la position des probesets de la puce HERV-V2 et sont colorées pour indiquer une détection positive.

Cette mise en œuvre est toutefois contrainte par le niveau de succès dans la conception de sondes spécifiques au sein des deux régions U3 et U5 de chaque LTR. La puce HERV-V2 comporte ainsi 1 513 LTRs dites fonctionnalisables, c'est-à-dire auxquelles il est possible d'associer un scénario

de fonction. Cela représente un tiers de l'effectif total des LTRs de la puce et signifie que seule une observation d'expression, non reliable à une fonction, sera le lot du reste de l'effectif.

II.2.1.5.1.2. Mise en place et utilisation de systèmes d'annotations

Le développement de puces à façon visant un répertoire génomique longtemps relégué au rang de poubelle se confronte fatalement au retard de développement pris dans la création d'outils d'étude adaptés. Une démarche d'annotation et le déploiement d'interfaces de représentation ont été mis en place durant ce travail de thèse, en collaboration avec le service de bio-informatique de bioMérieux, dans le but de faciliter l'analyse et l'interprétation biologique des résultats liés aux HERV. Deux niveaux de bases de données ont tout d'abord été définis. Au niveau sondes et probesets, une base de connaissance puce (BDCP) recense les informations de conception telles que : les séquences des sondes, les scores EDA associés, le nombre de sondes entrant dans la composition d'un probeset, le positionnement génomique des sondes, et des liens directs vers les principaux Genome Browsers. Au niveau des loci HERV, une base de connaissance copies (BDCC) est mise en place. La BDCC contient les annotations des régions fonctionnelles au sein d'une séquence HERV (gag/pol/env) et les bornes de découpage de ses LTRs (U3/R/U5) ; elle donne également le positionnement chromosomique de chaque locus, l'adjoint de la présence d'éléments SINE ou LINE le cas échéant, et intègre l'environnement nucléique à ± 10 kb. L'étude des gènes cellulaire présents dans l'environnement génomique des séquences HERV fait intervenir des scripts codés en perl et n'est réalisée que dans le cadre de problématiques de recherches ciblées. Ces différents niveaux de bases de données, auxquels s'ajoute le contenu initial de HERV-gDB3, sont compilés dans l'interface de représentation Geneious (Figure II-6).

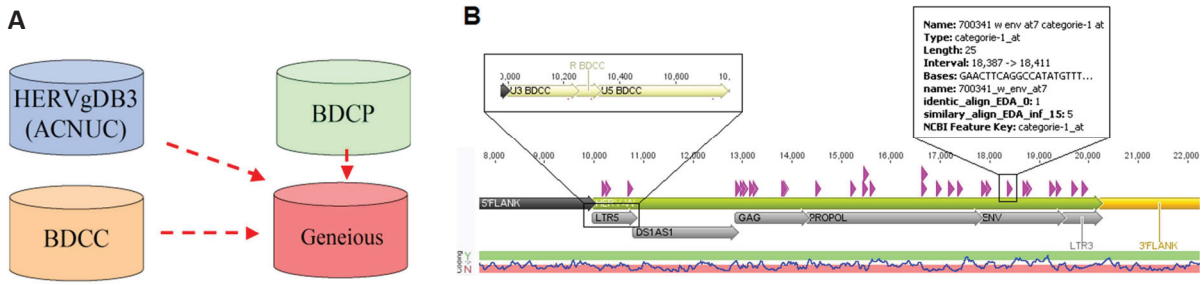


Figure II-6 Mise en place d'outils d'annotations destinés à l'analyse du répertoire HERV. (A) Procédure de regroupement des différentes bases de données internes. (B) Intégration et représentation sous Geneious. Exemple du locus ERVWE1 (700341_w). La séquence du locus HERV est représentée en vert entre l'ADN 5' et 3' flanquant, ses régions fonctionnelles sont délimitées en gris et le découpage de ses LTRs est illustré par le grossissement de la LTR 5'. Les sondes de détection de la puce associées au locus sont positionnées en rose au-dessus de la séquence, et un système d'annotations dynamiques en résume les principales caractéristiques. Une prédiction *in silico* des capacités codantes est indiquée par un profil de probabilité sous la séquence (vert : Yes, rouge : No).

Ainsi, les différents outils de connaissance associés à l'utilisation de la puce HERV-V2 permettent une certaine souplesse dans l'interprétation des données qui en découlent. Notamment, la conception d'amorces de PCR pour des étapes aval de validation est simplifiée par un mode de représentation intuitif.

II.2.2. Couplage RT-PCR et HRM pour la validation de l'expression des séquences HERV individuelles

En 2003, une équipe de chercheurs de l'université de l'Utah, en collaboration avec la société Idaho Technology, met au point une nouvelle méthode de fusion des produits de PCR permettant la détection de différences de séquences au nucléotide près (Gundry et al. 2003; Wittwer et al. 2003). La technique repose sur une acquisition haute résolution de la fluorescence lors de l'étape de dissociation de brins réalisée en fin de cycle de PCR. Ce gain de sensibilité parvient à être obtenu par l'application d'un gradient de température continu aux échantillons, et par l'utilisation de molécules fluorescentes intercalantes en quantité saturante sans effet inhibiteur de réaction. Ainsi, l'arrivée de nouvelles générations de thermocycleurs permettant des gradients de température de 0,1°C (citons le LighCycler 480 de Roche ou encore le RotorGene de Corbett racheté par Qiagen) et la mise sur le marché de fluorochromes saturants (LC Green, ResoLight, EvaGreen ou SYTO 9 par exemple), le tout associé à un coût opérationnel relativement faible et à une grande simplicité d'utilisation ont progressivement contribuer à la diffusion de cette technique à travers le monde. Les courbes de fusion à haute résolution (HRM) sont à présent largement utilisées en microbiologie dans des applications de typage de souches (Ganopoulos et al. 2012; Tong et al. 2011) ou mises en œuvre pour des études de polymorphismes ou de variants alléliques sur des gènes cellulaires (Chang et al. 2012; Nissen et al. 2012), méthodes qui nécessitaient traditionnellement l'utilisation de techniques d'électrophorèse ou de clonage/séquençage. Un des aspects de ce travail de thèse a été d'appliquer la technologie HRM à l'étude des éléments répétés, dans le cadre de la validation par RT-PCR de l'expression des séquences HERV.

II.2.2.1. Mise au point de systèmes de PCR HERV locus-spécifiques

La conception d'amorces de PCR destinées à la détection de transcrits HERV repose basiquement sur les mêmes contraintes physiques et thermodynamiques que pour n'importe quel transcrit (Abd-Elsalam 2003). L'utilisation du procédé de HRM exclut cependant la possibilité de travailler avec des systèmes de PCR basés sur des sondes (TaqMan, Molecular Beacon, Scorpion) puisque la fluorescence suivie lors de la fusion des produits d'amplification est directement reliée à la structure bi caténaire de l'ADN. Cette contrainte réduit les possibilités de gagner en spécificité de

détection et incite par conséquent à renforcer les étapes de contrôle lors de la sélection des amorces. Ainsi, dans le cadre des transcrits HERV à valider, la sélection d'une première amorce spécifique vient de l'utilisation de la séquence d'une sonde de la puce HERV-V2 ayant un score EDA nul. Sur cette base, la recherche du meilleur couple d'amorces est réalisée à l'aide de Primer-BLAST[9], en ayant pour critères de sélection la spécificité théorique de l'amplicon, et la spécificité de la seconde amorce seule quand cela est possible. Une PCR *in silico* est réalisée en contrôle complémentaire à l'aide des outils mis en ligne par UCSC[10]. Puis, les amorces synthétisées et reçues, une étape de mise au point ayant pour finalité de déterminer expérimentalement la température optimale de fixation des amorces (T_m) est réalisée par l'application de gradients de température sur cible d'ADN génomique (Figure II-7).

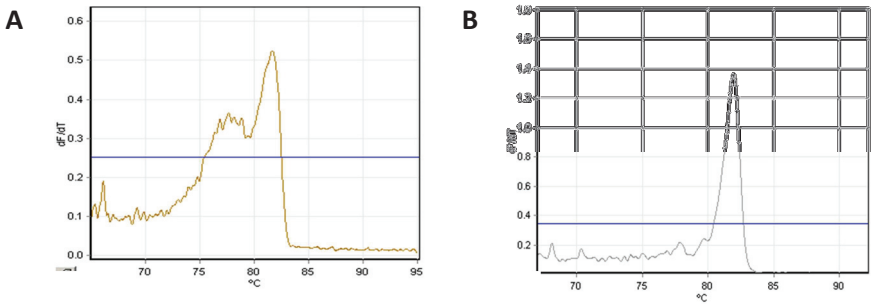


Figure II-7 Détermination de la température optimale de fixation des amorces de PCR par contrôle HRM des produits d'amplification. Exemples des pics de fusion HRM du produit de PCR de la séquence d'enveloppe du locus 700126_w, réalisés sur cible d'ADN génomique. (A) Température de fixation des amorces de 55°C. (B) Température de fixation des amorces de 61°C.

Le contrôle des produits de PCR par HRM indique la nature de l'amplicon. Un unique pic bien résolu signe la présence d'un amplicon homogène et valide l'utilisation de la paire d'amorces de PCR à une température T_m donnée. L'atout de la résolution HRM est de permettre d'individualiser des pics très proches (pouvant refléter des séquences très proches) qui auraient été englobés dans un seul pic par une technique de fusion classique. Il est à noter qu'à ce stade, l'application d'un gradient de température ne permet pas systématiquement d'aboutir à un produit d'amplification unique. Dans ce cas, la mise au point du système est abandonnée et de nouvelles amorces sont définies. Les produits d'amplification satisfaisant au contrôle HRM à une température T_m donnée font l'objet d'un séquençage pour qu'il soit établi, une fois, que l'observation d'un pic de fusion à une température T_{hrm} donnée correspond au produit d'amplification attendu. La HRM devient alors le seul contrôle d'amplification des expériences de RT-PCR qui suivent.

[9] Primer3 suivi d'une procédure BLAST, voir (Dunk et al. 2012)

[10] http://genome.ucsc.edu/cgi-bin/hgPcr

II.2.2.2. Mise en évidence de populations de produits de PCR HERV par la HRM

L'application du procédé de HRM au contrôle des produits d'amplification de PCR s'avère particulièrement utile pour l'étude des éléments répétés. En pouvant discriminer deux séquences qui ne diffèrent que d'une base, il devient en effet possible d'établir des signatures populationnelles de séquences homologues, par exemple dans le contexte de séquences appartenant à une même famille HERV. A titre d'exemple, nous avons cherché à discriminer 5 séquences *gag* de la famille HERV-E présentant un niveau d'homologie supérieur à 80 % (Figure II-8).

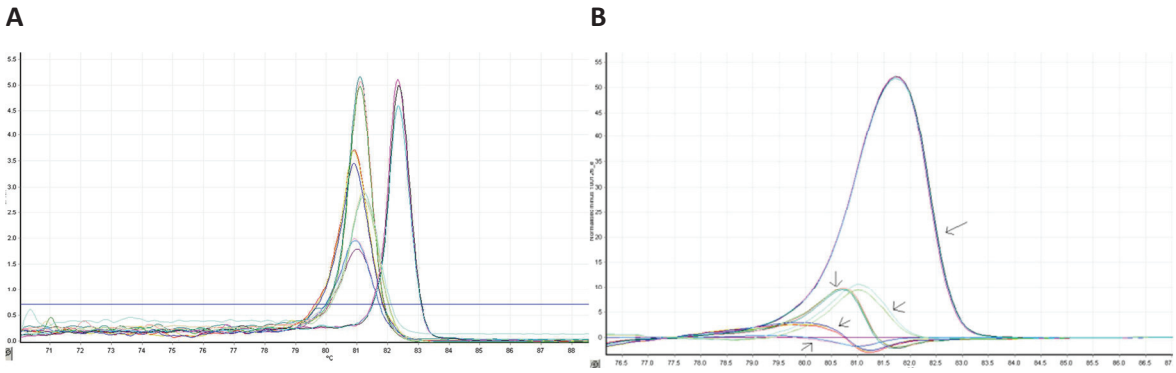


Figure II-8 Discrimination de populations de produits de PCR par la méthode de HRM. (A) Pics de fusion HRM associés à 5 séquences *gag* HERV-E présentant une homologie de séquence supérieure à 80 % (triplicats techniques). (B) Représentation de génotypage HRM appliquée aux pics, mettant en évidence 5 populations distinctes (flèches noires).

La fusion des produits d'amplification, dans cet exemple, met en évidence des pics uniques à des températures T_{hrm} parfois très rapprochées. Néanmoins, la résolution d'acquisition des valeurs (pas : 0,1°C) permet un traitement analytique des courbes qui fait apparaitre 5 comportements de fusion distincts, correspond à 5 populations de produits de PCR.

II.3. Etude du transcriptome HERV sur tissus

II.3.1. Etapes principales du protocole de réalisation des puces à ADN

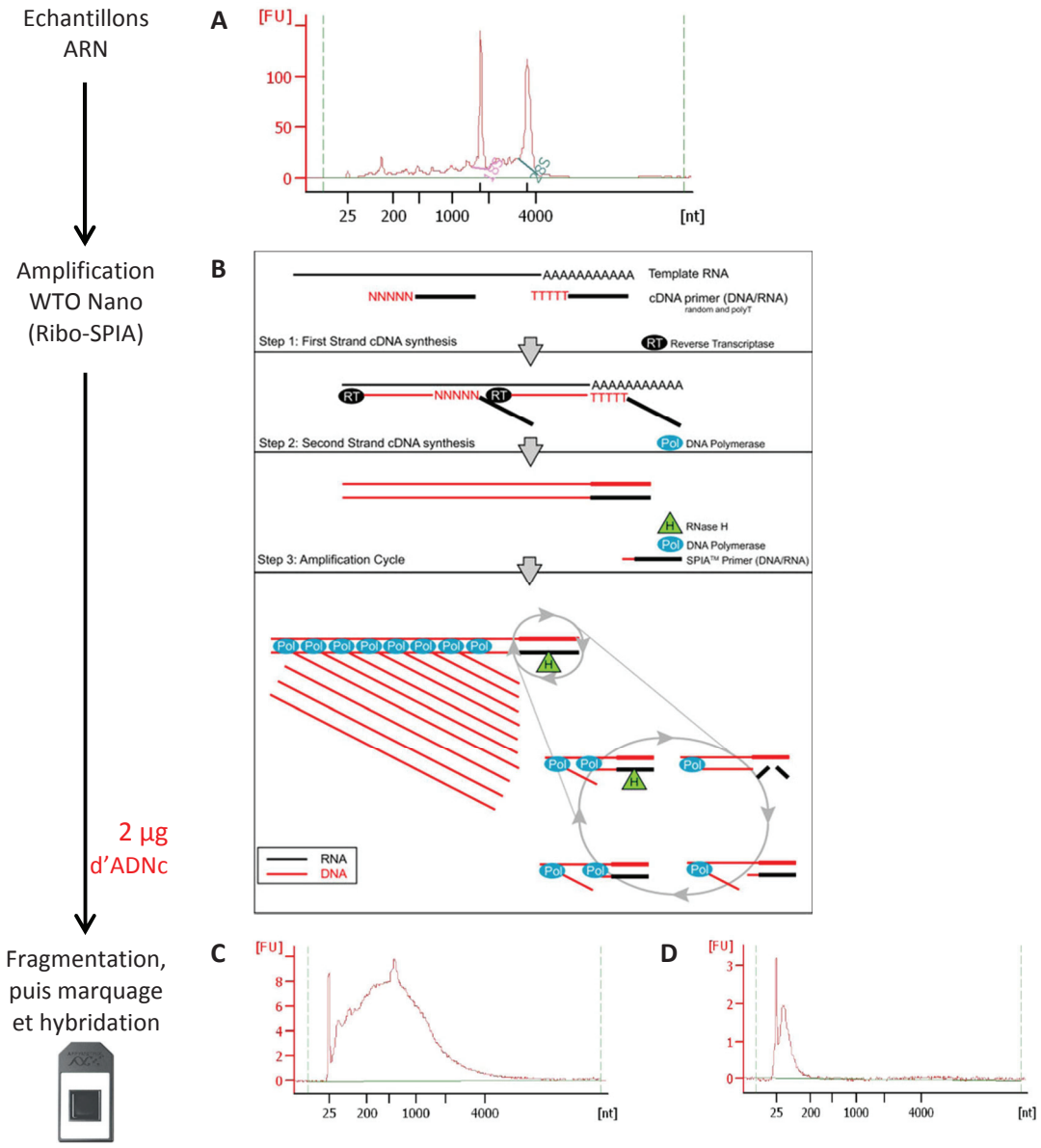


Figure II-9 De l'échantillon d'ARN à l'hybridation sur puces à ADN. (A) Caractérisation des ARNm par électrophorèse capillaire à l'aide du bioanalyzer. La qualité des échantillons d'ARNm est évaluée par le calcul du RIN (RNA Integrity Number). Cette valeur, basée sur la détection des ARN ribosomaux 18S et 28S, tend vers 10 lorsque l'ARNm est intègre. Ici, illustration d'un RIN égal à 8, considéré en pratique comme un échantillon de bonne qualité. (B) Procédé d'amplification Ribo-SPIA. Etape 1 : un premier brin d'ADNc est généré par reverse transcription à partir d'une matrice d'ARNm, en utilisant un mélange d'amorces aléatoires et d'oligo-dT. Etape 2 : l'ADN polymérase est ajoutée à la réaction est génère le second brin d'ADNc. Etape 3 : amplification SPIA par déplacement de brin. Les amorces hybrides ADN/ARN sont dégradées par l'activité RNAse H de l'ADN polymérase lorsqu'elles sont complexées avec la matrice d'ADNc. La synthèse d'ADN simple brin (complémentaire à la matrice ARNm) est amorcée et se poursuit, autorisant de nouvelles amorces SPIA à se fixer à la matrice d'ADNc, entretenant ainsi le processus répétitif de synthèse de brin. D'après (Watson et al. 2008) et adapté de NuGEN. (C) Profil d'amplification. La gamme de tailles des ADNc s'étale de 25 à 4 000 nucléotides. (D) Profil de fragmentation. La population d'acides nucléiques est centrée autour d'une taille de 100 nucléotides, ce qui est adapté à l'hybridation sur puces à ADN Affymetrix.

Des ARN tissulaires provenant de différentes sources (commerciales et partenariales) ont constitué un échantillonnage de travail représentatif de contextes biologiques variés et orientés vers des situations pathologiques cancéreuses. Des paires d'échantillons de tumeurs et de tissus normaux (adjacents à la tumeur) ont été recrutées auprès de 40 individus différents, où sont représentés (i) dans leurs composantes saine et cancéreuse : le côlon, le sein, le poumon, la prostate, l'ovaire, l'utérus, le testicule et (ii) dans sa seule composante saine : le placenta. Les contextes prépondérants sont le poumon (10 tissus sains et 10 cancers dont 9 formes épidermoïdes), le sein (8 tissus sains et 8 carcinomes invasifs à différents stades de différenciation) et la prostate (8 tissus sains et 8 adénocarcinomes à des stades d'évolution variés). Ensemble, 79 échantillons constituent le panel d'étude (40 tissus sains et 39 tumeurs).

Après avoir été caractérisés, les ARN ont été convertis en ADNc puis amplifiés de façon linéaire par une méthode isotherme décrite en 2005 par la société NuGEN Technologies (Kurn et al. 2005) (Figure II-9). Cette méthode, dite Ribo-SPIA (commercialisée sous le nom WTO Nano et désignée comme telle dans le manuscrit), basée sur le déplacement de brin et l'activité RNAse H de l'ADN polymérase, a été initialement évaluée dans des contextes d'amplification et de détection de petites quantités d'ARN. Elle s'est avérée être plus simple, plus rapide et à l'origine de brin d'ADNc plus stables que les méthodes d'amplification de transcription *in vitro* basées sur l'utilisation de la T7 ARN polymérase (Singh et al. 2005). L'amplification Ribo-SPIA a notamment l'avantage de conserver la polarité des transcrits en générant des ADNc simple brin complémentaires des ARNm. Pour ces raisons, ce protocole d'amplification a semblé adapté à l'étude du transcriptome HERV.

Deux microgrammes de cibles d'ADNc ont été fragmentés à la DNAse puis hybridés sur la puce HERV-V2 pour chacun des échantillons biologiques. Le jeu de données complet est composé de 113 puces, qui ont été réalisées dans des temps fractionnés, en nombre variable de réplicats et avec différents lots de réactifs. Pour ces raisons, l'analyse des données a eu recourt à une méthode corrective de l'impact des lots d'expériences (Johnson et al. 2007), en plus de la normalisation RMA (Irizarry et al. 2003) communément utilisée (l'annexe II présente une démarche type d'analyse de puces à ADN, de l'acquisition des données aux tests statistiques pour l'identification de gènes différentiellement exprimés). Cette correction dite COMBAT (pour COMbining BATches) est adaptée dans des cas où les structures de données sont clairement corrélées à des variables expérimentales. COMBAT, corrigeant la variabilité technique du jeu de données, fait ressortir sa variabilité biologique (Equation 2).

$$Y_{ijg} = \mu_g + X_j\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}$$

Equation 2 Modèle correctif COMBAT. Y_{ijg} est le signal d'expression mesuré pour le gène *g* lorsque l'échantillon *j* est traité dans le lot d'expérience *i* ; μ_g est la moyenne d'expression du gène *g* ; une unique co-variable biologique *X* est considérée (ici : une variable qualitative qui inclut le type et l'état du tissu) ; β_g définit le niveau d'expression différentielle associé aux catégories biologiques (le paramètre d'intérêt) ; les paramètres γ_{ig} et δ_{ig} sont des composantes d'erreurs additives et multiplicatives qui caractérisent le lot d'expérience *i* (elles sont spécifiques d'un gène) et ε_{ijg} est l'erreur résiduelle qui suit une loi $N(0, \sigma_g)$. D'après (Johnson et al. 2007).

II.3.2. Description du transcriptome HERV

Les données d'expression des puces ont été regroupées par contextes biologiques (e.g. : côlon tumoral, côlon normal, prostate tumorale, prostate normale, etc.). Pour un probeset donné, seule la valeur médiane des puces du groupe est étudiée. Dans une première approche descriptive de l'expression, un seuil de bruit de fond est appliqué et les évènements positifs de transcription, tous contextes biologiques confondus, sont comptés. Cette première analyse a conduit à l'identification de 1 718 loci HERV actifs, dont 808 LTRs solitaires (Tableau II-3). Cette activité transcriptionnelle représente un tiers du contenu de la puce HERV-V2, ce qui peut laisser penser, par extrapolation, qu'environ un tiers du répertoire HERV est actif.

	HERV-W	HERV-H	HERV-E	HERV-FRD	HERV-K HML-2	HERV-K HML-5	Total
LTRs solitaires	100	209	30	251	199	19	**808**
Provirus	101	587	91	39	75	17	**910**
LTRs 5'	26	154	10	8	10	4	212
LTRs 3'	43	182	10	11	35	7	288
gag	12	202	51	4	15	4	288
ppol	8	0	0	0	0	0	8
pol	0	170	28	1	9	2	210
env	49	71	5	16	12	2	155

Tableau II-3 Détection du transcriptome HERV au niveau locus. Le nombre d'éléments présentant une expression au-dessus du bruit de fond est indiqué pour chacune des 6 six familles HERV représentées sur la puce HERV-V2. La somme des deux valeurs de total en gras donne le nombre de loci HERV total actifs sur le panel de tissus étudié (soit 1 718).

Des phénomènes de transcription au niveau locus ont été observés pour l'ensemble des 6 familles HERV représentées sur la puce HERV-V2, ainsi que pour différentes régions fonctionnelles *gag*/*pol*/*env*. Nous avons cherché à comparer cette activité au contenu rétroviral du génome pour tenter de révéler des comportements privilégiés. Pour cela, un génome HERV virtuel a été défini comme la somme des éléments présents dans la base de données HERV du laboratoire, HERV-gDB3. Ce génome HERV, lu soit sous l'angle de la composition en familles soit sous celui de la structure des loci, est comparé au résultat global de transcriptomique (c'est-à-dire la somme des évènements individuels d'expression) (Figure II-10). Ceci a montré que la proportion d'expression relative des éléments de chaque famille est très comparable au contenu génomique. En revanche,

proportionnellement, la transcription des LTRs est supérieure à celles des gènes proviraux *gag*/*pol*/*env*, pouvant suggérer que l'activité d'un ensemble de LTRs est liée à une autonomie fonctionnelle, ce que nous avons par la suite cherché à objectiver (voir partie II.3.4 Lecture fonctionnelle du transcriptome en lien avec son environnement génomique).

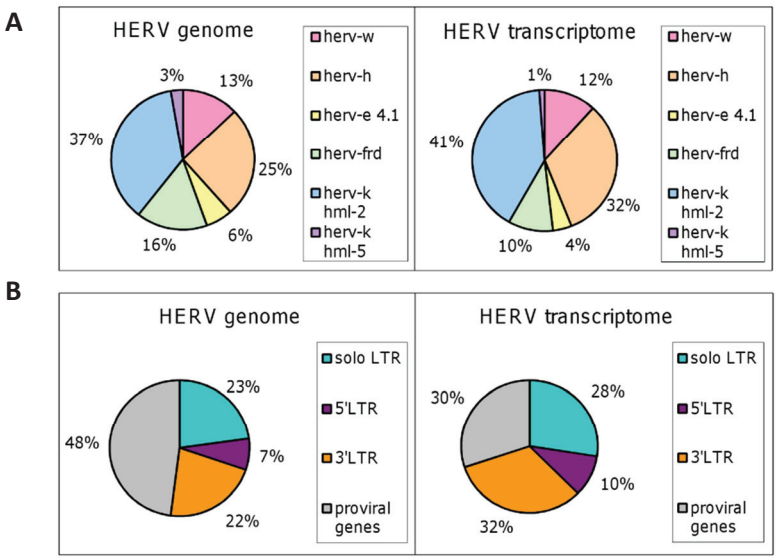


Figure II-10 Projections génomiques et transcriptomiques du répertoire HERV. L'image d'un génome HERV virtuel est basée sur le contenu de HERV-gDB3. L'extrapolation du transcriptome dérive de l'activité observée au niveau des loci, après l'application de facteurs correctifs qui prennent en compte le succès de conception des sondes. (A) Familles HERV. (B) Structures des loci HERV.

La structure des données d'expression a été explorée par une méthode de partitionnement hiérarchique euclidien. Cette procédure vise à regrouper les probesets qui, dans le jeu de données, présentent des comportements d'expression similaires. Dix profils d'expression ont pu être mis en évidence (Figure II-11, page suivante). Des expressions spécifiques de tissus ont ainsi été montrées pour le côlon tumoral, le testicule tumoral, le testicule sain, l'ovaire tumoral, le placenta normal et, dans une moindre mesure, la prostate tumorale et le poumon sain. A ces 7 groupes tissu-spécifiques, deux profils d'expression impliquant un ensemble de contextes biologiques ont été caractérisés : le premier regroupe des séquences HERV s'exprimant à la fois dans le testicule tumoral et le placenta normal, et le second rassemble les éléments qui s'expriment conjointement dans l'ovaire tumoral, l'utérus sain, l'utérus tumoral, la prostate saine et la prostate tumorale. Enfin, un profil recense toutes les séquences HERV qui s'expriment dans l'ensemble des tissus.

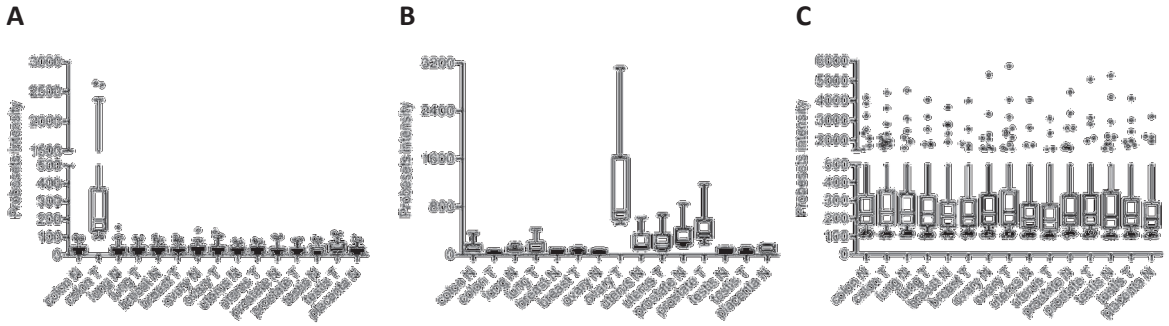


Figure II-11 Tropisme d'expression HERV (3 cas de figures sur 10 existants). Distribution des intensités d'expression des loci en fonction du contexte biologique. (A) Ensemble de loci HERV qui s'expriment spécifiquement dans un contexte (ici le côlon tumoral). (B) Ensemble de loci HERV qui s'expriment dans un sous-ensemble de tissus (ici l'ovaire tumoral, l'utérus sain, l'utérus tumoral, la prostate saine et la prostate tumorale). (C) Ensemble des loci HERV qui s'expriment à un fort niveau quelque soit le contexte biologique. Les médianes des distributions sont indiquées par une barre horizontale et les boites encadrent 90 % des valeurs de la distribution. Les points montrent les 'outliers'.

Il est intéressant de constater que le contenu des profils d'expression correspond parfois à des enrichissements de certaines familles HERV. Par exemple, sur les 43 probesets HERV qui composent le profil côlon tumoral, 42 appartiennent à la famille HERV-H. Cependant, cette remarque n'est pas généralisable et des profils d'expression comme le testicule tumoral impliquent des éléments appartenant à plusieurs familles. De tels résultats suggèrent que la régulation de l'expression en condition pathologique peut avoir une composante liée aux familles de rétrovirus endogènes (par exemple par la conservation des sites de fixation de facteurs de transcription) à laquelle s'ajoute vraisemblablement des mécanismes séquence-dépendants (par exemple des levées de méthylation locales ou des phénomènes d'embarquement transcriptionnels).

II.3.3. Eléments présentant une expression différentielle

Dans le but d'associer des évènements de transcription à un état pathologique, une recherche d'expression différentielle a été réalisée. La procédure SAM (Tusher et al. 2001) avec contrôle du taux de faux positifs (Storey and Tibshirani 2003) (voir annexe II) a mis en évidence un nombre important de probesets subissant un changement d'expression statistiquement significatif entre les états sains et cancéreux d'un même tissu. A ce titre, 41 probesets ont été identifiés pour le sein, 102 pour la prostate, 305 pour le poumon, 354 pour le côlon, 618 pour l'ovaire et plus de 1 000 pour le testicule. Par comparaison des listes obtenues dans les différents contextes biologiques, des probesets tissu-spécifiques ont pu être caractérisés (Figure II-12, page suivante).

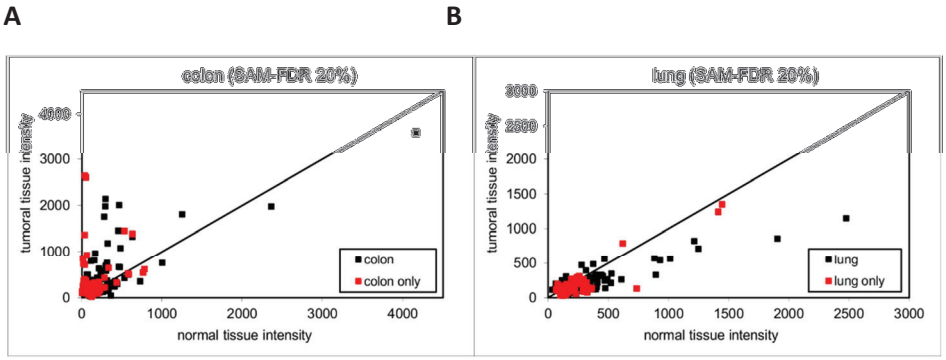


Figure II-12 Expression différentielle HERV entre tissu sain et tumoral (2 cas de figures sur 7 existants). Le test statistique SAM (Tusher et al. 2001) avec contrôle du taux de faux positifs (Storey and Tibshirani 2003) a mis en évidence des probesets dont les niveaux d'expression varient entre les états normaux et cancéreux. Les valeurs d'expression de ces probesets sont représentées sous forme de nuage de points, et une coloration rouge est appliquée lorsque l'expression différentielle d'un probeset n'a été retrouvée dans aucun autre couple de tissus. La ligne diagonale représente une absence de différentiel d'expression, et par conséquent les variations significatives sont positionnées loin de cette droite. Expression différentielle identifiée dans (A) le côlon et (B) le poumon.

La limite des tests statistiques tient à ce qu'ils ne répondent toujours qu'à une question prérequise (et si possible formulée correctement), à l'exclusion de beaucoup d'autres qui auraient pu s'imposer légitiment au cours d'une analyse par essais erreurs. En cela, la valeur de leurs résultats est complémentaire de ceux obtenus par des approches exploratoires de type partitionnement. Typiquement, dans le contexte de cette étude, l'approche SAM-FDR identifie beaucoup d'évènements statistiquement significatifs mais qui n'ont que peu, voire pas, de signification biologique (les nombreuses variations dans des valeurs d'expressions très faibles en sont une bonne illustration). Les approches exploratoires présentées plus haut, quant à elles, et dénuées d'indicateurs statistiques objectifs, permettent de faire des rapprochements opportuns sur la base de rationnels biologiques, par exemple en constatant qu'il est pertinent de créer un groupe avec un effectif certes très faible mais constitué exclusivement de probesets de la famille HERV-E. Ainsi, une hiérarchisation des résultats d'expression, dans une perspective d'identification de marqueurs de cancers, a été effectuée sur la base du recoupement des deux approches analytiques.

II.3.4. Lecture fonctionnelle du transcriptome HERV et lien avec son environnement génomique

Pour approcher la question de la régulation de la transcription par les éléments LTRs, nous avons utilisé la dichotomie de signaux des régions U3 et U5 de la puce HERV-V2, comme décrit dans le paragraphe II.2.5.1.1. Compte tenu de la dérive génétique (et de l'absence de critères objectifs pour définir une LTR active), la fonction de chaque LTR a été étudiée sans *a priori* vis-à-vis de sa structure provirale. Parmi les 1 513 LTRs de la puce HERV-V2 pour lesquelles il est possible de

conclure sur une activité fonctionnelle, 326 ont été identifiées comme promotrices et 209 comme polyA, toutes conditions biologiques confondues. En utilisant un seuil de négativité assez strict (fixé à 50, soit inférieur à 2^6-2^7 (64-128) qui est la fourchette de seuil généralement admise pour les puces Affymetrix commerciales HG-U133-PLUS2), 672 LTRs ont été déclarées silencieuses dans l'ensemble des tissus. Le reste de l'effectif des LTRs n'a pu être interprété en raison de signaux d'expression ambigus (valeurs comprises entre 50 et 100). De manière remarquable, la liste des LTRs promotrices est exclusive de celle des LTRs polyA, montrant ainsi qu'une LTR ne change pas de fonction (bien qu'elle puisse ne plus s'exprimer) selon le tissu dans lequel elle se trouve. Dit autrement, une LTR active est une LTR fonctionnellement spécialisée, soit promotrice, soit polyA. Pour cette raison, nous avons proposé le concept de déterminisme opérationnel puis questionné sa validité sur un jeu de données de culture cellulaire. Cette expérience annexe *in vitro* a corroboré les observations faites sur tissus et, par ailleurs, n'a ajouté que 2 % de nouvelles caractérisations fonctionnelles. Ensemble, ces résultats nous ont donc amené à conclure que nous avons identifié et délimité une population de LTRs actives et autonomes. Ces caractérisations fonctionnelles concernent en majorité des LTRs solitaires (247 promoteurs et 151 polyA) avec des distributions inégales d'une famille à l'autre (Tableau II-4). Par exemple, la famille HML-5 ne présente aucun cas de LTR solitaire polyA. D'autre part, les familles HERV-K HML-2, HERV-W, HERV-E 4.1 et HERV-FRD affichent une surreprésentation de LTRs solitaires promotrices par rapport aux LTRs solitaires polyA, ce qui est le contraire de la tendance observée pour la famille HERV-H. S'agissant des contextes de structures provirales, 34 LTRs 5' promotrices et 30 LTRs 3' polyA ont été identifiées, mais aussi, de manière moins attendue, 45 LTRs 3' promotrices et 28 LTRs 5' polyA.

	HERV-W			HERV-H			HERV-E 4.1			HERV-FRD			HERV-K HML-2			HERV-K HML-5			Total		
	Solo LTR	*5' LTR*	*3' LTR*	*Solo LTR*	*5' LTR*	*3' LTR*	*Solo LTR*	*5' LTR*	*3' LTR*	*Solo LTR*	*5' LTR*	*3' LTR*	*Solo LTR*	*5' LTR*	*3' LTR*	*Solo LTR*	*5' LTR*	*3' LTR*	*Solo LTR*	*5' LTR*	*3' LTR*
promoter	57	9	14	21	13	13	7	3	1	76	2	3	69	6	10	17	1	4	247	34	45
polyA	40	6	2	29	15	19	4	2	1	53	2	0	25	2	8	0	1	0	151	28	30
readthrough	0	0	1	5	4	3	1	0	0	6	0	0	3	0	0	1	0	1	16	4	5
silent	109	14	21	43	24	33	7	4	1	281	11	12	64	3	12	21	7	5	525	63	84

Tableau II-4 LTRs fonctionnelles identifiées à partir de tissus. Pour chaque famille, une recherche systématique de fonctions a été appliquée aux séquences LTRs pour lesquelles des probesets ont pu être préalablement conçus dans les régions U3 et U5.

Nous avons alors étudié l'environnement génomique des LTRs actives et silencieuses identifiées. Les noms et coordonnées génomiques des gènes présents dans les 50 kb adjacents à chaque LTR, accompagnés du pourcentage de bases G et C de la séquence, ont été extraits *in silico* à l'aide d'un script perl développé au laboratoire par Nathalie Mugnier. Ceci a permis de montrer que la densité génique est environ 1,5 fois plus élevée au voisinage des LTRs actives (promotrices ou

polyA) que dans celui des LTRs silencieuses, alors que le contenu en GC ne varie quasiment pas. Une telle observation avait déjà été réalisée pour des éléments promoteurs de la famille HML-2 (Buzdin et al. 2006a), et peut donner lieu à deux interprétations : (i) l'expression des séquences LTRs est stimulée par l'activité transcriptionnelle de la région chromosomique (par l'apport de facteurs de transcription, ou profitant d'une ouverture de brin) et n'est qu'un effet secondaire de la machinerie cellulaire, ou (ii) les LTRs actives et autonomes, par leur activité promotrices et polyA, contribuent à la régulation de l'expression des gènes d'une manière bénéfique pour l'hôte. Dans ce cas, la sous-représentation des LTRs silencieuses dans les régions riches en gènes pourrait prémunir d'une propagation de méthylation dont l'effet serait l'extinction des gènes cellulaires adjacents, comme cela a été suggéré pour les éléments transposables (Hollister and Gaut 2009). Pour apporter des arguments à ce questionnement, nous avons, pour les LTRs actives et silencieuses, opposé les représentations des environnements génomiques.

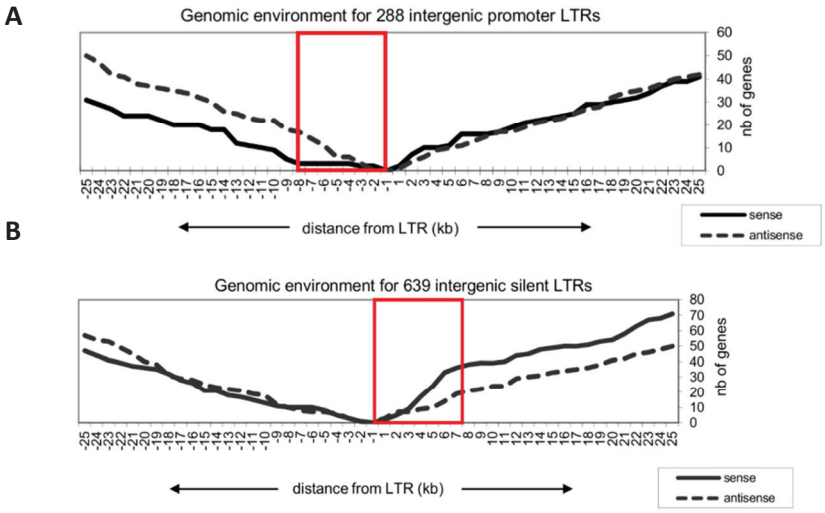


Figure II-13 Environnements génomiques des LTRs promotrices et silencieuses intergéniques. Le nombre de gènes à ± 25 kb des LTRs est obtenu à partir de la version du génome NCBI 36/hg18 et des annotations de la table RefGene (UCSC). Les gènes dans la même orientation que les LTRs (sense) ou en opposition des LTRs (antisense) sont comptés pour l'ensemble des LTRs promotrices et silencieuses. La valeur centrale en zéro représente la position des LTRs. Les occurrences cumulatives des gènes de l'environnement sont tracées en amont et en aval des LTRs. (A) Environnement génomique des 288 LTRs promotrices intergéniques. (B) Environnement génomique des 639 LTRs silencieuses intergéniques.

La comparaison des environnements des LTRs promotrices et silencieuses fait apparaitre des situations de déséquilibre au sein des effectifs de gènes sens et antisens (déséquilibre en amont des LTRs promotrices, déséquilibre en aval des LTRs silencieuses) (Figure II-13). Une telle situation n'aurait pas dû être observée si l'on suit l'hypothèse de neutralité de présence des LTRs. Plus remarquable, cette projection fait apparaitre une situation miroir (encadrés rouges), dans laquelle une zone de 8 kb semble délimiter des inflexions de comportements. La zone de 8 kb en amont des LTRs promotrices est caractérisée par une nette sous-représentation de gènes en orientation sens. Interprétée dans le cadre de l'hypothèse d'une contribution des LTRs à la régulation des gènes, cette

observation peut suggérer qu'une LTR autonome ne peut être tolérée par l'hôte que si elle se trouve à une certaine distance des gènes en orientation sens qui la précèdent, sans quoi l'utilisation de site accepteurs d'épissage ou l'ajout d'un signal polyA alternatif pourraient se traduire par un effet délétère sur le gène, comme cela a été proposé pour les éléments HERV introniques (van de Lagemaat et al. 2006). La zone de 8 kb en aval des LTRs silencieuses, pour laquelle les effectifs en gènes antisens croissent plus lentement que ceux des gènes sens, est plus délicate à analyser dans ce cadre théorique. Cependant, bien qu'aucune explication convaincante n'ait été proposée à ce jour, il est frappant de constater qu'une région symétrique, de 8 kb, autour des sites d'initiation de transcription de gènes spécifiques de tissus, a été récemment associée à un maximum de densité de séquences LTRs (Jjingo et al. 2011).

II.3.5. Validation des résultats

Les résultats d'expression ainsi que l'identification des fonctions des LTRs ont fait l'objet d'une stratégie de validation à plusieurs niveaux. La RT-PCR quantitative a tout d'abord été utilisée pour confirmer l'expression de 18 séquences HERV, appartenant à 8 loci impliqués essentiellement dans une expression tissu-spécifique. Sur l'ensemble de ces expériences, une corrélation moyenne de 0,922 (min : 0,606 ; max : 0,998) a confirmé les tendances de tropisme (Figure II-14). Pour deux loci HERV cependant, une expression, identifiée spécifiquement dans le colon tumoral par la puce, a été également associée, quoi qu'à un niveau plus faible, à l'ovaire tumoral. Cet écart peut refléter des différences de sensibilité de détection entre la puce et la RT-PCR, introduites par exemple par des biais d'amplification de la méthode Ribo-SPIA. La mise en œuvre d'une stratégie de validation des fonctions des LTRs par RT-PCR est plus délicate à mettre en place car elle se confronte à la difficulté de concevoir deux systèmes spécifiques, un dans la région U3 et un dans la région U5, sur une séquence d'ADN de courte taille. Suivant cette approche, trois LTR 5' promotrices ont pu être validées. Nous avons également confirmé que 5 LTR 5' promotrices validées en RT-PCR dans une étude précédente (Gimenez et al. 2010) ont été détectées sur la puce HERV-V2, ce qui va dans le sens d'une fiabilité de détection de la puce.

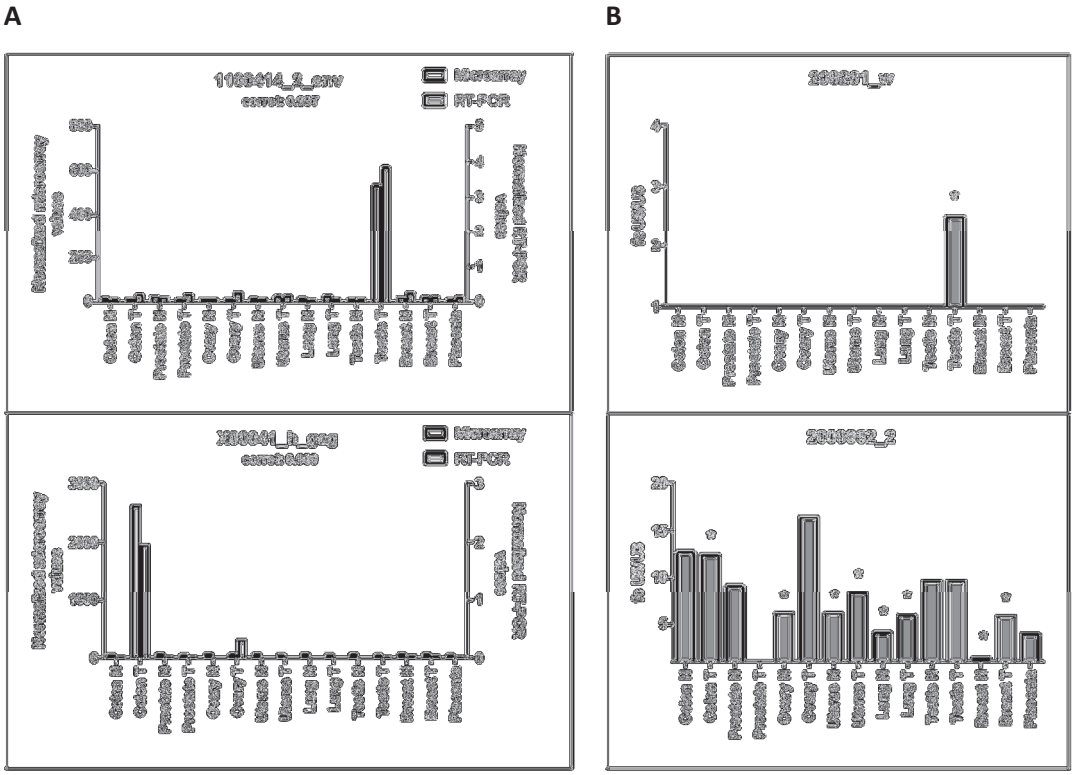


Figure II-14 Validation RT-PCR de l'expression et des fonctions identifiées par la puce HERV-V2 (extrait). (A) Corrélations d'expressions des données de puce et de RT-PCR pour deux systèmes HERV tissu-spécifiques : 1100414_2_env et X00041_h_gag. La valeur de corrélation est calculée par Correl = $Cov_{puce;RT\text{-}PCR}$ / (σ_{puce} x $\sigma_{RT\text{-}PCR}$). (B) Ratios d'expression de RT-PCR U5/U3 pour deux LTRs identifiées comme promotrices. L'expression relative U5/U3 est donnée par $Fc_{U5/U3}$ = (Eff_{U3}^{CtU3}) / (Eff_{U5}^{CtU5}). Les valeurs supérieures à 1 indiquent une activité promotrice. Une étoile (*) précise les tissus pour lesquels l'activité promotrice a été mise en évidence sans équivoque sur la puce.

Compte-tenu du nombre d'évènements de transcription détectés par l'approche puce, les méthodes de confirmation reposant sur la conception d'amorces de PCR ne peuvent être appliquées qu'à quelques candidats bien choisis, tout au plus. C'est pourquoi nous avons, parallèlement, exploité les données des banques d'ESTs. Les loci HERV à tropisme d'expression identifiés par la puce ont été alignés sur l'ensemble des banques de tissus pour en faire ressortir les enrichissements respectifs. Après normalisation des résultats, les loci HERV spécifiques du côlon, de l'ovaire, du placenta et du testicule ont pu être associés à un enrichissement au sein des bases d'ESTs correspondantes (Figure II-15). En revanche, le tropisme des éléments HERV dans le cas du poumon et de la prostate n'a pas été supporté par les banques d'ESTs, probablement en raison du niveau d'expression plus faible de ces éléments.

colon T	
EST	*Normalized hits*
colon	65,7
lung	39,3
ovary	35,6
placenta	21,0
prostate	11,7
testis	28,2

ovary T	
EST	*Normalized hits*
colon	48,0
lung	55,0
ovary	58,8
placenta	28,4
prostate	35,8
testis	43,0

placenta N	
EST	*Normalized hits*
colon	12,0
lung	14,8
ovary	16,9
placenta	20,6
prostate	7,8
testis	17,9

planceta N + testis T	
EST	*Normalized hits*
colon	17,7
lung	16,1
ovary	32,3
prostate	16,1
placenta + testis	48,0

testis N	
EST	*Normalized hits*
colon	44,1
lung	39,2
ovary	31,4
placenta	27,5
prostate	34,6
testis	40,5

testis T	
EST	*Normalized hits*
colon	17,6
lung	13,3
ovary	10,0
placenta	10,6
prostate	7,3
testis	14,6

lung N	
EST	*Normalized hits*
colon	53,8
lung	46,7
ovary	51,4
placenta	33,8
prostate	57,3
testis	62,0

prostate T	
EST	*Normalized hits*
colon	17,1
lung	13,6
ovary	22,9
placenta	15,8
prostate	20,3
testis	22,4

Figure II-15 Correspondance des séquences HERV tissu-spécifiques avec le contenu des banques d'ESTs. La banque CleanEST a été utilisée comme source d'ESTs dans le but de construire 6 groupes de référence : côlon (311122 ESTs), poumon (441913 ESTs), ovaire (123944 ESTs), placenta (321881 ESTs), prostate (69860 ESTs) et testis (264243 ESTs). Une procédure BLAST a été appliquée à partir de chaque groupe sur l'ensemble des séquences HERV composant les profils d'expression précédemment décrits. Le nombre d'alignements est alors normalisé par l'effectif de séquences HERV du groupe ainsi que par le nombre total d'ESTs formant le groupe de référence. Au sein de chaque tableau, le rang de la valeur d'intérêt est associé à une couleur pour souligner l'enrichissement en ESTs : vert (rang 1 sur 6), jaune (2 sur 6) ou rouge (supérieur à 2 sur 6).

Par ailleurs, l'analyse des fonctions des LTRs par les ESTs a été réalisée dans le seul cadre des éléments de la famille HERV-W. Ce choix s'appuie sur le fait que la famille HERV-W contient le locus domestiqué ERVWE1 pour lequel les fonctions promotrices et polyA des LTRs ont été expérimentalement démontrées (Blond et al. 1999; Cheng et al. 2004b; Gimenez et al. 2010; Mallet et al. 2004; Prudhomme et al. 2004), ce qui permet d'avoir un référent méthodologique pour juger de la pertinence générale des résultats des banques. 17 loci sur les 131 loci analysés ont des correspondances parmi les ESTs (un seuil de 97 % d'homologie a été retenu comme compromis entre la forte similarité qui existe entre deux séquences HERV de la même famille et le polymorphisme dans la population humaine), et 16 sont finalement interprétables à des niveaux de crédits différents. La corrélation entre les résultats d'expression de puce et les banques d'EST, pour la famille HERV-W, s'évalue ainsi entre 50 % et 75 %. L'étude des banques d'ESTs a montré que des évènements d'épissage qui excluent la région U3 de la LTR peuvent avoir lieu dans certains cas, ce qui peut avoir conduit à une surestimation du nombre de promoteurs. Mais à l'inverse, des fonctions promotrices n'ont pas été attribuées aux LTRs qui présentaient un fort signal dans U5 mais pour lesquelles aucune sonde spécifique n'a pu être définie dans la région U3.

Enfin, une comparaison de nos résultats avec les travaux menés par d'autres équipes, tant sur les fonctions que sur le tropisme d'expression, a été réalisée sur les éléments de la famille HML-2. Nous avons inclus les études basées sur le couplage PCR et clonage/séquençage (Flockerzi et al. 2008), sur l'identification à large échelle d'éléments promoteurs (GREM) (Buzdin et al. 2006b), sur l'exploitation des banques d'ESTs (Stauffer et al. 2004), et ajouté à cela le travail du laboratoire mené

avec la première génération de puce HERV (Gimenez et al. 2010). Sur les 327 loci HML-2 que nous avons identifiés, 25 sont partagés par au moins une des études précédentes. Sur cet ensemble, nous affichons des corrélations de 64 %, 63 % et 50 % avec les approches par ESTs, PCR-clonage/séquençage, et première génération de puce HERV, respectivement. Une plus faible corrélation, de 19 %, est associée à GREM. En particulier, notre étude a identifié 34 % d'éléments HML-2 promoteurs, ce qui est moins que la revendication de 50 % formulée par Buzdin à l'issue de son étude sur le parenchyme testiculaire (Buzdin et al. 2006a). Ceci peut être lié à des phénomènes de recombinaisons ayant lieu durant l'étude de PCR de GREM (Flockerzi et al. 2007), ou, plus simplement, à la variabilité inter individuelle entre les tissus utilisés par chaque équipe.

II.3.6. Etude de la variabilité inter individuelle de séquences HERV candidates dans le cancer du côlon

Les motifs d'expression du transcriptome décrits à l'aide de la puce HERV-V2 reposent sur un panel d'échantillons supposé être représentatif de la biologie humaine et, dans une certaine mesure, de la population. Il est pourtant criant que le faible nombre d'échantillons utilisés pour chaque contexte ne peut suffire à rendre compte des variations d'expressions pouvant exister entre les individus. Or, l'estimation des variations inter individuelles est une problématique centrale dans les étapes conduisant à l'identification de marqueurs de pathologies. Pour nous, il s'agit également d'un moyen de consolider la valeur de la démarche d'utilisation de la puce HERV-V2 dans un contexte d'identification de marqueurs de cancers. Nous avons donc fait le choix d'exploiter un des profils d'expression les mieux résolus, celui qui délimite un ensemble de séquences spécifiques du côlon tumoral. Pour rappel, ce groupe de 43 probesets contient 42 probesets de la famille HERV-H, ce qui en fait, à ce titre, un ensemble ayant probablement une signification biologique particulière.

Il convient ici de rappeler que plusieurs travaux ont précédemment associé l'expression de la famille HERV-H avec le cancer du côlon. L'exploitation des banques d'ESTs par Stauffer a permis d'identifier une expression plus importante de cette famille dans les cancers de l'estomac et du côlon que dans d'autres tumeurs, mais n'a pas conclu sur l'identification de loci individuels (Stauffer et al. 2004). La MD-PCR OLISA développée au laboratoire, basée sur des séquences consensus *pol*, a montré une forte activité de la famille HERV-H dans le côlon tumoral (Pichon et al. 2006). Les tentatives de caractérisation de loci HERV-H individuels, en association avec une expression dans le cancer du côlon, s'amorcent par un résultat un peu fortuit obtenu par une équipe allemande qui cherchait à identifier des différences d'expression de gènes cellulaires par une méthode d'hybridation soustractive (SSH) (Wentzensen et al. 2004). Un locus HERV-H précis a ainsi été caractérisé sur le chromosome X en position Xp22.3 et, évalué plus tard par la même équipe sur une

cohorte de 34 patients, s'est avéré être surexprimé dans 47 % des échantillons de cancer du côlon (Wentzensen et al. 2007). Des propriétés immuno-modulatrices de l'ORF d'enveloppe du locus Xp22.3 ont été récemment mises en évidence (Mullins and Linnebacher 2012). Dans sa suite, une équipe chinoise a étudié le locus Xp22.3, décrivant d'abord sa forte surexpression pour 7 sur 8 des échantillons de côlon cancéreux à disposition (Liang et al. 2007), puis confirmant sur une cohorte chinoise plus large la surexpression du locus pour 19 échantillons sur 20 (Liang et al. 2009a). De tels résultats peuvent notamment poser la question de la prédisposition populationnelle à l'expression du locus Xp22.3 dans le développement cancéreux.

Le locus Xp22.3 fait partie des séquences HERV associées à un tropisme d'expression côlon tumoral par la puce HERV-V2. Il est désigné sous l'identifiant X00041_h, et un système d'amplification spécifique dans sa région *gag* a été mis au point au laboratoire en utilisant la HRM (cf. II.2.2). Pour cette étude de variabilité inter individuelle, nous avons également retenu deux autres loci HERV-H montrant de forts signaux d'expression spécifiques sur la puce : le locus 2000045_h et le locus 1400177_h. Des amorces d'amplification ont été définies dans les régions *gag* et *pol* de ces loci. Enfin, un locus HERV-H dont l'expression a été spécifiquement associée au séminome (1900006_h) a été utilisé comme témoin négatif (Figure II-16).

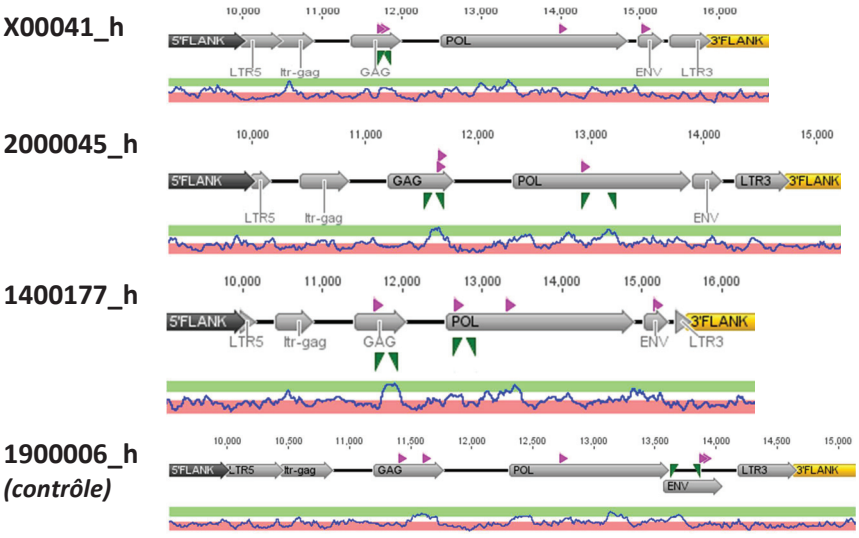


Figure II-16 Loci HERV-H candidats associés au cancer du côlon. Représentation Geneious des 3 loci HERV-H dont l'expression a été évaluée en RT-PCR quantitative sur un ensemble de tissus de côlon (X00041_h, 2000045_h et 1400177_h) et d'un locus HERV-H contrôle surexprimé spécifiquement dans le séminome (1900006_h). Les sondes de détection de la puce associées au locus sont positionnées en rose au-dessus de la séquence, et les amorces de PCR qui ont été conçues sont représentées par des triangles verts. Une prédiction *in silico* des capacités codantes est indiquée par un profil de probabilité sous la séquence (vert : probables capacités codantes, rouge : probable absence de capacités codantes).

L'obtention d'échantillons s'est faite par deux sources : 10 préparations d'ARN de tumeurs du côlon ont été achetées auprès de la société américaine BioChain Institute, et 13 paires de tissus (tumeur / zone saine adjacente) ont été recrutées au centre anticancéreux de l'Université Fudan de

Shanghai, en partenariat avec les équipes de BioMérieux en Chine. Ceci porte à 28 le nombre d'échantillons de cancers du côlon, accompagnés de 18 échantillons sains, auxquels s'ajoutent les 14 contextes tissulaires ayant permis la description du transcriptome HERV. C'est sur ce panel qu'a été évaluée la variabilité d'expression des loci HERV-H candidats.

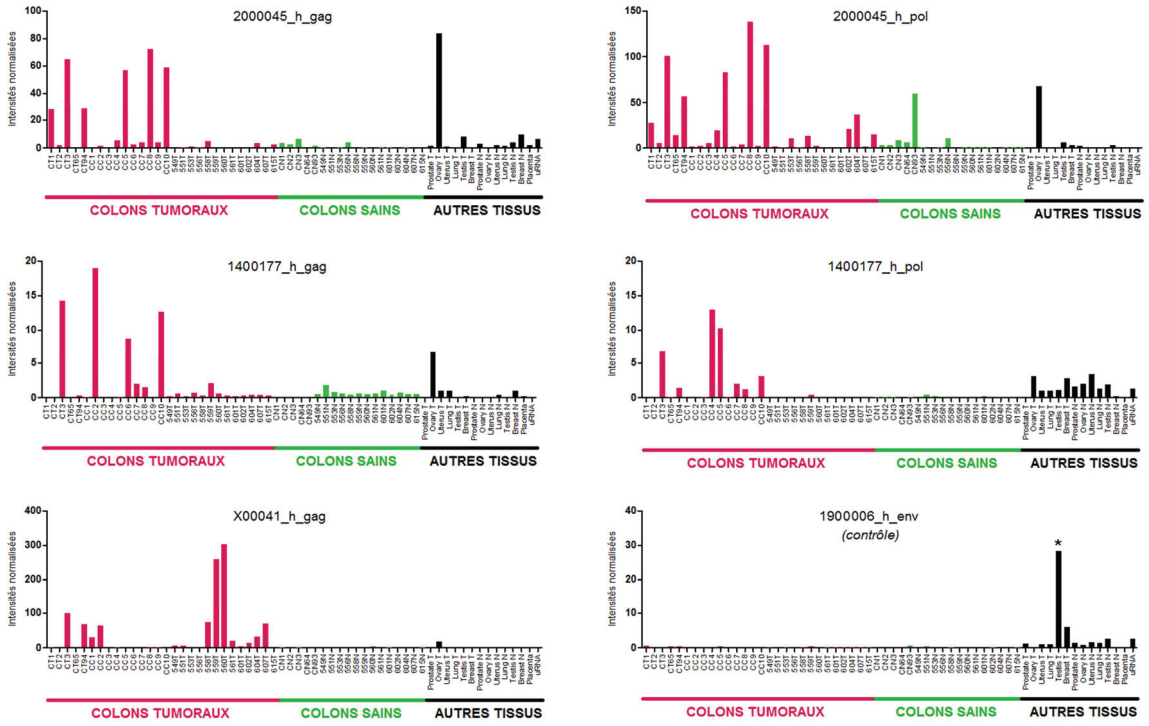


Figure II-17 Valeurs d'expression de RT-PCR quantitative des loci HERV-H candidats sur un ensemble de tissus. Les expressions des loci associés au cancer du côlon (200045_h, 1400177_h et X00041_h) et d'un locus HERV-H contrôle dont la transcription a été associée au testicule tumoral par l'approche puce (l'enveloppe de 1900006_h) ont été mesurées sur 28 échantillons de côlons tumoraux (rouge), 18 échantillons de côlons sains (vert) et 14 contextes tissulaires additionnels (noir). L'échantillon uRNA correspond à un mélange d'une vingtaine d'ARN extraits à partir de tissus humains sains. Une étoile (*) indique l'échantillon de séminome pour le système de détection contrôle 1900006_h_env.

Les trois loci candidats présentent une spécificité d'expression globalement bonne (Figure II-17) : à l'exception d'une détection unique dans un échantillon de côlon sain (pour le système 2000045_pol) et d'une détection quasiment systématique dans un autre contexte tumoral (l'utérus), une surexpression claire est associée à certains des échantillons de cancer du côlon testés. Plus précisément, dans le cas du locus 2000045_h, les profils d'expression des séquences *gag* et *pol* sont assez comparables, mettant en évidence entre 6 et 10 cas de surexpression sur 28 tumeurs, en fonction du seuil de positivité que l'on fixe (ici : estimé entre 5 et 10 sur les échelles normalisées). A ce titre, le système *gag* semble générer moins de bruit de fond vis-à-vis des échantillons de côlon sains. Les résultats du locus 1400177_h indiquent 4 cas de surexpression pour la région *gag* et 4 cas également pour la région *pol* mais, de façon intéressante, la détection recouvre en partie des échantillons différents : CC2 et CC6 ne sont détectés que par le système *gag*, quand CC4 et CC5 ne le sont que par le système *pol*. Une telle observation suggère qu'il existe des stratégies d'épissages ou

du polymorphisme au sein du locus. Enfin, le locus X00041_h (Xp22.3) affiche les valeurs d'expression les plus fortes de l'expérience, détectant entre 10 et 13 échantillons de cancer du côlon. La combinaison des différents systèmes d'amplification permet une détection de 16 à 22 des 28 tumeurs du côlon (57 % à 78 %), pour, au maximum, un évènement de détection positive dans les côlon sain.

Il est frappant d'opposer les détections du système X00041_h_gag avec celles des deux autres loci candidats. Pour ce locus du chromosome X, la série d'échantillons de 551T à 615T, venant du recrutement de Shanghai, donne des détections positives en plus grand nombre et un des niveaux plus forts que les échantillons de CC1 à CC10 obtenus auprès de BioChain ; ce motif de détection est pour ainsi dire inversé lorsque l'on considère les loci 2000045_h et 1400177_h. Malheureusement, si cette observation peut suggérer l'existence d'une susceptibilité populationnelle, nous avons appris récemment par le fournisseur américain BioChain que les échantillons qui nous ont été vendus proviennent tous d'un recrutement effectué dans des hôpitaux chinois[11]. Ceci n'exclut toutefois pas qu'il puisse y avoir des zones populationnelles en Chine liées, par exemple à des habitudes alimentaires, comme cela a été proposé pour expliquer l'augmentation de l'incidence des cancers colorectaux qui s'observe à Shanghai depuis les années 1970 (Chiu et al. 2003; You et al. 2002).

Pour clore cette partie expérimentale consacrée à l'étude de l'expression des rétrovirus endogènes humains, et au terme de ces éléments de discussion sur la variabilité inter individuelle, on retiendra que la puce HERV-V2 permet d'identifier des loci candidats en association à un état cancéreux. Le cas de la famille HERV-H est assez révélateur des avancées progressives réalisées par différents groupes. Avant la caractérisation du locus Xp22.3 (Wentzensen et al. 2004), les seules descriptions de l'activité des éléments rétroviraux HERV-H se font au niveau de la famille. Puis des tentatives, parfois infructueuses (Liang et al. 2009a), de caractérisation de nouveaux loci HERV sont réalisées et se traduisent aujourd'hui par la connaissance d'un ensemble de séquences HERV-H associées à un état pathologique. Il a ainsi tout récemment été montré que le locus 1400177_h, que nous avons identifié dans ce travail, est fortement actif dans une lignée cellulaire de cancer du côlon (Liang et al. 2012). Si la puissance de l'approche par puces à ADN permet de conserver une certaine avance dans l'identification de séquences HERV actives et fonctionnelles, il est clair que l'écart se réduit progressivement dans la course à la maitrise du répertoire HERV. L'enjeu pour nous est donc de capitaliser rapidement sur le savoir-faire de conception des puces HERV. Nous verrons dans la seconde partie expérimentale de la thèse l'application qui en a été faite dans le cadre du partenariat clinique entre les Hospices Civils de Lyon et BioMérieux.

[11] Voir J. Stiglitz, Le triomphe de la cupidité.

OPEN ACCESS Freely available online

PLoS one

Microarray-Based Sketches of the HERV Transcriptome Landscape

Philippe Pérot[1], Nathalie Mugnier[2], Cécile Montgiraud[1], Juliette Gimenez[1], Magali Jaillard[2], Bertrand Bonnaud[2], François Mallet[1]*

1 Joint Unit Hospices Civils de Lyon, bioMérieux, Cancer Biomarkers Research Group, Centre Hospitalier Lyon Sud, Lyon, France, **2** BioMérieux, Data and Knowledge Laboratory, Marcy l'Etoile, France

Abstract

Human endogenous retroviruses (HERVs) are spread throughout the genome and their long terminal repeats (LTRs) constitute a wide collection of putative regulatory sequences. Phylogenetic similarities and the profusion of integration sites, two inherent characteristics of transposable elements, make it difficult to study individual locus expression in a large-scale approach, and historically apart from some placental and testis-regulated elements, it was generally accepted that HERVs are silent due to epigenetic control. Herein, we have introduced a generic method aiming to optimally characterize individual loci associated with 25-mer probes by minimizing cross-hybridization risks. We therefore set up a microarray dedicated to a collection of 5,573 HERVs that can reasonably be assigned to a unique genomic position. We obtained a first view of the HERV transcriptome by using a composite panel of 40 normal and 39 tumor samples. The experiment showed that almost one third of the HERV repertoire is indeed transcribed. The HERV transcriptome follows tropism rules, is sensitive to the state of differentiation and, unexpectedly, seems not to correlate with the age of the HERV families. The probeset definition within the U3 and U5 regions was used to assign a function to some LTRs (i.e. promoter or polyA) and revealed that (i) autonomous active LTRs are broadly subjected to operational determinism (ii) the cellular gene density is substantially higher in the surrounding environment of active LTRs compared to silent LTRs and (iii) the configuration of neighboring cellular genes differs between active and silent LTRs, showing an approximately 8 kb zone upstream of promoter LTRs characterized by a drastic reduction in sense cellular genes. These gathered observations are discussed in terms of virus/host adaptive strategies, and together with the methods and tools developed for this purpose, this work paves the way for further HERV transcriptome projects.

Citation: Pérot P, Mugnier N, Montgiraud C, Gimenez J, Jaillard M, et al. (2012) Microarray-Based Sketches of the HERV Transcriptome Landscape. PLoS ONE 7(6): e40194. doi:10.1371/journal.pone.0040194

Editor: Richard Cordaux, University of Poitiers, France

Received February 2, 2012; **Accepted** June 2, 2012; **Published** June 28, 2012

Funding: This work was supported by bioMérieux SA and the French public agency OSEO (Advanced Diagnostics for New Therapeutic Approaches, a French government-funded program dedicated to personalized medicine). PP was supported by doctoral grants from bioMérieux and the Association Nationale de la Recherche et de la Technologie (ANRT); JG and CM were supported by a doctoral fellowship from bioMérieux; Funding for open access charge: bioMérieux SA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: PP, NM, MJ, BB and FM are employees of bioMérieux SA. PP, NM, CM and FM have submitted patent applications covering the findings of this paper. This does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials.

* E-mail: francois.mallet@biomerieux.com

Introduction

The concept of endogenous retroviruses (ERV) dates back to the 1970's and particle-budding observations in the years that followed have gradually provided evidence that mammal genomes serve as reservoirs for retroviral elements [1,2,3]. Later, the sequencing of distinct species unveiled the contribution of the ERV subset within transposable elements (TE), and highlighted in particular a similar proportion of retrovirus-like sequences in human and mouse genomes (8–10%) [4,5,6,7]. The endogenous retrovirus pool is thought to originate from ancestral and independent infections within the germ line [8,9], before complex re-infection, retro-transposition, propagation and error-prone steps occurred during evolution. In humans, the definition of at least 31 HERV families is commonly accepted in reference to putative ancestors [10]. As a result, each family contains tens to thousands of distinct loci scattered throughout the human genome.

To date, all the HERV elements that have been characterized are defective for viral replication. Nevertheless, the discovery that some HERV proteins may contribute to biological events has quickly generated interest in open reading frame (ORF) sequences. The Syncytin-1 and Syncytin-2 envelope glycoproteins are encoded by full-length HERV sequences belonging to the HERV-W and HERV-FRD families, respectively and, through cell differentiation mechanisms, these proteins are presumably essential for human placentation (reviewed in [11]). Syncytin-1 is also associated with epithelial cancers [12,13] and was recently detected in the peripheral blood of leukemia and lymphoma patients [14]. Among the HERV-K HML-2 family, full-length proviruses can encode either Rec or Np9 proteins, which are known to interact with cellular partners and ultimately may affect cancer signaling pathways [15,16,17,18]. Although HERVs match the self-antigen concept, the immune response directed against HERV-K HML-2 Env and Gag proteins is remarkably detectable in the blood of patients with seminoma up to six months before diagnosis [19,20,21,15] and thus may form a basis for molecular tools for early germ cell tumor detection.

However, the role of HERV in human biology should not only be reduced to ORF and putative coding genes (in reference to oncoviruses) since TE also contribute to genome plasticity. Duplication of Alu sequences, recombination, and transduction of LINE elements may have led to multigenic families, gene duplication and exon shuffling [22,23,24]. In particular, the long terminal repeat (LTR) sequences of HERV elements may be the source of inter-element recombination phenomena resulting in chimerical proviruses, tandem structures and solitary LTRs [25,26,27]. Current estimates indicate that the human genome harbors around 200,000 HERVs (excluding MaLR), mainly composed of sequences resembling LTRs [5,28,29]. Taking into account that LTRs exert natural transcription functions within a retrovirus, it is likely that some have now retained the potential to act as regulatory elements [30,31].

In this context, many studies have established a role for LTRs as a promoter [32,33,34,35,36,37,38], bidirectional promoter [39,40], enhancer [41,42], polyadenylation signal [43] and antisense transcript negative modulators [44] of cellular genes in different biological contexts (for a full review see [45]). On the basis of serendipitously and case-by-case identifications, knowledge of functional interactions between HERV elements and cellular environment has gradually grown and is increasingly based on systematic approaches. As different works now estimate that more than 50% of human genes use alternative promoters [46,47], the importance of accurately identifying distinct HERV elements in transcriptome-wide studies, documenting their expression in a variety of biological contexts and finally assessing the question of their regulation in connection with their genomic environment is a strong argument for the need for a HERV transcriptome project [48].

Over the last 10 years, most of the attempts for HERV expression measurement used RT-PCR techniques either to focus on a specific locus [49,50,51,52,53] or to evaluate general trends within HERV families or genera [54,55,56,57]. Yet the inherent limitations in the development of reliable PCR systems to discriminate individual HERV elements in a holistic approach require fairly laborious work [58,48]. On the other hand, methods based on expressed sequence tags (ESTs) provided a more comprehensive view of the HERV transcriptome but generally ran into trouble for identifying the unique genomic source of expression [59,60].

We previously developed an early high-density microarray generation dedicated to the HERV transcriptome, given promising results in terms of tropism and individual locus identification notwithstanding high risks of cross-reactions [61]. Following this attempt, in this work, we introduced a new methodology suitable for repeated element probe design aiming to minimize cross-reactions. At the same time, we expanded the content of the chip to 6 HERV families: HERV-W, HERV-H, HERV-E 4.1, HERV-FRD, HERV-K HML-2 and HERV-K HML-5, providing the user with a collection of 2,690 distinct proviruses (complete or partial) and 2,883 distinct solo LTRs ready for expression monitoring. Additionally, independent probesets within U3 and U5 regions made it possible to assign a function (i.e. promoter or polyA) to 1,513 LTRs. We used this next generation microarray to gain insights into the HERV transcriptome using a composite panel of 40 normal and 39 tumor RNA samples. We found that HERV expression patterns are highly dependent on tissue type and differentiation state and accordingly we established a list of potential HERV biomarkers. We also identified 326 and 209 LTRs with putative promoter and polyA activity, respectively, and highlighted extensive operational determinism for active LTRs. We finally emphasized the trend for promoter LTRs to be associated with an upstream 8 kb zone characterized by a poor sense cellular gene density, compared to silent and polyA LTRs. Taken together, these data allowed us to discuss the adaptive relationship between viruses and host and to prepare a first draft of the HERV transcriptome that could help renew the role of the HERV repertoire in the context of what was improperly named 'junk' DNA.

Results

Detection of the HERV Transcriptome

We constructed a database grouping 10,035 distinct HERV elements that belong to 6 HERV families (Table 1*a*), and we used it as an input collection for the design of a new and suitable HERV-dedicated microarray, called HERV-V2. For this purpose, we developed a scoring function which assesses the ability of a 25-mer probe/target pair to hybridize in Affymetrix-based technology format. This function, referred to as EDA+, allowed us to exclude candidate probes that did not meet specificity criteria. The resulting HERV-V2 chip can discriminate 5,573 distinct HERV elements (23,583 probesets) that can reasonably be assigned to a unique genomic position, including functional U3/U5/gag/pol/env parts, either for provirus structures or solo LTR elements (Table 1*b*).

As an initial view of the HERV transcriptome, we performed a study based on a diversified panel composed of both normal and tumor tissues, including testis, colon, ovary, prostate, breast, uterus, lung and placenta samples. Noteworthy, all samples except placenta are matched normal/tumor tissues obtained from the same individual. The set of data revealed transcriptional activity for 1,718 distinct HERV elements (Table 1*c*), which is about one third of the HERV-V2 chip contents and may suggest a similar proportion of active elements among the human genome. We then sought (i) to determine whether HERV expression varies depending on the tissue and thus follows tropism rules or not, (ii) to find out the extent to which HERV elements are sensitive to the state of differentiation and may serve biomarker research, (iii) to gain insight into transcriptional mechanisms in the light of genomic environment and (iv) to reinforce the comprehensive role of the HERV repertoire in our biology.

Characterization of the HERV Transcriptome

Tropism of Active HERVs. To determine whether the nature of a tissue affects HERV expression, we classified active probesets according to their expression pattern. Although a large proportion shows either no expression or weak unclassifiable signals (data not shown), 10 expression profiles were obtained from partitioning clustering (Figure 1A). The final sizes and the resolving power vary from one profile to another in accordance with data structure. Among the 10 profiles, 2 main types should be distinguished, whether the profile involves only one tissue, or more than one. In profiles 1, 2, 3, 4, 6, 7 and 8, the probesets have a single-tissue expression and consequently can be considered to be tissue-state sensitive. On the other hand, profiles 5, 9 and 10 are dedicated to active probesets expressed in more than one tissue (even being expressed in all tissues such as in profile 10), and thus must reflect a more complex tropism. A detailed list of all HERV loci composing the groups of expression, including genomic coordinates, is provided in Table S3.

In an attempt to unveil a particular behavior in such expression patterns, the number of probesets is summarized taking into account the 6 HERV families (Figure 1B). Interestingly, some profiles coincide with a predominant family representation. This is the case among the colon tumor group (profile 2) where the

Table 1. Detection of the HERV transcriptome.

Repertoire	Elements[e]	HERV-W	HERV-H	HERV-E 4.1	HERV-FRD	HERV-K HML-2	HERV-K HML-5	Total
Genome[a]	**solo LTRs**	**464**	**1079**	**158**	**1259**	**1000**	**87**	**4047**
	complete or partial proviruses	**823**	**1492**	**455**	**349**	**2685**	**184**	**5988**
	5′ LTRs[d]	128	1036	41	36	52	22	1315
	3′ LTRs[d]	219	1062	39	45	2482	22	3869
	gag[d]	199	1093	246	88	117	126	1869
	ppol[d]	234	0	0	96	0	0	330
	pol[d]	0	1315	330	75	155	147	2022
	env[d]	240	1173	67	154	2548	97	4279
Chip[b]	**solo LTRs**	**432**	**553**	**120**	**1189**	**512**	**77**	**2883**
	complete or partial proviruses	**304**	**1354**	**427**	**218**	**215**	**172**	**2690**
	5′ LTRs[d]	120	444	29	33	29	18	673
	3′ LTRs[d]	171	485	29	43	85	19	832
	gag[d]	162	787	228	80	85	125	1467
	ppol[d]	222	0	0	0	0	0	222
	pol[d]	0	1154	307	35	93	135	1724
	env[d]	205	513	63	127	66	97	1071
Transcriptome[c]	**solo LTRs**	**100**	**209**	**30**	**251**	**199**	**19**	**808**
	complete or partial proviruses	**101**	**587**	**91**	**39**	**75**	**17**	**910**
	5′ LTRs[d]	26	154	10	8	10	4	212
	3′ LTRs[d]	43	182	10	11	35	7	288
	gag[d]	12	202	51	4	15	4	288
	ppol[d]	8	0	0	0	0	0	8
	pol[d]	0	170	28	1	9	2	210
	env[d]	49	71	5	16	12	2	155

[a]Number of distinct genomic HERV loci included in HERV database HERV-gDB3. The database contains 6 HERV families with unequal input. The search for distinct elements belonging to each family is performed by systematic BLAST genome coverage, allowing a maximum 20% divergence with prototype elements.
[b]Number of distinct genomic HERV loci present in the chip. Each element of the database is processed through home-made EDA+ algorithm to find probes that match optimal hybridization criteria. The candidate probes are then checked against the entire human genome (NCBI 36/hg18) using the KASH algorithm to control their cross-hybridizing ability and non-specific sequences are removed. Probes are ultimately assembled into probesets to discriminate individual genomic HERV sequences. Differences between database and chip mark the success in designing HERV-specific probes and probesets. For clarity, the probeset content is not detailed.
[c]HERV transcriptome results: number of active elements in all tissues tested. After the experiments were normalized using the COMBAT method and an arbitrary positive threshold was applied (value = 100), elements that are active in at least one tissue are enumerated.
[d]Subsets of complete or partial proviruses.
[e]One element can be composed of several probesets.
doi:10.1371/journal.pone.0040194.t001

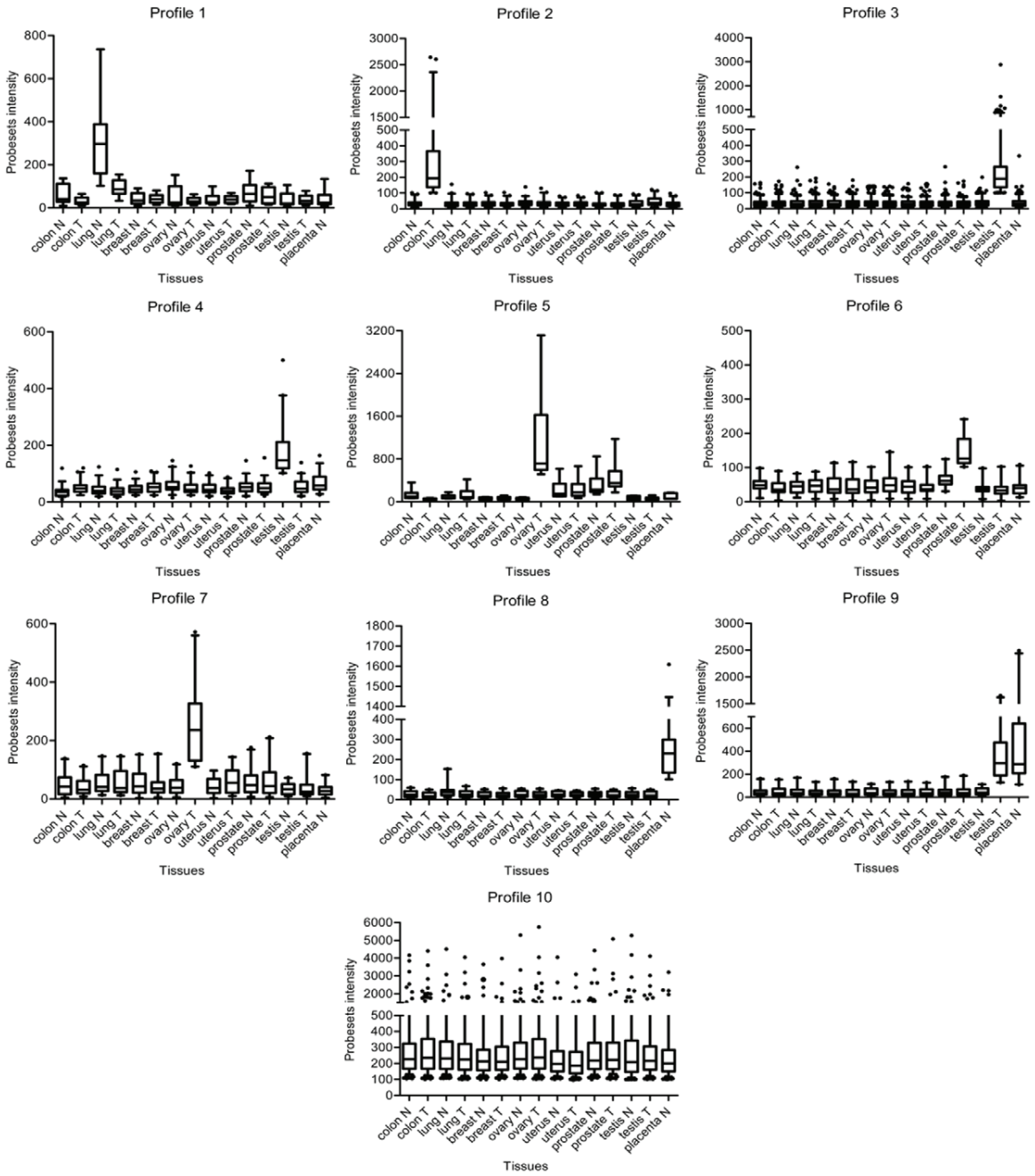
HERV-H family is almost exclusively present, as well as for profile 5 that is entirely described by the HERV-E 4.1 family, and for profile 9 which mostly involves HERV-W probesets.

Differential Expression Associated with Tissue State Changes. To gain insight into the variation of expression associated with tissue state, we performed supervised statistical analysis in pairwise tissues using the SAM method with FDR correction. In order to easily compare the results from the different tissues, we used a high and constant false discovery rate (FDR = 20%). Each paired tissues (i.e. normal tissue versus adjacent tumor tissue) gives an independent number of differentially expressed probesets, ranging from zero (in uterus, data not shown), to 1,092 in testis (Figure 2A). Additionally, the lists are compared to highlight tissue-specific probesets.

To get a better view of the relevance of the results obtained with the SAM-FDR procedure, we drew scatter plots of normal versus tumor expression values for each tissue pair (Figure 2B). The density of plots, the relative amount of tissue-specific probesets (in red) as well as the deviation from the reference straight line together serve to distinguish valuable probesets from non-relevant results. Thereby colon, testis and in a lesser extent ovary and lung involve numerous probesets showing both significant variation of expression and tissue specificity. A list of all the HERV loci that show differential expression together with their associated genomic coordinates is provided in Table S3.

LTR Functions. To approach the question of HERV transcription mechanisms, we focused on LTR signals. In the context of the distribution of a substantial number of retroviral sequences throughout the human genome, we assessed the question of LTR functions regardless of the original provirus structure. Based on the fact that one LTR can theoretically assume different functions depending to its environment, we systematically tested whether the transcription initiates, ends within the LTR, or

A



B

Profile number	Cluster	HERV-W	HERV-H	HERV-E 4.1	HERV-FRD	HERV-K HML-2	HERV-K HML-5	Total
1	lung N	0	7	1	0	1	4	13
2	colon T	0	42	0	0	1	0	43
3	testis T	41*	87	5	32	37	0	202
4	testis N	1	17	4	9	7	1	39
5	ovary T - uterus N - uterus T - prostate N - prostate T	0	0	6	0	0	0	6
6	prostate T	5	7	8	0	1	0	21
7	ovary T	1	13	5	0	2	0	21
8	placenta N	7	2	2	11	0	0	22
9	placenta N - testis T	17	0	1	1	1	0	20
10	all tissues	40*	324	18	56	76	6	520

Figure 1. Tropism of active HERVs. (**A**) Active probesets 'cluster' into 10 expression profiles. The final number of profiles is estimated after iterative corrections combining Euclidean partitioning algorithm and fine manual adjustment steps. Box plots indicate the distribution and the

median of probeset intensity, whiskers are 5–95 percentiles, dots show outliers. The order of profiles is not important. (**B**) Profile description. Each profile refers to a specific cluster of tissues, and involves a number of probesets detailed by families. By definition, a probeset is classified in a unique profile, except for the asterisk (*) where a single probeset is willingly shared by both profiles 3 and 10.
doi:10.1371/journal.pone.0040194.g001

A

Paired tissues N *vs* T	HERV-W	HERV-H	HERV-E 4.1	HERV-FRD	HERV-K HML-2	HERV-K HML-5	Total
colon	34 (19)	219 (90)	21 (6)	37 (16)	37 (16)	6 (2)	354 (149)
lung	31 (18)	161 (61)	29 (13)	36 (20)	38 (15)	10 (4)	305 (131)
breast	3 (0)	28 (5)	4 (1)	0	5 (0)	1 (1)	41 (7)
ovary	43 (25)	318 (147)	60 (31)	87 (52)	95 (60)	15 (9)	618 (324)
prostate	13 (8)	40 (16)	23 (11)	3 (2)	20 (6)	3 (1)	102 (44)
testis	140 (110)	540 (344)	56 (28)	149 (121)	193 (134)	14 (7)	1092 (744)

B

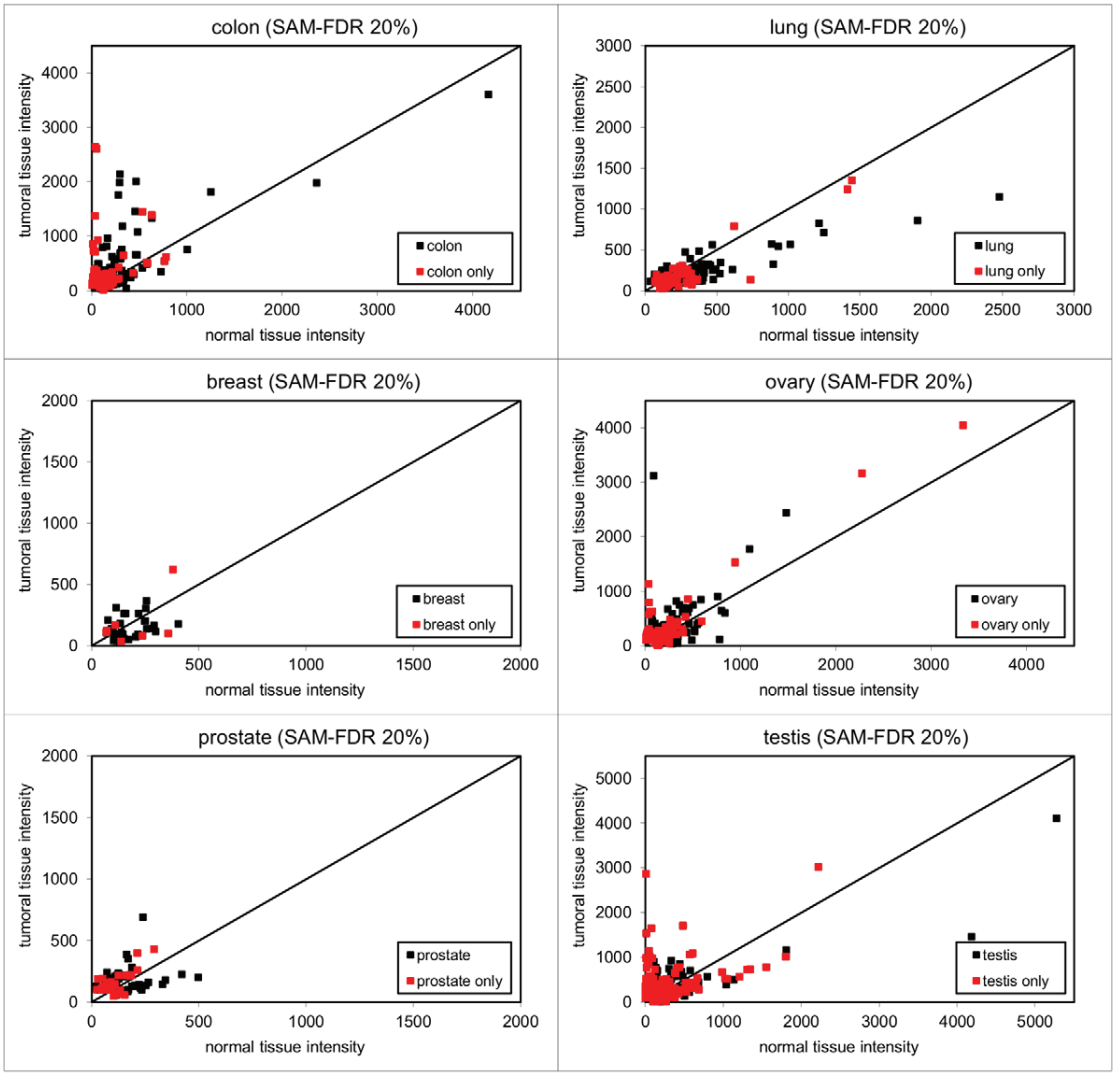


Figure 2. Differential expression induced by tissue state changes. (**A**) Pairwise analysis. For each paired tissue, the SAM-FDR method is applied and leads to the identification of a number of probesets that show significant differential expression (FDR = 20%). The red number in brackets indicates how many differential expressed probesets are specific to the tissue. Uterus normal versus tumor comparison gives no result and consequently does not appear in the table. (**B**) Scatter plots of expression values. Normal versus tumor normalized expression values of differential expressed probesets are draw for each tissue pair. The statistically significant absence of differential expression is represented by the diagonal line (y = x). Red plots refer to tissue-specific probesets (previously mentioned using red numbers in bracket in Figure 2A).
doi:10.1371/journal.pone.0040194.g002

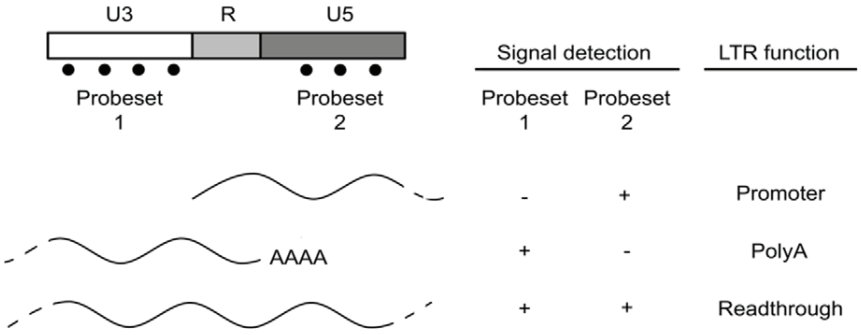
passes through the element with no incidence on the transcription process (Figure 3A).

We used the dichotomy of signals acquired from probesets distributed along U3 and U5 regions to assign the functions: U3-associated negative signals and U5-associated positive signals for promoter, U3 positive signals and U5 negative signals for polyA, U3 and U5 double positive signals for readthrough. Loci which exhibited double negative U3 and U5 signals were classified as silent (expression level cutoff = 50 for negative signal). Expression levels between 50 and 100 delineate an indeterminate grey area where a function is assigned only if the ratio between U3 and U5 signals is greater than 4 (see legend of Figure 3A for details). Due to the general LTR sequence homology and the large share of partial and complex structures, only a small fraction of LTRs meet the requirement to infer a function. These LTRs are referred to as 'attributable LTR' - aLTR in the text - in Figure 3B and represent one third of the chip LTR content.

Of all the tissue samples tested, we finally identified a total of 326 distinct autonomous 'promoter' LTRs (21% of aLTR) and 209 distinct 'polyA' LTRs (13% of aLTR). Very surprisingly, there is no overlap between these two LTR lists except one which belongs to the grey area. This highlights that active LTRs cannot switch from promoter to polyA function even if the tissue changes, which we will refer to as operational determinism. To enhance this opinion we repeated LTR function analysis using a completely different set of data coming from cell lines that were subject to chemical and oncogenic transformations (data not shown). Using 6 different cell lines, we found a parallel list of 67 promoter and 46 polyA LTRs and still no overlap between promoter and polyA LTRs exists. Moreover, the cell culture-derived list closely matches the results from tissues: among the 113 (67 promoters +46 polyA) active LTRs unambiguously characterized from the cell culture, only 7 LTRs did not intersect with the 535 LTR list (326 promoters +209 polyA) characterized from tissues. This means that less than 2% of new characterizations have been gained by diversifying biological records. Consequently, this result prompts us to conclude that we have delineated a stable pool of active and functional LTRs.

Most of the function characterizations concern solo LTRs with 247 promoter solo LTRs and 151 polyA solo LTRs (26% of aLTR), but the function distribution seems to be unbalanced between families: although there is generally a low number of output cases, we observed for instance that the HERV-K HML-5 family has no occurrence of polyA solo LTRs, and we noted that the HERV-K HML-2, HERV-W, HERV-E 4.1 and HERV-FRD families have more promoter solo LTRs cases than polyA solo

A



B

	HERV-W			HERV-H			HERV-E 4.1			HERV-FRD			HERV-K HML-2			HERV-K HML-5			Total		
	Solo LTR	*5′ LTR*	*3′ LTR*	*Solo LTR*	*5′ LTR*	*3′ LTR*	*Solo LTR*	*5′ LTR*	*3′ LTR*	*Solo LTR*	*5′ LTR*	*3′ LTR*	*Solo LTR*	*5′ LTR*	*3′ LTR*	*Solo LTR*	*5′ LTR*	*3′ LTR*	*Solo LTR*	*5′ LTR*	*3′ LTR*
Genome [a]	464	128	219	1079	1036	1062	158	41	39	1259	36	45	1000	52	2482	87	22	22	4047	1315	3869
Chip [b]	432	120	171	553	444	485	120	29	29	1189	33	43	512	29	85	77	18	19	2883	673	832
Attributable LTR [c]	274	34	45	112	60	77	22	9	6	500	16	18	199	13	41	58	16	13	1165	148	200
promoter	57	9	14	21	13	13	7	3	1	76	2	3	69	6	10	17	1	4	247	34	45
polyA	40	6	2	29	15	19	4	2	1	53	2	0	25	2	8	0	1	0	151	28	30
readthrough	0	0	1	5	4	3	1	0	0	6	0	0	3	0	0	1	0	1	16	4	5
silent	109	14	21	43	24	33	7	4	1	281	11	12	64	3	12	21	7	5	525	63	84

Figure 3. LTR functions. (**A**) Schematic view of LTR structure and associated theoretical transcription events. Top to bottom: the LTR is a natural or alternative promoter when the transcription starts between U3 and R/U5; the LTR ends an upstream transcription event by the addition of polyA tail at the end of the R region; the transcription passes through the LTR with no incidence in the progression of the polymerase, which results in the detection of U3, R and U5 transcripts. Rules for function assignment are *promoter*: U3−/U5+; *polyA*: U3+/U5-; *readthrough*: U3+/U5+ and *silent*: U3−/U5-; with expression levels: + >100; - <50. Expression levels between 50 and 100 delineate an indeterminate grey area where a function is assigned if the ratio between U3 and U5 is greater than 4 (for instance U3 = 80 and U5 = 321 is counted as promoter). Otherwise, the LTR function is declared to be unknown. (**B**) Assignment of functions. [a,b,c] Loss of information from HERV database to understandable functions. [a] Summary of Table 1*a* [b] Summary of Table 1*b* [c] Enumeration of LTRs whose function is attributable, i.e. defined as LTR combining both complete structure on the genome and existing probesets on the chip, that can ultimately allow a discrimination between U3 and U5 expression signals.
doi:10.1371/journal.pone.0040194.g003

LTRs cases, whereas the HERV-H family, on the contrary, seem to involve a greater amount of polyA solo LTRs compared to promoter solo LTRs. Focusing on provirus structures, we identified 34 promoter 5′LTRs and 30 polyA 3′LTRs. In some cases, we associated both 5′ promoter and 3′ polyA activities within a given provirus. We also discovered 45 promoter 3′LTRs and 28 polyA 5′LTRs.

Besides these 34% comprehensive active aLTR, we also showed that a high proportion of the LTR population always remains silent (672 cases; 44% of aLTR). In addition to that, we identified a smaller number of readthrough LTRs (25 cases; <2% of aLTR).

Validation Analyzes. We first compared our results with previously published data focusing on the HML-2 family as this family has been widely studied using various methodologies. We included data derived from EST study [59], genomic repeat expression monitoring (GREM) for experimental genome-wide identification of promoter-active repetitive elements [62], PCR-sequencing [48] and array-based approaches [61]. Of the 327 HML-2 elements we analyzed, 25 elements were shared by at least one of the previous studies (Table S4). On this subset, Affymetrix-based format analysis gave 64% and 63% correlation with the EST approach and the PCR-sequencing-based study, respectively. A poor correlation of 19% was observed with GREM.

To confirm the tropism of active HERV, we then tested whether the elements we classified within expression groups using HERV-V2 correlate with tissue-related EST libraries. A fairly clear enrichment of the expected EST population was observed in the case of colon, ovary and placenta and can also reasonably be claimed in the case of testis taking into account that testis-associated HERV sequences were initially distributed into 3 expression groups (Table S5). In contrast, results depicted for lung and prostate were not supported by ESTs, probably due to an overall less-pronounced expression level of related HERV elements. We then picked 33 candidate loci and designed PCR primer pairs which were evaluated for sequence specificity using high resolution melting and sequencing (see Materials and Methods, RT-PCR). Eighteen highly specific primer pairs corresponding to 8 loci were eventually selected and tested on samples (Table S6). An overall good correlation of 0.926 (min 0.606; max 0.998) between arrays and RT-PCR was observed, essentially confirming the attributed tropism (Figure S2). Nevertheless, unexpected expression was found twice for two HERV-H proviruses, in cancerous colon in addition to the expected expression in tumor testis, and in cancerous ovary in addition to the expected expression in tumor colon.

The LTR functions were assessed using U3 versus U5 RT-PCR assays. Using this strategy, we previously validated the promoter function of 6 loci expressed in testicular cancer identified by the first version of the HERV microarray [61], which was confirmed in this study (data not shown). Such a strategy was used again and confirmed two new tropism-related promoters (200261_w and 1100414_2) as well as one ubiquitous promoter (2000062_2) as presented in Figure S3. Then, to broaden the scope of such analysis, we sought to confirm LTR functions by analyzing the U3 versus U5 distributions of LTR-associated ESTs for a subset of the HERV-W family consisting of 21 proviruses and 110 solo LTRs or LTRs associated with truncated proviruses. We focused on the HERV-W family because it contains the ERVWE1 domesticated locus in which the 5′LTR promoter and the 3′LTR polyA functions have been exhaustively demonstrated [63,64,65,66,61]. Results are depicted in Table S7 and alignments are provided in Figure S4**.** In brief, only 17 loci among the 131 loci analyzed exhibited significant LTR-associated ESTs and only 16 loci are ultimately interpretable. 8 EST-deduced functions (7 promoters, 1 polyA) were consistent with those we identified following microarray results. Two other promoter functions were compatible with an upstream alternative transcription initiation site (see locus 1200505_w and locus 600462_w in Table S7). One additional promoter function was plausible (locus 400207_w) although an alternative splicing event excluding U3 could be involved. Finally, one readthrough identified using microarray (locus 700126_w) could be classified either as polyA or readthrough with regard to EST data. Four comparisons were discrepant, opposing readthrough to promoter function, and putatively identifying a removal of the U3 region in mRNAs due to a splice occurrence. Altogether, the overall correlation between array and EST-deducible functions ranged between 50% and 75%.

Influence of the Genomic Environment. We extended our investigation to the genomic environment encompassing the newly-identified functional and silent LTRs. For each LTR, we performed a search for gene presence and %GC content in the surrounding 50 kb, starting from the limits of the LTR. When the position of the LTR overlaps with the position of the gene, the LTR is counted as intronic. The total number of neighboring genes normalized by the initial number of LTRs gives a gene density ratio detailed for each category (Figure 4A). The gene density ratio is almost 1.5 times higher for active LTRs than for silent LTRs although the %GC barely varies. Meanwhile, the proportion of intronic LTRs is largely in favor of antisense representation for all categories of LTRs.

The case of intergenic LTRs is subject to a more detailed description in Figure 4B. For promoter, polyA and silent LTR groups, a cumulative gene distribution function is drawn upstream (5′) and downstream (3′) of the LTR limits (vertical bar) emphasizing whether the genes found away from the LTR have the same orientation as the LTR (sense) or not (antisense). This revealed a strikingly low occurrence of genes in sense orientation up to 8 kb upstream of promoter LTRs while the upstream 8 kb for silent and polyA LTRs shows no difference regarding the gene orientation. Besides, the downstream environment also appears to be linked to the LTR function but in a kind of mirror situation in which the sense genes occurrence apparently rises faster than for antisense genes in the downstream 8 kb zone of silent LTRs compared to promoter and polyA LTRs.

Lessons Learned from the HERV Transcriptome. The different results were finally used to construct a comparative view of HERV genome and transcriptome. To achieve this goal, we used the term 'HERV genome' to refer to the entirety of our HERV genome database content (i.e. 6 HERV families), and we opposed the HERV transcriptome resulting from our experiments (Figure 5). Since the HERV-V2 content reflects the success in designing specific probes and probesets, which varies from one family and one element to another, we had to apply correction factors to raw transcriptome results. Accordingly, the following outcome must be regarded as an extrapolation.

The first observation tends to show there is no difference between the contribution of each family to genome and transcriptome sharing (Figure 5A). However, the transcription seems to be impacted by the structure of HERV elements in a trend that aims to reduce proviral gene expression (30% to 16%) (Figure 5B). More flagrantly, the genomic environment appears to exert a major influence as the expression of HERV elements that map close to human genes (<10 kb) is twice constricted and, at the same time, the expression of intronic HERV sequences in sense orientation reduces dramatically (9% to 3%) (Figure 5C). Focusing on LTRs and regardless of the tissue tested, almost 50% of the LTR elements remain silent, while active LTRs are roughly

A

LTR function	nb of LTR	nb of neighbouring genes (+/- 50kb)	Mean % GC content (+/- 50kb)	Ratio genes/LTR	Intronic LTR		Intergenic LTR
					sense	antisense	
promoter	326	304	40.8	0.93	6	32	288
polyA	209	177	40.7	0.85	7	18	184
readthrough	25	23	42.0	0.92	1	2	22
silent	672	432	40.1	0.64	7	26	639

B

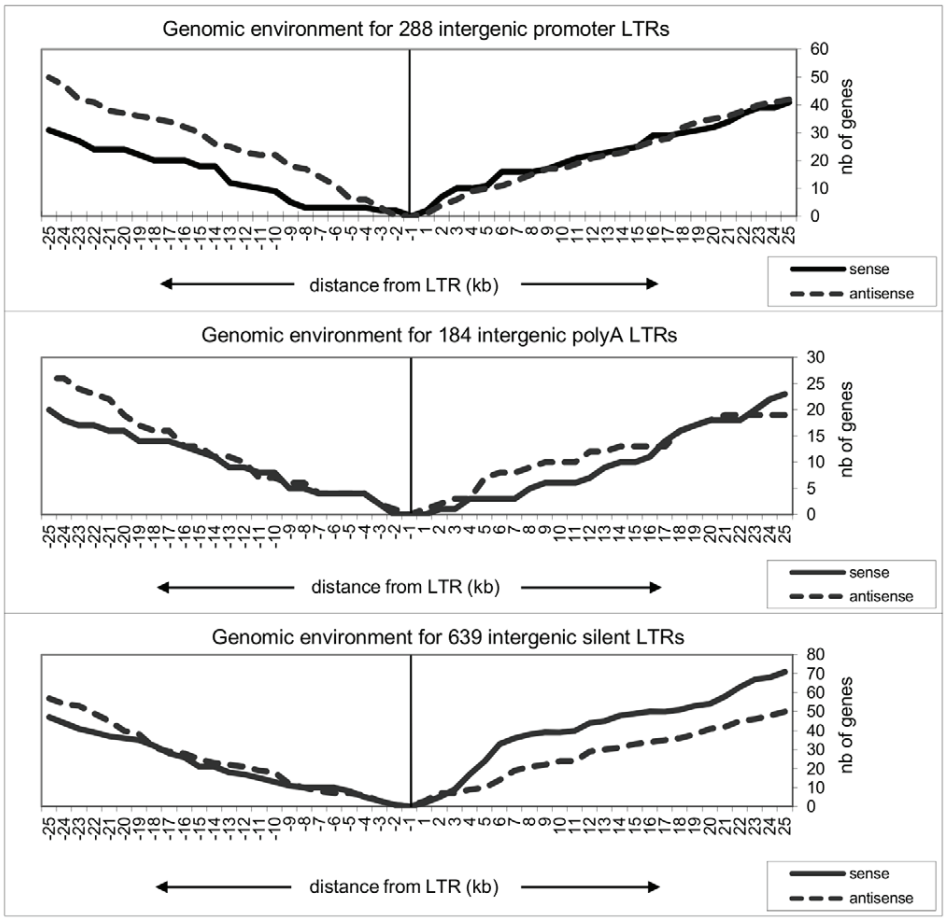


Figure 4. Genomic environment of functional and silent LTRs. (**A**) Overview of genomic and chromatin composition (%GC) of functional LTR neighborhood. For all promoter, polyA, readthrough and silent LTRs, the number of neighboring genes in the surrounding +/−50 kb is obtained from NCBI 36/hg18 using annotations from the RefGene table (UCSC), then the DNA sequences are extracted *in silico* for %CG content calculation. The table includes the number of intronic functional LTRs, defined as LTRs that overlap gene limits (NCBI 36/hg18 RefGene table), and ends with the number of intergenic LTRs. *Sense*: LTR and gene are in the same orientation; *antisense*: LTR and gene are in opposition. (**B**) Genomic environment for intergenic functional LTRs. Genes in the same orientation (*sense*) or in opposition (*antisense*) with the LTRs are counted in the case of promoter, polyA and silent intergenic LTRs. Read-through LTRs are not included as their number, which is too low, does not fit with the representation. Vertical bar centered on zero should be interpreted as an ellipse of the LTR sequence. Away from the bar, the cumulative gene occurrence is shown up to +/−25 kb starting from the LTR limits. Curve tendencies beyond 25 kb do not change significantly and are not represented.
doi:10.1371/journal.pone.0040194.g004

equally divided between promoter and polyA functions (Figure 5D). We also identified very few cases of readthrough LTRs (3%).

Discussion

HERV Transcriptome Views

We provided a microarray-based description of the HERV transcriptome based on the analysis of a set of cancerous and non-cancerous tissues that reflect a range of diversity. Different works have been conducted to discover the contribution of HERV to the human transcriptome [38,60,59,62,48,61]. In this study we identified 1,718 active HERV elements suggesting that about 30% of the retroviral sequences spread across the genome are transcribed. Despite the fact that it is usually thought that HERVs colonize the genome and consequently are tightly controlled to avoid gene disruption [67], our observation of a substantial basal HERV transcriptional activity is partly supported by others. In 2008 Conley *et al.* analyzed high-throughput expression data to claim that transcribed HERV sequences correspond to 1.16% of

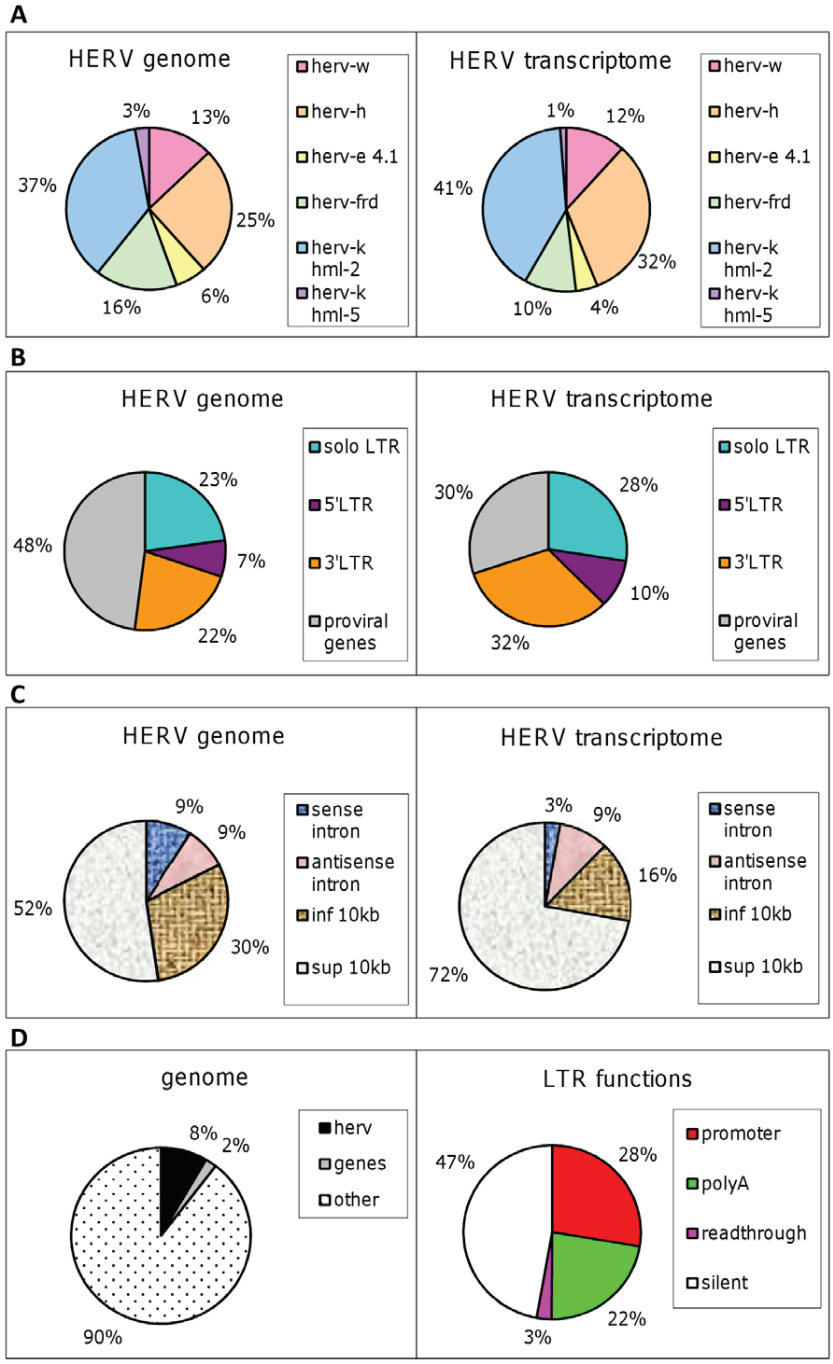


Figure 5. Genomic and transcriptomic projections of the HERV repertoire. (**A**) HERV families. The 6 HERV families studied in this work are voluntarily depicted as 100% of HERV human genome, in the proportions described in Table 1*a*. The transcriptome picture is obtained from results detailed in Table 1*c* after applying a correction factor that takes into consideration chip content in Table 1*b*. (**B**) HERV structures. Solo LTR and proviruses account for 100% of the HERV genome in the proportion described in Table 1*a*. The transcriptome part is based on Table 1*c* after correction taking into consideration the chip content presented in Table 1*b*. (**C**) HERV environment. A systematic search for genomic environment is performed for elements present in the HERV database (genome) and the active elements described in Table 1*c* (transcriptome). The proportion of the 3 types of LTR is based on Table 1*a* and the transcriptome from Table 1*c* after correction using Table 1*b*. (**D**) The role of junk DNA. HERV sequences represent approximately 8% of the human genome. The graph of LTR functions is based on Figure 3B after a correction based on the number of attributable functions and chip content.
doi:10.1371/journal.pone.0040194.g005

the human genome sequence [38], which would mean that approximately 15% of HERV sequences are active. Previous analyses of HERV activity based on ESTs led Oja and colleagues to estimate that 7% of the HERV sequences are transcribed [68]. More generally, the fact that the human genome might be more or less pervasively transcribed, including sequences previously

thought to be silent, was a key outcome of the ENCODE pilot project and led to the proposal of the 'warehouse' concept for natural selection [69]. This suggests how HERV may regulate human transcription on a large scale.

EST data appear to be insufficient to describe the transcriptional activity of HERVs and therefore to unambiguously characterize promoter functions, as previously discussed [60,48]. Moreover, for the most active HERV elements, Oja reported hundred to thousand-fold over-representation of *pol* and *env* regions (as opposed to LTRs). Such poor EST detection in either 5′ or 3′ LTRs could be due to the nature of the EST methodology which may be sensitive to low level of expression or end-location of secondary structured LTRs on mRNA or even the occurrence of polypurine tracks within retrovirus genes [70]. Notably, EST strategy failed to identify ERVWE1/Syncytin-1 3′LTR as a polyA signal as discerned using HERV-V2, although the full-length Syncytin-1-containing polyadenylated cDNA has been isolated [63]. Nevertheless, focusing on the 22 well-described HML-2 elements we shared with the EST study conducted by Stauffer, we obtained a correlation of 64%. Similarly, the circumscribed EST analysis conducted on HERV-W elements confirmed up to 75% of our promoter elements.

HERV Tropism and Implication in the Biomarkers Field

The HERV transcriptome presented herein was generated using a set of tissues selected in order to support the hypothesis that individual HERV can serve the biomarker field. Among cancerous and non-cancerous tissues, we characterized expression patterns supporting that testis and placenta are privileged places of HERV expression. Syncytin-1, a functional envelope glycoprotein belonging to the HERV-W family, is expressed in the placenta and in the testis [63,71,49,61]. Syncytin-2, a member of the HERV-FRD family, takes part in the placenta expression cluster [72,49,73] and numerous envelope and capsid elements related to the HERV-K HML-2 family formed the testicular tumor group as described previously [2,74,75,19,76]. In a recent work, we reported the expression of 6 HERV-W elements in testicular tumor using an early version of HERV chip [61]. This second generation of the HERV chip allowed to confirm the overexpression of 5 out of 6 elements (the 6^{th} locus belongs to the grey area as defined above) and, at the same time, we identified numerous new HERV-W elements specific to the testicular cancer sample with high expression levels. The association of HERV-H elements with colon cancer [59,77,78,79] and the finding of HERV-E 4.1 sequences in a group composed of prostate, uterus and ovary samples has also been reported [61] and is confirmed here. Taken together, these findings argue in favor of non-random behavior of HERV elements and families and thus suggest a strong HERV tropism acting within human organs.

In line with this idea, we focused on differential expression between normal and tumor tissues in pairwise analyses. The use of SAM-FDR gold standard statistical tests [80] led to the identification of a variable number of elements that are sensitive to the state of differentiation. We took the responsibility of false-positive results using a high FDR value but we also assumed that, by using a test with low stringency, we did not miss any interesting elements. Testis here again appears to be the most predisposed context to HERV differential activity with more than 1,000 DEP composed of almost two-thirds of tissue-specific probesets. Notably we highlighted a significant number of probesets with strong and specific expression variation between normal and cancerous colon samples. The RT-PCR experiments we set up to validate HERV tropism and differential expression showed that HERV-V2 overall trends are accurate. Nevertheless, discrepancies between micro-array and RT-PCR have also been observed, which may reflect a lower sensitivity of the chip as opposed to RT-PCR, e.g. due to the intrinsic sensitivity of the whole transcriptome amplification or to a target-dependent unbalanced amplification. For ovary and lung analysis, although the number of DEP seems impressive, only a few probesets deviate from low values. In addition, we did observe variable levels of genomic DNA contamination within lung samples, which may have biased the result of analysis. Altogether, although promising, the transfer of these results into biomarkers will require further clinical studies based on relevant dedicated procedures [81], notably taking into account inter-individual variations.

Specialization of Human LTR Function

After a retrovirus has integrated the host genome, its two flanking LTR sequences are strictly identical, yet the alteration of HERV structures and the genetic drift over time may provide a favorable context for both natural and alternative LTR functions. As a result in the current human genome, the estimated 200,000 HERV LTRs can be seen as a wild collection of promoter and polyA elements. Based on this concept, we identified 326 promoter LTRs, 209 polyA LTRs, 25 readthrough LTRs and 672 silent LTRs among the 1513 evaluated LTRs. Confirmation analysis based on HERV-W-associated ESTs revealed that putative splicing events excluding U3 regions occurred in some cases, which may lead to an overestimation of promoter functions. Conversely, we did not assign promoter functions to LTRs lacking probes in U3 but exhibiting high positive signals in U5. In particular, we identified 34% of active HML-2 promoters. This is slightly less than the GREM experimental method that showed at least 50% of HERV-K HML-2 LTR serve as in vivo promoters [62,36]. Some of the elements identified with GREM were found in our study but it is somewhat disturbing to find only a poor correlation (19%). This could be due to inter individual variations among tissue samples in both studies, as only one testicular parenchyma was used to implement the GREM methology [62]. Alternatively, given that GREM is a PCR-based method, the analysis of transcribed HERV sequences can be more sensitive than with microarrays but conversely can be complicated by recombination events during PCR [58].

Most of the function characterizations concern solo LTRs (398 out of 535; 74%). In detail, we characterized 247 promoter solo LTRs and 151 polyA solo LTRs. If we look at solo LTRs regardless of their family, we are inclined to consider that these structures, originating from recombination phenomena, are more likely to exert promoter rather than polyA functions. However, the relative amount of promoter and polyA solo LTRs varies remarkably from one family to another. Within the HERV-K HML-5 family, we only characterized promoter solo LTRs. The HERV-W, HERV-E 4.1, HERV-FRD and HERV-K HML-2 families similarly showed a predominant set of solo LTRs with promoter functions. It is noteworthy that among the 6 HERV families we studied, the oldest, HERV-H, gives the most significant example of polyA solo LTR overrepresentation. The observed biases in solo LTR specialization may result from an intrinsic property of the natural history of each family, as exemplified in a different context by the LINE-1-mediated spreading of a significant proportion of the HERV-W family [82]. Alternatively we cannot exclude an orientated and irreversible genetic drift within LTR sequences. Further functional comparative analysis of evolutionary-conserved solo LTRs may permit to address these hypotheses.

We also examined the 5′ end of the 45 promoter 3′LTR elements. The proportion of 5′-truncated structures in this subset

is not higher compared to other HERV proviruses. However, when a function can be attributed, the existing 5′LTR is silent. This observation can suggest a loss of fixation of transcription factors. Indeed, different works on proto-oncogene activation induced by retrovirus insertion have showed that the 3′LTR can initiate alternative transcription of cellular genes only if the insertion was accompanied by an inactivation of the 5′LTR of the provirus [83], a concept referred to as promoter occlusion [84]. Thus, the description of 45 promoter 3′LTRs in this study appears consistent with the concept of promoter occlusion.

Astonishingly, promoter and polyA lists have no LTR in common, a strong trend we called operational determinism. This was observed using both the 79 normal et cancerous tissue panel and the 6 cell lines. Thus, despite environmental changes over time, active LTRs seem to feature unique specialized functions. Nevertheless, HERV-W-associated ESTs showed that in some contexts, only a readtrough phenomenon can replace or be added to promoter or polyA function. This finding is compatible with operational determinism but suggests the presence of weak promoter or polyA activities. In addition, attempts to validate LTR functions by leveraging EST data have faced the possibility of alternative transcription initiations. Indeed, alternative initiation sites have been proposed for the promoter of ERVWE1 following mung bean nuclease protection assays [65]. These two alternative sites are located 71 bp and 75 bp upstream from the site we defined by RACE as the R border [63,61], respectively. Moreover, due to genetic drift, the location of initiation sites within HERV LTRs may be more flexible than for exogenous retroviruses.

HERV Functions and Genomic Environment

Gene density in the environment of active promoter LTRs is significantly higher than for silent LTRs as previously observed for the HML-2 family [36]. Notably, this behavior was also shared by LTRs exhibiting polyA function. Such observations could be interpreted in two ways: either chromosomal regions with high transcriptional activity promote HERV activity as a side effect (e.g.: bringing transcription factors together with DNA strand opening), or there is a functional contribution of active LTRs to human gene regulation in a way that would be of benefit to the genome. Conversely, exclusion of methylated silent LTRs from gene-rich regions preclude methylation spreading and then silencing of conventional genes as previously suggested for transposable elements [85]. The set of 99 intronic LTR elements investigated here presented a 3.7 fold bias in favor of antisense-oriented insertion, similar to the 2 to 4.5 range previously described [86,87]. As previously proposed, this suggests a strong selection against LTR elements in the sense direction and consequently argues that LTRs found in the same transcriptional orientation are much more likely to have a detrimental effect [87]. It is noteworthy that the antisense orientation bias appears similar for silent and transcriptionally active LTRs. Regarding surrounding genes, this may reflect an overall weak transcriptional activity as observed for a set of proviruses and solo LTRs belonging to the HERV-W family [88]. Alternatively this could represent substantial and therefore gene-independent transcription events in altered cellular contexts.

Among the 1133 intergenic LTR elements, 288 (25%) were promoter LTRs, 184 (16%) polyA LTRs and 639 (56%) silent LTRs. Comparison of the gene environment of those intergenic LTRs highlighted two points. Unexpectedly, an approximate 8 kb interval upstream of intergenic promoter LTRs was characterized by a drastic under-representation of sense genes. This result was considered relevant due to the significant number of LTRs (n = 228) and the absence of LTR-associated multigene families which may skew the results. This suggests that a sense-intergenic promoter LTR can only survive at a certain distance of a sense gene, otherwise it would have a detrimental effect on the gene. Such a location may contribute to the usage of acceptor donor sites together with alternative polyA signal which may alter the original transcript as proposed for intronic elements [87]. Second, a mirror situation consisting in an 8 kb window was observed upstream from silent LTRs, showing a decrease in antisense genes compared to sense genes. Although no obvious explanation can be provided to date, it is striking to note that such a symmetrical 8 kb region was recently shown to correspond to the maxima of LTR density around transcription start site of tissue-specific genes [89].

Conclusion

This microarray-based approach unveiled the expression of 1,718 distinct HERV loci and identified 326 promoter LTRs and 209 polyA LTRs in a broad range of tissues. Further systematic quantitative analysis is required to gain insight on the relative variation of expression of HERV sequences and their adjacent cellular genes. In particular, looking at different stages of cell differentiation may accelerate the identification of alternative promoters as already documented for a subset of genes in the mouse embryo [90].

In addition to the preservation of transcription factor binding sites, two important features determining the control of HERV expression consist of the LTR methylation status [91,92,93,94,61,95] and the chromatin context associated to post-translational histone modifications [95]. Locus-specific LTR hypomethylation was observed both during placental development [91,92,93] and in testis and colon cancers [94,61,95]. Thus, such whole transcriptome approach together with LTR function identification and further characterization of associated epigenetic marks may help to discriminate between *statu quo*, conflict and cooperation, the components of a many-facetted relationship between retrotransposons and their metazoan hosts.

Materials and Methods

Chip Design

HERV database. A database for genomic HERV elements was constructed following a 4-step process: (i) for each HERV family, we defined a prototype by choosing the most representative and complete HERV element present in the human genome. (ii) Functional U3/U5/gag/pol/env parts were labeled on the prototypes. (iii) These sequences were then used as an input reference library for RepeatMasker [96] (see the details of the prototypes in Table S1). The search for HERV functional sequences was extended to the entire human genome (NCBI 36/hg18) allowing a maximum 20% divergence with prototype sequences. (iv) The functional sequences identified were lastly assembled into annotated HERV elements and were implemented in an owner database, so-called HERVgDB3. HERVgDB3 contains 10,035 distinct HERV elements belonging to 6 HERV families, including complete and partial proviruses (Table 1*a*).

Probeset Design

The probe design steps aimed ultimately to define probesets for the functional parts of each HERV element that belongs to HERVgDB3. We first generated all possible and overlapping 25-mer tracks for any given HERV sequence of HERVgDB3, leading to an initial pool of candidate probes. We then evaluated the cross-hybridization risk of each candidate probe using local alignment versus the entire human genome (NCBI 36/hg18) as a model of hybridization, supported by an internally developed alignment

scoring function called EDA+. The EDA+ principle is based on instability induced by any mismatch within the hybridization between probe and target. Using EDA+, the impact of mismatches is cumulative and modulated regarding their type, their position and the size of the interval between two mismatches. A threshold on the cumulative weight is then defined to consider the hybridization as probable or not. Note that no specific thermodynamic parameter was added to the model. The relevance of this score was evaluated independently (data not shown). EDA+ was applied to any local alignment between a candidate probe and the human genome, computed using the KASH algorithm [97]. Probes that meet the alignment-EDA+ criteria were definitively selected to enter the design process. This last step finally grouped the selected probes in order to constitute probesets for any given functional part of the HERV elements collection. When more than 10 probes can be used to create a probeset, we make a selection to obtain a homogeneous distribution of probes along the functional part.

Custom HERV GeneChip Microarray

The custom HERV GeneChip integrates 23,583 HERV probesets (88,592 probes) and can discriminate 5,573 distinct HERV elements, composed of complete and partial proviruses (Table 1*b*). In addition to the HERV repertoire, a set of mismatch declinations (37,200 probes), initially based on 19 perfect match (PM) probesets belonging to the commercial Affymetrix HG_U133_PLUS2 chip, serves to evaluate and improve the EDA+ hybridization scoring function (data not shown). The standard Affymetrix control probes for unbiased amplification and hybridization were also included in the microarray.

Sample Description

Tissue samples and cell lines. Matched-pair tumor/normal RNA samples of colon (3), breast (8), ovary (3), uterus (3) and prostate (1) were purchased from Clinisciences. Additional First Choice human tumor/normal RNA samples of colon (1), ovary (1), uterus (1), testis (1), lung (1) and prostate (2), plus normal placenta sample (1), were obtained from Life Technologies. The Centre de Ressources Biologiques of Nancy provided epidermoid carcinoma and normal adjacent lung RNA samples (9) and the Centre Hospitalier Lyon-Sud performed macro dissections on radical prostatectomy specimens (5) to isolate cancer tissue from normal tissue. Details on samples are provided in Table S2.

The human prostate epithelial cell line RWPE1 and the chemically stressed-derived WPE1-NA22, WPE1-NB14, WPE1-NB11, WPE1-NB26 [98] as well as the v-Ki-Ras-transformed RWPE2 [99] cell lines were obtained from the CelluloNet of the UMS3444/US8 BioSciences Gerland Lyon-Sud.

Ethical considerations. The human tissue specimens provided by the Centre Hospitalier Lyon-Sud and by the Centre de Ressources Biologiques of Nancy were obtained in compliance with the ICH-GCP regulations, current European and French legislations. A 'non-interventional' biomedical research protocol for tissue samples conservation after a prostate surgery has been set-up at the Centre Hospitalier Lyon-Sud with the approval of the Ethics Committee in Lyon (CPP Sud-Est 2). Therefore, patients admitted to the urology department in the Centre Hospitalier Lyon-Sud were informed and gave voluntary, signed informed consent prior to any tissue sample conservation and for research use. Patients admitted to the Nancy Hospital were informed that their sample tissue after the lung surgery will be conserved at the Centre de Ressources Biologiques de Nancy for research use according to the French bioethics law (2004). Clinisciences and Life Technologies signed an agreement to ensure that the tissue samples were obtained in compliance with ICH-GCP standards.

Molecular Biology Analysis

RNA extraction. RNA was extracted from macro-dissected radical prostatectomies following the Trizol protocol (Invitrogen) and was purified on Rneasy columns (Qiagen). The quality of all RNA samples was assessed with the Bioanalyser 2100 capillary electrophoresis device using the RNA Nano Chips kit (Agilent).

Target amplification, labeling and microarray hybridization. cDNA synthesis and amplification were performed using 50 ng of RNA, using the WT-Ovation RNA Amplification System kit (Nugen). Briefly, amplification was initiated both at the 3′ end and randomly throughout the whole transcriptome, and this was followed by reverse transcriptase/RNAse H mix step before SPIA linear and single strand amplification. Amplified ssDNA products were purified using the QIAquick purification kit (Qiagen), total DNA concentration was measured using the NanoDrop 1000 spectrophotometer (Termo Scientific) and the product quality was checked on the Bioanalyser 2100. Two micrograms of purified ssDNA were fragmented into 50–200 bp fragments by DNAseI treatment and were 3′-labeled using a terminal transferase recombinant kit (Roche). The resulting target was mixed with standard hybridization controls and B2 oligonucleotides following the recommendations of the supplier. The hybridization cocktail was heat-denatured at 95°C for 2 minutes, incubated at 50°C for 5 minutes and centrifuged at 16,000 g for 5 minutes to pellet the residual salts. The HERV GeneChip microarrays were prehybridized with 200 µl of hybridization buffer and placed under stirring (60 rpm) in an oven at 50°C for 10 minutes. The hybridization buffer was then replaced by the denatured hybridization cocktail. Hybridization was performed at 50°C for 18 hours in the oven under constant stirring (60 rpm). Washing and staining were carried out according to the protocol supplied by the manufacturer, using a fluidic station (GeneChip fluidic station 450, Affymetrix). The arrays were scanned using a fluorometric scanner (GeneChip scanner GS 3000, Affymetrix).

Real-time PCR. A set of locus-specific PCR primers was designed using Primer3 and the NCBI Primer-BLAST software and then checked *in silico* at UCSC (http://genome.ucsc.edu). Primers were ordered from Eurogentec. For each individual PCR system, a range of amplifications, followed by High Resolution Melting (HRM) analysis and product sequencing, was performed on genomic DNA to control the specificity of the products and to determine optimal experimental melting temperature (Tm). For each tissue, individual samples were pooled in order to compare results from RT-PCR with the data from microarrays. 50 ng of total RNA of each sample were DNAse-treated and reverse-transcribed using the QuantiTec Reverse Transcription Kit (Qiagen). Reverse-transcriptase-free reactions were carried out to verify the absence of contaminating genomic DNA. SYBR green experiments were set up using the Type-it HRM PCR kit (Qiagen) in 10 µL final reaction volume with 5 µM primers and a 20-fold dilution of the cDNA. PCR amplifications were carried out in Rotor-disc 100 wrapped discs devised for the Rotor Gene Q (Qiagen). Housekeeping genes G6PD, GAPDH and HPRT were analyzed in the same experiment as the target transcripts. Amplifications of cDNA were performed as follows: a 5-min denaturation step at 95°C, followed by 45 cycles (95°C for 10 s, Tm for 30 s, 72°C for 10 s) and HRM analysis (from 65°C to 95°C, 0.1°C increments every 2 s) to control the product purity. Each reaction was performed in duplicate. The second derivative method was used to assess the amplification efficiency (Eff).

Relative expression (RLE) for each system is $Eff^{\Delta Ct}$ ($\Delta Ct = Ct_{min} - Ct_{sample}$). All data were normalized by the geometric mean of the RLE of the three housekeeping genes.

Bioinformatics

Chip analysis. The quality control (QC) of the microarrays was assessed using the standard Affymetrix controls to verify that the chips met the criteria. In addition, the dataset was explored to highlight unexpected batch effects and to correct them before statistical analysis. The distributions of intensities for probes and probesets were plotted to test different putative covariate effects (e.g.: dates of amplification and hybridization, people in charge of the experiment, lots of reagents). The following representations were used: the log intensity value distribution (density plots and box plots), the median absolute deviation (MAD) versus the intensity median (MAD-Med plots), the background plots and nuse plots and finally the relative log expression (RLE) plots. A strong batch effect related to the experimental operator was identified, as well as residual batch effects related to the amplification dates within each operator. A customized pre-processing strategy was thus selected to correct these technical and undesired effects.

The data pre-processing included a background correction based on the tryptophan probe baseline signal, followed by normalization and summarization steps involving a double batch effect correction. In brief, the background of each chip was estimated as the 15^{th} percentile of the intensity values of the tryptophan probes, then the robust microarray averaging (RMA) process [100] was applied within each operator batch using the configuration of quantile normalization followed by median polish summarization. This process was applied independently for each amplification date batch. After that, a two-step combining batches (COMBAT) method [101] was applied, first within each operator dataset in order to merge the date effects of a given operator, and second within the entire chips set in order to merge the operator datasets together (see Figure S1). The COMBAT method constructs a model for each gene, formally written as:

$$Y_{ijg} = \mu_g + X_j\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}$$

Y_{ijg} is the signal measured for the gene g when the sample j is processed in the batch I; μ_g is a mean expression level for the gene g; we considered a single biological covariate X (here a qualitative variable including the origin and the state of the tissue); β_g defines the level of differential expression related to biological categories (the parameter we are looking for); parameters γ_{ig} and δ_{ig} are the additive and multiplicative error components that define the batch i (they are gene-specific) and ε_{ijg} is the error that follows $N(0, \sigma_g)$.

After all the chips were normalized, expression values of individual chips were grouped into sets of samples as described in Table S2. If no precision is given, all the results illustrated and discussed in this study are based on the values of the sets of samples.

Partitioning clustering was applied to the normalized expression values using a Euclidean distance function algorithm to determine similarities between observations. The final number of clusters was decided after iterative corrections combining algorithm auto-decisions and fine adjustments through direct observations of the resulting dataset arrangement. The minimum number of probesets required to form an expression cluster was empirically set at 6.

The search for differentially expressed genes (DEG) implied a classical significant analysis of microarray (SAM) procedure [80] followed by a false discovery rate (FDR) correction [102]. The dataset was filtered to exclude the probesets for which expression values were less than 2^6 in all tissues. A FDR cut-off of 20% was applied.

Genomic environment. Homemade perl scripts were developed to request and extract information from the human genome build NCBI 36/hg18 and the RefGene annotation table (UCSC). Gene density and %GC content calculation were evaluated by default in the +/−50 kb surrounding environment, starting from the HERV element ends.

Expressed Sequenced Tag (EST) analysis. The blastn algorithm (NCBI blastn v.2.2.25) was used to compare HERV sequences to EST libraries. A cut-off of 97% was retained as a compromise between the extreme similarity existing between loci of the same family and the polymorphism in the human population, ranging from 1 out of 0.31 kb in repeats to 1 out of 1.8–2.0 kb in coding regions (Nickerson, 1998}. If no precision is given, the default parameters used for alignment were: alignment length >200 bp; EST/sequence alignment coverage >85%.

Software and data. QC, pre-processing and DEG analysis were performed using R statistical software [103], packages from the Bioconductor project [104] and homemade R packages. The clustering algorithm used for this study is implemented in Partek Genomics Suite 6.5. Geneious 5.0 was used for primer design and EST analysis. The complete experimental set comprises 113 microarrays. Affymetrix data files (.cel) are available upon request.

Supporting Information

Figure S1 Effect of RMA-COMBAT normalization. Distribution of intensities within the dataset before (upper part) and after (lower part) RMA-COMBAT normalization. Each boxplot represents a single chip and the colors refer to experimental batches.
(PDF)

Figure S2 Correlations between microarray and RT-PCR results. Normalized values of microarray and RT-PCR experiments are given for 12 independent HERV sequences that belong to 8 distinct HERV loci. Correlations close to 1 indicate a strong positive linear relationship and therefore confirm the findings. Correl = $Cov_{microarray;RT\text{-}PCR}/(sd_{microarray} * sd_{RT\text{-}PCR})$.
(PDF)

Figure S3 RT-PCR analyses of LTR promoter functions. The promoter activity of 3 independent LTRs was evaluated in RT-PCR. Relative expression of U5 *vs* U3 is given by $Fc_{U5/U3} = (Eff_{U3}^{CtU3})/(Eff_{U5}^{CtU5})$. Values greater than 1 indicate a promoter activity. An asterisk (*) highlights tissues for which the promoter activity has been unequivocally found using the microarrays. In the particular case of 1100414_2 no probeset was defined within the LTR and consequently the promoter activity in testicular tumor could not be detected using microarrays.
(PDF)

Figure S4 Pictures illustrating alignments of HERV-W loci with their best EST counterpart. Each alignment is designated by the name of the locus as it stands on the microarray, followed by the name of the most similar EST. The alignment explicitly states the retroviral structure including LTR U3, R and U5 subdomains, as well as flanking regions. Probes defined on the array are indicated by grey arrows. The sequence used for the query is represented as well as the EST retained for analysis, as developed in Table S7. Accession number and EST count are shown. Arbitrary blue numbering of HERV subdomain and the aligned EST together with blue vertical bars are indicated when required to facilitate the reading, e.g. clones overlapping U5 and 5′ flanking region for 400207_w-AI738459.jpg. Best score EST

aligned with 5′ (700341_w-ERVWE1_5LTR.jpg) and 3′ (700341_w-ERVWE1_3LTR.jpg) LTRs of the ERVWE1 locus are included to highlight the limits of information provided by ESTs.
(PDF)

Table S1 HERV prototypes used for the construction of HERV-gDB3. Accession numbers, genomic localizations and the limits of the functional region within the prototype sequences (U3, R, U5, gag, pol env) are given for the 6 HERV families studied. [a] for HERV identification and gene cutting out. [b] for LTR sub region cutting out.
(PDF)

Table S2 Biological samples included in the study. List of biological samples included in the study (samples) and used in the composition of analysis groups (set of samples). Information on pathological status, age and sex are provided when available. Matched tumoral/normal samples are indicated (paired with). An asterisk (*) highlights samples that were not used for the microarray study.
(PDF)

Table S3 Genomic coordinates of active and functional HERV sequences. Genomic coordinates refer to the human genome version NCBI 36/hg18. Each HERV locus is designated by a single identifier (locus id). The table summarizes the different observations mentioned in the study, i.e. whether the locus shows expression patterns (tropism), is differentially expressed between normal and cancer samples (DGE) or exhibits functional LTRs (LTR functions). Two 'x' in the "LTR functions" box associated with one locus reflect distinct functions for each LTR of the same provirus.
(XLS)

Table S4 Identification of HML-2 repetitive elements characterized by independent methods. From left to right the genomic location (NCBI 36/hg18), the individual HML-2 locus sequence name, the tropism of expression deduced from the microarrays, the differential expression and the LTR functions as depicted in Table S3, the references from which data were obtained taking into consideration either EST analysis [59], genomic repeat expression monitoring (GREM) for experimental genome-wide identification of promoter-active repetitive elements [62], PCR-sequencing [48] or array-based approach [61] are given. The original designation of the HERV loci is given for each study. We added April 2012 EST query information obtained using the method developed in Table S7. Statistics concerning this analysis are given at the bottom of the table and include, for each study, the number of elements, the number of shared elements, the number of active elements and the correlation between our work and each individual study.
(XLS)

Table S5 Matching of tissue-specific HERV sequences with Expressed Sequenced Tag (EST) databases. The CleanEST database [105] was used to retrieve ESTs associated with tissues of interest in order to construct 6 reference EST groups: colon (311122 ESTs), lung (441913 ESTs), ovary (123944 ESTs), placenta (321881 ESTs), prostate (69860 ESTs) and testis (264243 ESTs). Each EST group was blasted against the HERV sequences composing the expression profiles shown in Figure 1, following the procedure detailed in the EST analysis part of the materials and methods section. Hits were normalized by the total number of HERV loci of the expression profile and by the total number of ESTs forming the reference group. The ranking of the value is associated with a color code highlighting the enrichment of tissue-associated ESTs: green (1/6), yellow (2/6) and red (>2/6).
(PDF)

Table S6 Primers used for RT-PCR experiments. Forward and reverse primer sequences used for RT-PCR analyses. The Tm of each primer pair was determined as described in the related materials and methods section. The domain of application is indicated (normalization, tropism, promoter function).
(PDF)

Table S7 Identification of Expressed Sequenced Tags (ESTs) putatively associated with active HERV-W repetitive elements. We used Megablast to compare HERV sequences to EST libraries using Geneious 5.0 software and NCBI libraries. A cut-off of 97% was retained as a compromise between the extreme similarity existing between loci of the same family and the polymorphism in the human population ranging from 1 out of 0.31 kb in repeats to 1 out of 1.8–2.0 kb in coding regions [106]. From left to right, the genomic location (NCBI 36/hg18), the individual HERV-W locus sequence name, the tropism of expression deduced from the microarrays, the differential expression and the LTR functions as depicted in the Table S3, the LTR associated structure (i.e: provirus, solo LTR, partial provirus with either 5′ or 3′ LTR), the EST scores, the reference accession numbers of the ESTs, the EST length in bp, the EST coverage of the LTR query (i.e: 100% or numbering when <100%), the LTR-element covered regions (i.e: U3, R, U5, gag, pol, env, 5′ or 3′ flanking region), the information concerning the additional coverage of clones and the existence of additional clones in flanking regions, the previous identification and designation of the locus, the EST-associated proposed function, and the name of the pictures illustrating alignments of HERV loci with their best EST counterpart as detailed in Figure S4 are given. Parameters used for Megablast query with Geneious 5.0 are indicated at the bottom of the table, as well as statistics concerning the query and a color code highlighting the correlation between array and EST LTR-deduced functions.
(XLS)

Acknowledgments

We are grateful to Dr. Alain Ruffion and Dr. Myriam Decaussin-Petrucci from the Centre Hospitalier Lyon Sud for providing us with specimens of radical prostatectomies, and to Dr. Nadine Martinet and Stéphanie Lacomme from the Centre de Ressources Biologiques of Nancy for giving us access to postoperative lung samples. We also wish to thank Laurent Duret for his advice, Amandine Campan for her contribution to databases, Hader Haidous for his guidance on ethical considerations and Isabelle Grosjean from the CelluloNet of the UMS3444/US8 BioSciences Gerland Lyon-Sud.

Author Contributions

Conceived and designed the experiments: FM. Performed the experiments: PP JG CM. Analyzed the data: PP NM FM. Contributed reagents/materials/analysis tools: FM BB JG NM CM MJ PP. Wrote the paper: PP FM.

References

1. Boller K, Frank H, Lower J, Lower R, Kurth R (1983) Structural organization of unique retrovirus-like particles budding from human teratocarcinoma cell lines. J Gen Virol 64 (Pt 12): 2549–59.

2. Boller K, Konig H, Sauter M, Mueller-Lantzsch N, Lower R, et al. (1993) Evidence that HERV-K is the endogenous retrovirus sequence that codes for the human teratocarcinoma-derived retrovirus HTDV. Virology 196: 349–53.

3. Lyden TW, Johnson PM, Mwenda JM, Rote NS (1994) Ultrastructural characterization of endogenous retroviral particles isolated from normal human placentas. Biol Reprod 51: 152–7.
4. Smit AFA (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 9: 657–663.
5. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
6. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science (New York, N Y) 291: 1304–1351.
7. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–62.
8. Bannert N, Kurth R (2004) Retroelements and the human genome: new perspectives on an old relation. Proc Natl Acad Sci U S A 101 Suppl 2: 14572–14579.
9. Bannert N, Kurth R (2006) The evolutionary dynamics of human endogenous retroviral families. Annu Rev Genomics Hum Genet 7: 149–73.
10. Katzourakis A, Tristem M (2004) Phylogeny of Human Endogenous and Exogenous Retroviruses. In:Sverdlov E, editors. Retroviruses and Primate Genome Evolution. Georgetown: Landes Bioscience. 1–18.
11. Pérot P, Montgiraud C, Lavillette D, Mallet F (2011) A Comparative Portrait of Retroviral Fusogens and Syncytins. In: Larsson LI, editors. Cell Fusions: Regulation and Control. Springer Netherlands. 63–115.
12. Bjerregaard B, Holck S, Christensen IJ, Larsson LI (2006) Syncytin is involved in breast cancer-endothelial cell fusions. Cell Mol Life Sci 63: 1906–1911.
13. Strick R, Ackermann S, Langbein M, Swiatek J, Schubert SW, et al. (2007) Proliferation and cell-cell fusion of endometrial carcinoma are induced by the human endogenous retroviral Syncytin-1 and regulated by TGF-beta. J Mol Med 85: 23–38.
14. Sun Y, Ouyang DY, Pang W, Tu YQ, Li YY, et al. (2010) Expression of syncytin in leukemia and lymphoma cells. Leukemia research 34: 1195–1202.
15. Boese A, Sauter M, Galli U, Best B, Herbst H, et al. (2000) Human endogenous retrovirus protein cORF supports cell transformation and associates with the promyelocytic leukemia zinc finger protein. Oncogene 19: 4328–36.
16. Armbruester V, Sauter M, Roemer K, Best B, Hahn S, et al. (2004) Np9 protein of human endogenous retrovirus K interacts with ligand of numb protein X. J Virol 78: 10310–9.
17. Denne M, Sauter M, Armbruester V, Licht JD, Roemer K, et al. (2007) Physical and functional interactions of human endogenous retrovirus proteins Np9 and rec with the promyelocytic leukemia zinc finger protein. J Virol 81: 5607–16.
18. Kaufmann S, Sauter M, Schmitt M, Baumert B, Best B, et al. (2010) Human endogenous retrovirus protein Rec interacts with the testicular zinc-finger protein and androgen receptor. The Journal of general virology 91: 1494–1502.
19. Sauter M, Schommer S, Kremmer E, Remberger K, Dolken G, et al. (1995) Human endogenous retrovirus K10: expression of Gag protein and detection of antibodies in patients with seminomas. J Virol 69: 414–21.
20. Boller K, Janssen O, Schuldes H, Tonjes RR, Kurth R (1997) Characterization of the antibody response specific for the human endogenous retrovirus HTDV/HERV-K. J Virol 71: 4581–8.
21. Goedert JJ, Sauter ME, Jacobson LP, Vessella RL, Hilgartner MW, et al. (1999) High prevalence of antibodies against HERV-K10 in patients with testicular cancer but not with AIDS. Cancer Epidemiol Biomarkers Prev 8: 293–6.
22. Rearden A, Magnet A, Kudo S, Fukuda M (1993) Glycophorin B and glycophorin E genes arose from the glycophorin A ancestral gene via two duplications during primate evolution. J Biol Chem 268: 2260–7.
23. Schwartz A, Chan DC, Brown LG, Alagappan R, Pettay D, et al. (1998) Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. Hum Mol Genet 7: 1–11.
24. Long M (2001) Evolution of novel genes. Curr Opin Genet Dev 11: 673–80.
25. Johnson WE, Coffin JM (1999) Constructing primate phylogenies from ancient retrovirus sequences. Proc Natl Acad Sci U S A 96: 10254–60.
26. Hughes JF, Coffin JM (2001) Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat Genet 29: 487–9.
27. Hughes JF, Coffin JM (2005) Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. Genetics 171: 1183–94.
28. Paces J, Pavlicek A, Paces V (2002) HERVd: database of human endogenous retroviruses. Nucleic Acids Res 30: 205–206.
29. Belshaw R, Watson J, Katzourakis A, Howe A, Woolven-Allen J, et al. (2007) Rate of recombinational deletion among human endogenous retroviruses. J Virol 81: 9437–42.
30. Brosius J (1999) Genomes were forged by massive bombardments with retroelements and retrosequences. Genetica 107: 209–238.
31. Jern PCoffin JM (2008) Effects of retroviruses on host genome function. Annual review of genetics 42: 709–732.
32. Samuelson LC, Wiebauer K, Gumucio DL, Meisler MH (1988) Expression of the human amylase genes: recent origin of a salivary amylase promoter from an actin pseudogene. Nucleic Acids Res 16: 8261–76.
33. Medstrand P, Landry JR, Mager DL (2001) Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. J Biol Chem 276: 1896–903.
34. Dunn CA, Medstrand P, Mager DL (2003) An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon. Proc Natl Acad Sci U S A 100: 12841–6.
35. van de Lagemaat LN, Landry JR, Mager DL, Medstrand P (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 19: 530–6.
36. Buzdin A, Kovalskaya-Alexandrova E, Gogvadze E, Sverdlov E (2006) At least 50% of human-specific HERV-K (HML-2) long terminal repeats serve in vivo as active promoters for host nonrepetitive DNA transcription. J Virol 80: 10752–62.
37. Romanish MT, Lock WM, van de Lagemaat LN, Dunn CA, Mager DL (2007) Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. PLoS Genet 3: e10.
38. Conley AB, Piriyapongsa J, Jordan IK (2008) Retroviral promoters in the human genome. Bioinformatics 24: 1563–7.
39. Dunn CA, Romanish MT, Gutierrez LE, van de Lagemaat LN, Mager DL (2006) Transcription of two human genes from a bidirectional endogenous retrovirus promoter. Gene 366: 335–342.
40. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, et al. (2009) The regulated retrotransposon transcriptome of mammalian cells. Nat Genet 41: 563–571.
41. Ruda VM, Akopov SB, Trubetskoy DO, Manuylov NL, Vetchinova AS, et al. (2004) Tissue specificity of enhancer and promoter activities of a HERV-K(HML-2) LTR. Virus Res 104: 11–6.
42. Prudhomme S, Oriol G, Mallet F (2004) A retroviral promoter and a cellular enhancer define a bipartite element which controls env ERVWE1 placental expression. J Virol 78: 12157–68.
43. Mager DL, Hunter DG, Schertzer M, Freeman JD (1999) Endogenous retroviruses provide the primary polyadenylation signal for two new human genes (HHLA2 and HHLA3). Genomics 59: 255–63.
44. Gogvadze E, Stukacheva E, Buzdin A, Sverdlov E (2009) Human-specific modulation of transcriptional activity provided by endogenous retroviral insertions. J Virol 83: 6098–6105.
45. Cohen CJ, Lock WM, Mager DL (2009) Endogenous retroviral LTRs as promoters for human genes: a critical assessment. Gene 448: 105–114.
46. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM (2006) Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. Genome research 16: 1–10.
47. Kimura K, Wakamatsu A, Suzuki Y, Ota T, Nishikawa T, et al. (2006) Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. Genome research 16: 55–65.
48. Flockerzi A, Ruggieri A, Frank O, Sauter M, Maldener E, et al. (2008) Expression patterns of transcribed human endogenous retrovirus HERV-K(HML-2) loci in human tissues and the need for a HERV Transcriptome Project. BMC Genomics 9: 354.
49. de Parseval N, Lazar V, Casella JF, Benit L, Heidmann T (2003) Survey of human genes of retroviral origin: identification and transcriptome of the genes with coding capacity for complete envelope proteins. J Virol 77: 10414–22.
50. Wang-Johanning F, Frost AR, Jian B, Azerou R, Lu DW, et al. (2003) Detecting the expression of human endogenous retrovirus E envelope transcripts in human prostate adenocarcinoma. Cancer 98: 187–97.
51. Smallwood A, Papageorghiou A, Nicolaides K, Alley MK, Jim A, et al. (2003) Temporal regulation of the expression of syncytin (HERV-W), maternally imprinted PEG10, and SGCE in human placenta. Biol Reprod 69: 286–93.
52. Okahara G, Matsubara S, Oda T, Sugimoto J, Jinno Y, et al. (2004) Expression analyses of human endogenous retroviruses (HERVs): tissue-specific and developmental stage-dependent expression of HERVs. Genomics 84: 982–90.
53. Buscher K, Trefzer U, Hofmann M, Sterry W, Kurth R, et al. (2005) Expression of human endogenous retrovirus K in melanomas and melanoma cell lines. Cancer Res 65: 4172–80.
54. Seifarth W, Frank O, Zeilfelder U, Spiess B, Greenwood AD, et al. (2005) Comprehensive analysis of human endogenous retrovirus transcriptional activity in human tissues with a retrovirus-specific microarray. J Virol 79: 341–52.
55. Forsman A, Yun Z, Hu L, Uzhameckis D, Jern P, et al. (2005) Development of broadly targeted human endogenous gammaretroviral pol-based real time PCRs Quantitation of RNA expression in human tissues. J Virol Methods 129: 16–30.
56. Muradrasoli S, Forsman A, Hu L, Blikstad V, Blomberg J (2006) Development of real-time PCRs for detection and quantitation of human MMTV-like (HML) sequences HML expression in human tissues. Journal of virological methods 136: 83–92.
57. Pichon JP, Bonnaud B, Mallet F (2006) Quantitative multiplex degenerate PCR for human endogenous retrovirus expression profiling. Nat Protoc 1: 2831–8.
58. Flockerzi A, Maydt J, Frank O, Ruggieri A, Maldener E, et al. (2007) Expression pattern analysis of transcribed HERV sequences is complicated by ex vivo recombination. Retrovirology 4: 39.
59. Stauffer Y, Theiler G, Sperisen P, Lebedev Y, Jongeneel CV (2004) Digital expression profiles of human endogenous retroviral families in normal and cancerous tissues. Cancer immunity : a journal of the Academy of Cancer Immunology 4:2.

60. Oja M, Peltonen J, Blomberg J, Kaski S (2007) Methods for estimating human endogenous retrovirus activities from EST databases. BMC Bioinformatics 8 Suppl 2: S11.
61. Gimenez J, Montgiraud C, Pichon JP, Bonnaud B, Arsac M, et al. (2010) Custom human endogenous retroviruses dedicated microarray identifies self-induced HERV-W family elements reactivated in testicular cancer upon methylation control. Nucleic Acids Res 38: 2229–2246.
62. Buzdin A, Kovalskaya-Alexandrova E, Gogvadze E, Sverdlov E (2006) GREM, a technique for genome-wide isolation and quantitative analysis of promoter active repeats. Nucleic Acids Res 34: e67.
63. Blond JL, Beseme F, Duret L, Bouton O, Bedin F, et al. (1999) Molecular characterization and placental expression of HERV-W, a new human endogenous retrovirus family. J Virol 73: 1175–85.
64. Prudhomme S, Oriol G, Mallet F (2004) A retroviral promoter and a cellular enhancer define a bipartite element which controls env ERVWE1 placental expression. J Virol 78: 12157–68.
65. Cheng YH, Richardson BD, Hubert MA, Handwerger S (2004) Isolation and characterization of the human syncytin gene promoter. Biol Reprod 70: 694–701.
66. Mallet F, Bouton O, Prudhomme S, Cheynet V, Oriol G, et al. (2004) The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. Proc Natl Acad Sci U S A 101: 1731–6.
67. Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284: 601–603.
68. Ehlhardt S, Seifert M, Schneider J, Ojak A, Zang KD, et al. (2006) Human endogenous retrovirus HERV-K(HML-2) Rec expression and transcriptional activities in normal and rheumatoid arthritis synovia. J Rheumatol 33: 16–23.
69. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816.
70. Hu C, Saenz DT, Fadel HJ, Walker W, Peretz M, et al. (2010) The HIV-1 central polypurine tract functions as a second line of defense against APOBEC3G/F. Journal of Virology 84: 11981–11993.
71. Mi S, Lee X, Li X, Veldman GM, Finnerty H, et al. (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403: 785–9.
72. Blaise S, de Parseval N, Benit L, Heidmann T (2003) Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. Proc Natl Acad Sci U S A 100: 13013–8.
73. Blaise S, de PN, Heidmann T (2005) Functional characterization of two newly identified Human Endogenous Retrovirus coding envelope genes. Retrovirology 2: 19.
74. Lower R, Lower J, Tondera-Koch C, Kurth R (1993) A general method for the identification of transcribed retrovirus sequences (R-U5 PCR) reveals the expression of the human endogenous retrovirus loci HERV-H and HERV-K in teratocarcinoma cells. Virology 192: 501–11.
75. Lower R, Tonjes RR, Korbmacher C, Kurth R, Lower J (1995) Identification of a Rev-related protein by analysis of spliced transcripts of the human endogenous retroviruses HTDV/HERV-K. J Virol 69: 141–9.
76. Armbruester V, Sauter M, Krautkraemer E, Meese E, Kleiman A, et al. (2002) A novel gene from the human endogenous retrovirus K expressed in transformed cells. Clin Cancer Res 8: 1800–7.
77. Liang Q, Ding J, Xu R, Xu Z, Zheng S (2009) Identification of a novel human endogenous retrovirus and promoter activity of its 5′ U3. Biochem Biophys Res Commun 382: 468–72.
78. Pichon JP, Bonnaud B, Cleuziat P, Mallet F (2006) Multiplex degenerate PCR coupled with an oligo sorbent array for human endogenous retrovirus expression profiling. Nucleic Acids Res 34: e46.
79. Liang Q, Xu Z, Xu R, Wu L, Zheng S (2012) Expression Patterns of Non-Coding Spliced Transcripts from Human Endogenous Retrovirus HERV-H Elements in Colon Cancer. PLoS ONE 7: e29950.
80. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America 98: 5116–5121.
81. Ptolemy AS, Rifai N (2010) What is a biomarker? Research investments and lack of clinical integration necessitate a review of biomarker terminology and validation schema. Scandinavian journal of clinical and laboratory investigation Supplementum 242: 6–14.
82. Costas J (2002) Characterization of the intragenomic spread of the human endogenous retrovirus family HERV-W. Mol Biol Evol 19: 526–33.
83. Cullen BR, Lomedico PT, Ju G (1984) Transcriptional interference in avian retroviruses–implications for the promoter insertion model of leukaemogenesis. Nature 307: 241–245.
84. Rabson AB, Graves BJ (1997) Synthesis and processing of viral RNA. In:Coffin JM, Hughes SH, Varmus HE, editors. Retroviruses. New-York: Cold Spring Harbor laboratory press. 205–261.
85. Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome research 19: 1419–1428.
86. Martin U, Steinhoff G, Kiessig V, Chikobava M, Anssar M, et al. (1999) Porcine endogenous retrovirus is transmitted neither in vivo nor in vitro from porcine endothelial cells to baboons. Transplant Proc 31: 913–914.
87. van de Lagemaat LN, Medstrand P, Mager DL (2006) Multiple effects govern endogenous retrovirus survival patterns in human gene introns. Genome Biology 7: R86.
88. Li F, Nellaker C, Yolken RH, Karlsson H (2011) A systematic evaluation of expression of HERV-W elements; influence of genomic context, viral structure and orientation. BMC Genomics 12: 22.
89. Jjingo D, Huda A, Gundapuneni M, Marino-Ramirez L, Jordan IK (2011) Effect of the transposable element environment of human genes on gene length and expression. Genome biology and evolution 3: 259–271.
90. Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, et al. (2004) Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Dev Cell 7: 597–606.
91. Matouskova M, Blazkova J, Pajer P, Pavlicek A, Hejnar J (2006) CpG methylation suppresses transcriptional activity of human syncytin-1 in non-placental tissues. Exp Cell Res 312: 1011–20.
92. Reiss D, Zhang Y, Mager DL (2007) Widely variable endogenous retroviral methylation levels in human placenta. Nucleic Acids Res 35: 4743–54.
93. Gimenez J, Montgiraud C, Oriol G, Pichon JP, Ruel K, et al. (2009) Comparative Methylation of ERVWE1/Syncytin-1 and Other Human Endogenous Retrovirus LTRs in Placenta Tissues. DNA Res 16: 195–211.
94. Wentzensen N, Coy JF, Knaebel HP, Linnebacher M, Wilz B, et al. (2007) Expression of an endogenous retroviral sequence from the HERV-H group in gastrointestinal cancers. Int J Cancer 121: 1417–23.
95. Trejbalova K, Blazkova J, Matouskova M, Kucerova D, Pecnova L, et al. (2011) Epigenetic regulation of transcription and splicing of syncytins, fusogenic glycoproteins of retroviral origin. Nucl Acids Res 39: 8728–8739.
96. Smit AFA, Hubley R, Green P (1996) RepeatMasker Open-3.0.
97. Navarro G, Raffinot M (2002) Flexible Pattern Matching in Strings: Practical On-Line Search Algorithms for Texts and Biological Sequences.Cambridge University Press.
98. Webber MM, Quader ST, Kleinman HK, Bello-DeOcampo D, Storto PD, et al. (2001) Human cell lines as an in vitro/in vivo model for prostate carcinogenesis and progression. The Prostate 47: 1–13.
99. Bello D, Webber MM, Kleinman HK, Wartinger DD, Rhim JS (1997) Androgen responsive adult human prostatic epithelial cell lines immortalized by human papillomavirus 18. Carcinogenesis 18: 1215–1223.
100. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England) 4: 249–264.
101. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics (Oxford, England) 8: 118–127.
102. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 100: 9440–9445.
103. Team RDC (2008) R: A language and environment for statistical computing.
104. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.
105. Lee BShin G (2009) CleanEST: a database of cleansed EST libraries. Nucl Acids Res 37: D686–D689.
106. Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, et al. (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat Genet 19: 233–240.

II.4. Application clinique : recherche de marqueurs diagnostiques et pronostiques du cancer de la prostate à l'aide de puces à ADN à partir de prélèvements urinaires

II.4.1. Rationnel

Est-il techniquement faisable de réaliser des études transcriptomiques basées sur des puces à ADN à partir des ARN extraits d'un recueil de cellules de prostate dans des urines post massage prostatique ? La mise au point des étapes pré-analytiques d'un protocole adapté implique notamment un travail sur : (i) le recueil des urines de patients, (ii) les rendements d'extraction et la qualité des acides nucléiques extraits, (iii) l'adaptation, si nécessaire, des différentes étapes du protocole de réalisation de puces ADN Affymetrix et (iv) l'évaluation de la reproductibilité générale des étapes pré-analytiques et de détection. Deux études méthodologiques ont été menées dans ce but. La première s'est attachée à comparer plusieurs procédés de recueil de cellules urinaires ainsi que plusieurs kits d'extraction d'ARN afin de déterminer la meilleure combinaison de procédés permettant d'aller jusqu'à l'utilisation d'une puce Affymetrix en détection : il s'agit d'une étude de faisabilité. La deuxième a retenu les choix méthodologiques trouvés par l'étude de faisabilité et a évalué la reproductibilité de l'ensemble du protocole, du recueil des urines jusqu'à la détection des signaux sur puces. C'est une étude de reproductibilité.

II.4.2. Mode d'obtention des échantillons urinaires

Les prélèvements urinaires sont réalisés dans les services de consultation d'urologie des HCL dans le cadre de la procédure de suivi des patients. L'examen par toucher rectal (DRE pour Digital Rectum Examination) fait partie de la prise en charge normale du patient. Cette procédure a été adaptée pour induire le relarguage des cellules tumorales de la prostate. Ainsi, lors du toucher rectal, l'urologue réalise trois palpations par lobe, de la base vers l'apex et de la ligne latérale vers la ligne médiane pour chaque lobe, ce qui provoque une pression suffisante pour créer en surface une dépression sur environ 1 cm. Le premier jet de la miction du patient est recueilli rapidement (25 mL maximum) et deux fois 2,5 mL d'urine sont prélevés dans des tubes *ad hoc* (Urine Specimen Kit, Gen-Probe) et servent au calcul du nombre de copies PSA et PCA3. Les urines restantes sont transférées dans un poudrier et envoyées à température ambiante par la transitique. Le service de biologie de l'hôpital réceptionne alors les prélèvements et réalise des mesures biochimiques sur bandelettes Combur 10 (pH, nitrite, glucose, bilirubine et leucocytes entre autres) avant d'appliquer le protocole de traitement de l'échantillon tel que discuté dans le paragraphe suivant.

II.4.3. Etude de faisabilité

II.4.3.1. Enjeux de l'étude et protocoles

Les contraintes pré-analytiques associées aux études transcriptomiques ont en effet été une cause d'adaptation des protocoles de traitement de l'échantillon. Ces optimisations ont porté à la fois sur le traitement des urines par le service de biologie et sur la recherche, au laboratoire, des procédés d'extraction d'acides nucléiques les plus adaptés à notre problématique. Cet aspect du travail constitue donc autant une étude de faisabilité au sens strict (c'est-à-dire cherchant à montrer si l'on on peut détecter une activité transcriptionnelle sur puces ADN à partir d'ARN extraits d'urines) qu'un travail technique de choix de protocoles de référence avant d'engager un travail sur de larges cohortes. En concertation avec le service de biologie, deux protocoles de traitement des urines ont été évalués en parallèle sur trois échantillons (Tableau II-5). Les différences entre ces protocoles se justifient principalement par la tentative de concilier plusieurs modes de préparations des échantillons destinés aux projets de recherche clinique menés plus largement à l'hôpital.

Protocole A : dans l'heure qui suit la réception du prélèvement, centrifugation des urines à 800 g pendant 8 minutes à 4°C, reprise du culot dans du PBS et nouvelle centrifugation à 200 g pendant 5 minutes. Puis reprise du culot dans du RNA Cell Protect (Qiagen) et stockage à –80°C.

Protocole B : dans l'heure qui suit la réception du prélèvement, centrifugation des urines à 1000 g pendant 15 minutes à température ambiante. Reprise du culot dans du PBS et nouvelle centrifugation à 200 g pendant 5 minutes. Puis reprise du culot dans du RNA Cell Protect (Qiagen) et stockage à –80°C.

Patient (urine)	Copies PCA3/mL	Copies PSA/mL	Score PCA3	Score de Gleason
667 A/B	332366	25673	77	G6 (3+3)
668 A/B	1738	69362	25	Pas de cancer
670 A/B	956	76353	12	G6 (3+3)

Tableau II-5 Echantillons utilisés pour l'étude de faisabilité. Les dosages de PCA3, du PSA ainsi que le score de Gleason lu sur biopsies sont fournis par le service d'urologie de l'hôpital. Le score PCA3 correspond au rapport du nombre de copies PCA3 sur le nombre de copies PSA x 1 000.

Au laboratoire, trois protocoles d'extraction des ARN ont été évalués sur ces échantillons : All Prep DNA/RNA Mini kit (Qiagen), All Prep DNA/RNA Micro kit (Qiagen) et PicoPure ARN (Arcturus). Ces trois protocoles utilisent des colonnes de purification mais leur gamme de sensibilité et la nature des acides nucléiques extraits varient. Notamment, les protocoles Qiagen permettent d'isoler ADN et ARN (ARNm et microARN) alors que PicoPure est destiné à la seule purification des ARNm. Suite à l'extraction des ARN, la procédure d'amplification, de fragmentation et de marquage conduisant à l'hybridation du matériel nucléique sur puce ADN a été suivie (Figure II-18, page suivante).

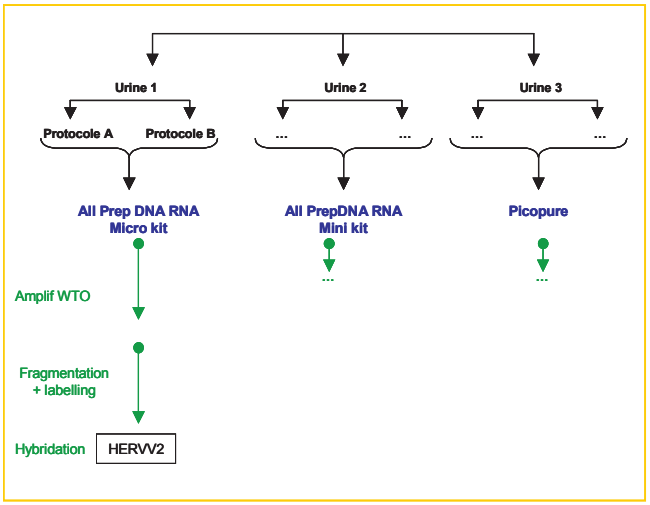


Figure II-18 Stratégie d'étude de faisabilité des étapes pré-analytiques et de détection. Des échantillons d'urines à l'hybridation sur puce Affymetrix HERV-V2, deux variables importantes sont évaluées : le protocole A ou B de traitement de l'échantillon et le choix du procédé d'extraction des ARN.

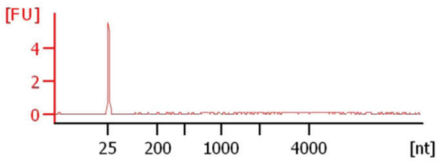
II.4.3.2. Comparaison de trois protocoles d'extraction

Les ARN extraits sont dosés par deux systèmes de mesure, le Nanodrop (Thermo) et le Qubit (Invitrogen) choisis pour leur limite de détection basse de l'ordre de 1 ng/µL ; les spectrophotomètres usuels en cuve ou les électrophorèses sur micro puce ne permettant pas ici d'obtenir de valeur de dosage fiable sur ces échantillons. Le principe de dosage diffère cependant entre les appareils : le Nanodrop mesure l'absorbance des acides nucléiques présents dans une micro goutte en tension alors que le dosage Qubit se base sur une fluorescence apportée par une molécule intercalante spécifique (ADN, ARN ou protéine). Parallèlement, les profils de tailles des ARN extraits sont mesurés par une électrophorèse en système micro puces (Bioanalyzer) dans le but d'évaluer les populations d'ARN en présence, leur quantité relative ainsi que le niveau général de dégradation des acides nucléiques (Figure II-19).

A

	Echantillons	Nanodrop (ng/µL)	Qubit (ng/mL)	Quantité totale d'ARN (ng)
DNA RNA micro	U667 A	2,6	too low	37,1
	U667 B	5,6	too low	79,3
DNA RNA Mini	U668 A	4,3	too low	174,0
	U668 B	2,5	too low	77,4
PicoPure	U670 A	16,1	too low	323,4
	U670 B	16,0	too low	320,2

B



C

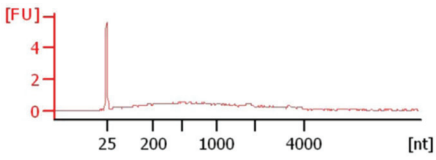


Figure II-19 Dosage et caractérisation des ARN extraits de prélèvements urinaires de trois patients. (A) Dosages Nanodrop et Qubit des ARN extraits par trois méthodes. Too low < 1ng/µL. (B) Profils de tailles Bioanalyzer des ARN extraits des échantillons 667 A/B et 668 A/B. (C) Profils de tailles Bioanalyzer des ARN extraits des échantillons 670 A/B. Le pic à 25 nt est un marqueur interne.

Les quantités d'ARN recouvrées à l'issue des extractions sont à la limite de détection des systèmes et l'indice d'intégrité des ARN (RIN, voir paragraphe II.3.1) ne peut pas être calculé. Toutefois, un épaulement est détectable sur la ligne de base du profil de tailles des échantillons extraits avec PicoPure, en cohérence avec des valeurs de dosage qui sortent du bruit de fond. Dans ce cas, l'extraction a donc conduit à l'isolement de quelques centaines de nanogrammes d'ARN caractérisés par un fort niveau de dégradation (absence complète de détection des ARN ribosomaux 18S et 28S). Notez que le plan de l'expérience ne permet pas *stricto sensu* de mettre en évidence le meilleur procédé d'extraction dans la mesure où il pourrait y avoir des variations inter individuelles importantes sur l'échantillon de départ. Néanmoins, ces résultats préliminaires ont conduit à retenir le procédé PicoPure qui, réutilisé par la suite sur des échantillons surnuméraires, a systématiquement surpassé le rendu d'autres procédés d'extraction. La reproductibilité de la méthode est quant à elle évaluée dans le paragraphe suivant. Par ailleurs, aucune incidence du protocole A ou B de traitement des urines n'est quantifiable à cette étape.

II.4.3.3. Amplification des ARN extraits

Suite à cela, les ARN des différents échantillons ont été amplifiés par le procédé WTO Nano, basé sur l'utilisation d'amorces polydT en mélange avec des amorces aléatoires, et conduisant à la synthèse d'ADNc simple brin positif (voir étape SPIA dans partie II.3.1). La quantité initiale d'ARN étant inaccessible par dosage, nous avons travaillé à volume constant en réalisant l'amplification à partir de 2 µL. Les produits d'amplifications ont été dosés au Nanodrop et caractérisés à l'aide du Bioanalyzer (Figure II-20).

A

	Echantillons	Nanodrop (ng/µL)	Quantité totale d'ADN (ng)
DNA RNA micro	667 A	47,3	995
	667 B	45,0	990
DNA RNA Mini	668 A	25,5	612
	668 B	21,8	523
Picopure	670 A	84,8	1018
	670 B	76,1	990

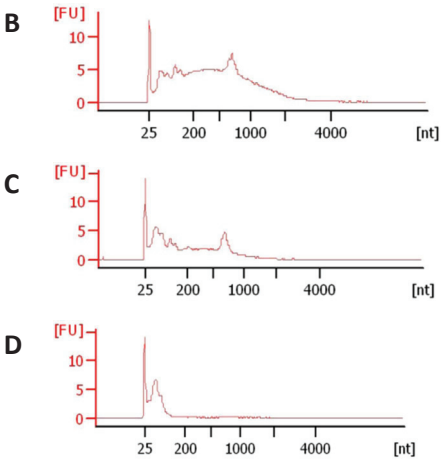


Figure II-20 Dosage et caractérisation des ADNc après amplification WTO Nano à partir des ARN extraits de prélèvements urinaires de trois patients. (A) Dosages Nanodrop des ADNc des échantillons. (B) Profils de tailles Bioanalyzer des ADNc des échantillons 670 A/B. (C) Profils de tailles Bioanalyzer des ADNc des échantillons 667 A/B. (D) Profils de tailles Bioanalyzer des ADNc des échantillons 668 A/B. Le pic à 25 nt est un marqueur interne.

Les rendements d'amplification varient de 0,5 µg à 1 µg en fonction du procédé d'extraction utilisé. DNA/RNA Micro et PicoPure occupent quantitativement tous les deux le haut de la gamme. D'un point de vue qualitatif cependant, la distribution des tailles d'ADNc est meilleure (et standard, cf. II.3.1) dans le cas d'une extraction PicoPure, pour laquelle la population s'étale de 25 à 2000 nucléotides, que pour les deux autres procédés d'extraction pour lesquels les tailles sont réduites entre 25 et 1000 nucléotides et entre 25 et 150 nucléotides. Ces résultats reflètent probablement le niveau de dégradation initial de l'ARN auquel s'ajoute un problème de quantités infraliminaires. Notons que le protocole standard Affymetrix nécessite d'obtenir 2 µg de produits d'amplification pour une hybridation sur puce. Cette expérience suggère qu'il sera nécessaire de faire deux amplifications WTO Nano en parallèle sur le même échantillon d'ARN extraits par PicoPure pour arriver à un total de 2 fois 1 µg. A ce sujet, les résultats de deux expériences sont proposés dans l'annexe III, qui évaluent l'impact de la quantité de matériel hybridé et l'effet d'un pool de deux amplifications WTO Nano sur l'intensité des signaux détectés.

II.4.3.4. Incidence de l'étape de fragmentation

L'hybridation sur puce ADN Affymetrix nécessite de travailler avec une population d'ADNc simple brin ayant une taille moyenne de 50 nucléotides. Les sondes de capture ayant une taille de 25 nucléotides, utiliser des cibles au-delà de 200 nucléotides peut défavoriser la dynamique d'hybridation, former des structures secondaires ou des ponts entre les sondes et, finalement, induire un biais de détection des signaux. Sur un échantillon d'ARN de bonne qualité et après une étape d'amplification, un traitement à la DNAse est donc systématiquement réalisé pour obtenir une distribution de taille d'ADNc autour de 50 nucléotides (cf. II.3.1). Ici, l'hétérogénéité des produits d'extraction et d'amplification a posé la question de l'utilisation du protocole standard de fragmentation. Autrement dit, la nature de l'échantillon post amplification permet-elle de recoller au protocole classique de réalisation de puce ADN, ou bien doit-on adapter (voire renoncer à) l'étape de fragmentation. Une fragmentation sans adaptation de protocole a donc été réalisée sur les différents produits d'amplification.

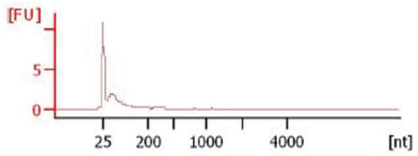


Figure II-21 Profils de tailles Bioanalyzer d'ADNc fragmentés. Les 6 échantillons présentent le même profil. Le pic à 25 nt est le marqueur interne.

Quelque soit l'échantillon d'ADNc, le traitement à la DNAse conduit à un profil-type de produit de fragmentation (Figure II-21), conforme aux recommandations de tailles. Là on l'on pouvait craindre que des échantillons déjà fortement fragmentés soient complètement détruits, l'action de la

DNAse a finalement épargné les petites tailles et aligné les différents profils sur la valeur attendue de 50 nucléotides.

II.4.3.5. Hybridation sur la puce HERV-V2

Nous avons donc procédé au marquage et à l'hybridation de six puces HERV-V2. En marge de son répertoire HERV, rappelons que la puce HERV-V2 possède des sondes de détection pour 319 gènes reflétant la biologie générique ou spécifique d'une cellule ou ayant précédemment été associés à des états cancéreux (cf. II.2.1.4). En particulier, l'expression de la kallicréine 3 (KLK3) est utile pour signer ici spécifiquement la présence de transcrits d'origine prostatique. Les puces ont été normalisées par la méthode RMA[12] et la distribution des signaux d'intensité est représentée (Figure II-22).

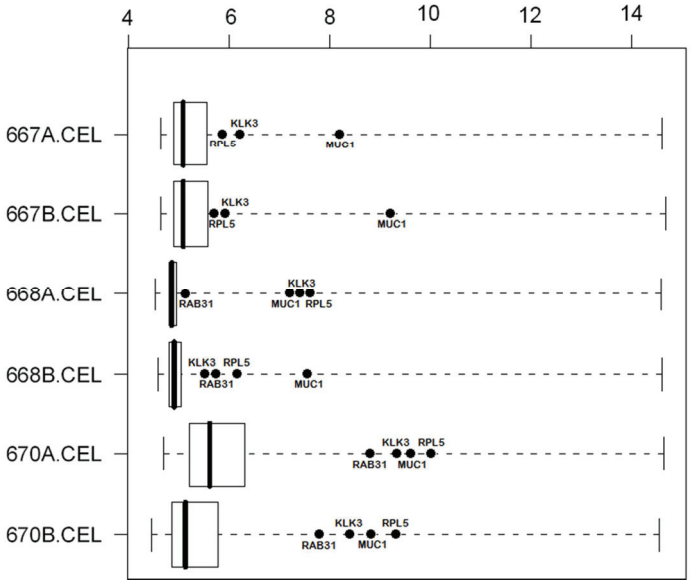


Figure II-22 Distribution des intensités de signaux des 6 puces de l'étape de faisabilité. Les valeurs d'intensité sont données en $\log_2$, la médiane est représentée par une barre verticale et les boites représentent 90 % des valeurs de la distribution. Les valeurs d'expression de KLK3, RPL5, RAB31 et MUC1, lorsqu'elles sortent des boites à moustache, sont indiquées par des cercles noirs.

Le bruit de fond d'une puce Affymetrix est de l'ordre de 2^6-2^7, comme nous avons eu l'occasion de le mentionner plus tôt (cf. II.3.4). Le niveau général d'expression des différentes puces est donc très faible et la majorité des signaux peuvent être considérés comme négatifs. Cependant quelques dizaines à quelques centaines de probesets sortent du bruit de fond dans chaque condition expérimentale, affichant exceptionnellement des valeurs élevées. En particulier, KLK3 est détecté sans ambiguïté dans les deux échantillons extraits avec PicoPure et sort légèrement du bruit de fond dans l'échantillon 668 A. La valeur d'expression de RPL5, transcrit de la protéine ribosomale 60S, peut être vu comme un indicateur de la quantité totale d'ARN cellulaire et donc comme un reflet de

[12] Robust Multi-array Average. Voir (Irizarry et al. 2003) et l'annexe II.

la qualité de l'échantillon utilisé. Le transcrit de la mucine 1 (MUC1) et celui d'un oncogène apparenté à RAS (RAB31) donnent quant à eux des niveaux d'expression qui peuvent être interprétés comme participant d'un état cellulaire pathologique.

Ensemble, ces résultats concourent à privilégier l'utilisation de PicoPure plutôt que des kits DNA/RNA Micro et Mini pour l'extraction d'ARN à partir de cellules urinaires et montrent qu'il est possible, dans ces conditions, de mener à bien une étude transcriptomique.

II.4.3.6. Conclusions de l'étude de faisabilité

La comparaison de deux protocoles de traitement des urines et de trois procédés d'extraction d'acides nucléiques a montré qu'une extraction PicoPure permet d'extraire des quantités d'ARN compatibles avec des étapes d'amplification et de détection en aval. Cependant, la qualité des ARN ne peut pas être validée par le calcul du RIN au Bioanalyzer. De plus, l'extraction des acides nucléiques par PicoPure exclut la possibilité de travailler sur l'ensemble des répertoires nucléiques ADN, ARNm et miRNA dans la mesure où seuls les ARNm sont purifiés. Les signaux d'hybridation sur la puce HERV-V2 obtenus avec les ARN extraits par PicoPure sont plus élevés que les signaux obtenus avec les procédés de purification sur colonnes Qiagen (DNA/RNA Micro et Mini). Une détection des transcrits de KLK3 fait notamment la preuve de concept que le culot des urines contient des cellules épithéliales de prostate. L'expression de gènes constitutifs (RPL5) ou liés au cancer (MUC1, RAB31) est observée et valide la faisabilité d'une approche transcriptomique sur ARN extraits de cellules urinaires, soulignant toutefois les risques liés à des valeurs d'expression globalement faibles. La comparaison de méthodes de traitement des urines ne permet pas de faire un choix objectif entre deux protocoles ; ainsi, pour des raisons extérieures, liées au partage d'échantillons biologiques avec d'autres projets de recherche, le protocole B est retenu pour la suite.

II.4.4. Etude de reproductibilité

II.4.4.1. Enjeux de l'étude et protocoles

L'originalité de la mise en œuvre d'une approche transcriptomique à partir d'échantillons complexes (ARN dégradés et en petites quantités), et en utilisant une puce ADN non commerciale, impose d'avoir entre les mains un protocole fiable, c'est-à-dire *a minima* reproductible. Après avoir identifié une combinaison de procédés qui semble adaptée aux contraintes techniques, nous avons conduit une expérience de répétition, à partir d'un pool d'urines, afin d'évaluer la reproductibilité de chaque étape du protocole. Ainsi, la variabilité des opérations pré-analytiques (extraction, amplification) et de détection des signaux (hybridation) permettra-t-elle d'augurer des chances de succès dans la recherche et l'identification d'une expression différentielle liée à un statut clinique. La

valeur d'un résultat biologique ne doit en effet pas s'inscrire dans la marge d'erreur de la variation intrinsèque de la méthode qui conduit à ce résultat. Une question sous-jacente à cette évaluation générale de la reproductibilité était notamment de savoir si la puce à façon HERV-V2, créée au laboratoire, pouvait donner des coefficients de variation (CV) comparables à ceux de la puce commerciale HG-U133-PLUS2, utilisées dans ce cadre expérimental. Les répertoires des rétrovirus endogènes humains et des gènes dits conventionnels ont donc été étudiés en parallèle sur un pool de quatre échantillons (Tableau II-6).

Patient (urine)	Copies PCA3/mL	Copies PSA/mL	Score PCA3	Score de Gleason
U865	322881	1219528	265	G7 (3+4)
U857	4074	281695	14	G7 (3+4)
U854	26662	822279	32	G7 (3+4)
U844	221299	1815238	122	G7 (3+4)

Tableau II-6 Echantillons utilisés pour l'étude de reproductibilité. Les dosages de PCA3, du PSA ainsi que le score de Gleason lu sur pièce de prostatectomie radicale sont fournis par le service d'urologie de l'hôpital de Lyon-Sud. Le score PCA3 correspond au rapport du nombre de copies PCA3 sur le nombre de copies PSA x 1000.

Le pool d'échantillons a été réalisé à partir de 4 culots urinaires, mélangés puis séparés en 4 réplicats. De l'extraction des ARN avec PicoPure à l'hybridation sur puce, la démarche de répétition se réalise en parallèle (Figure II-23).

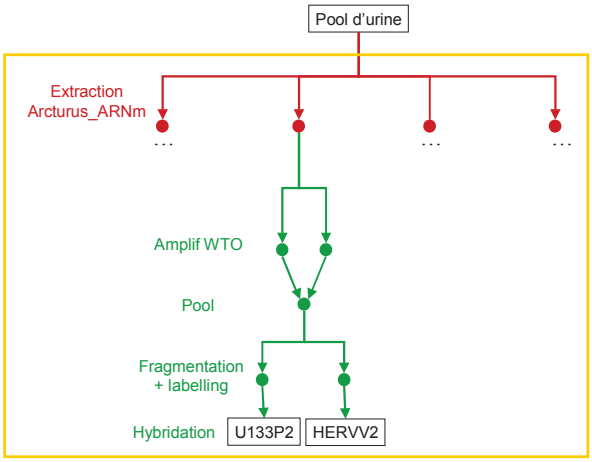


Figure II-23 Stratégie d'étude de la reproductibilité des étapes pré-analytiques et de détection. D'un pool d'urines à la détection de deux répertoires nucléiques, quatre expériences identiques sont menées en parallèles dans le but de calculer les coefficients de reproductibilité (CV = (écart-type / moyenne) x 100).

II.4.4.2. Etape d'extraction

Les dosages et la caractérisation des ARN extraits ont été réalisés, comme précédemment, à l'aide du Nanodrop et du Bioanalyzer (Figure II-24).

A

Echantillons	Nanodrop (ng/µl)	Quantité totale d'ARN (ng)
Ext 1	22,2	445,0
Ext 2	21,7	436,6
Ext 3	18,9	379,0
Ext 4	21,8	437,8

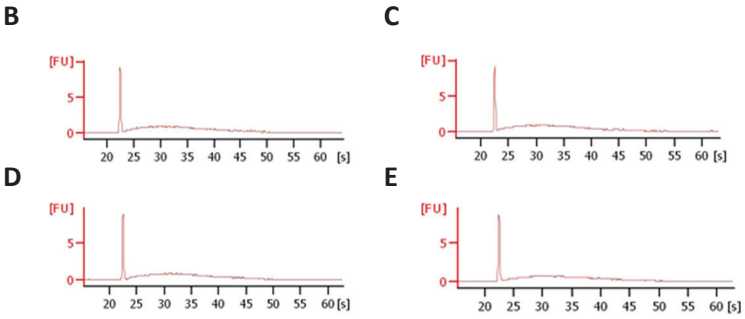


Figure II-24 Dosage et caractérisation des ARN extraits des réplicats techniques. (A) Dosages Nanodrop et Qubit des ARN. (B-E) Profils de tailles Bioanalyzer des ARN. Le pic à 25 nt est le marqueur interne.

Ces dosages, réalisés sur des échantillons indépendants de l'étude de faisabilité, confirment que l'extraction PicoPure permet d'obtenir des quantités d'ARN de l'ordre de 0,3 µg à 0,4 µg. Le coefficient de variation de l'extraction est égal à 7,2 % et fait de cette étape une source de variation non négligeable, tout est restant acceptable pris isolément. Les quatre profils de tailles sont très semblables entre eux et en cohérence avec le profil PicoPure obtenu dans l'étude de faisabilité, confirmant l'épaulement caractéristique attendu et l'absence de détection d'ARN 18S et 28S permettant de calculer un RIN.

II.4.4.3. Etape d'amplification des ARN

A partir d'un volume constant (2µl), deux amplifications WTO Nano ont été réalisées sur les ARN purifiés à partir des 4 extractions, conduisant à 8 produits d'amplification notés de 1-1' à 4-4'. Les ADNc post amplification sont dosés et caractérisés sur gel comme précédemment (Figure II-25).

A

Echantillons	Dosage Nanodrop (ng/µl)	Quantité totale (ng)
Ext 1	100	2820
Ext 1'	99	2772
Ext 2	103	2885
Ext 2'	110	3097
Ext 3	96	2693
Ext 3'	105	2962
Ext 4	98	2745
Ext 4'	128	3587

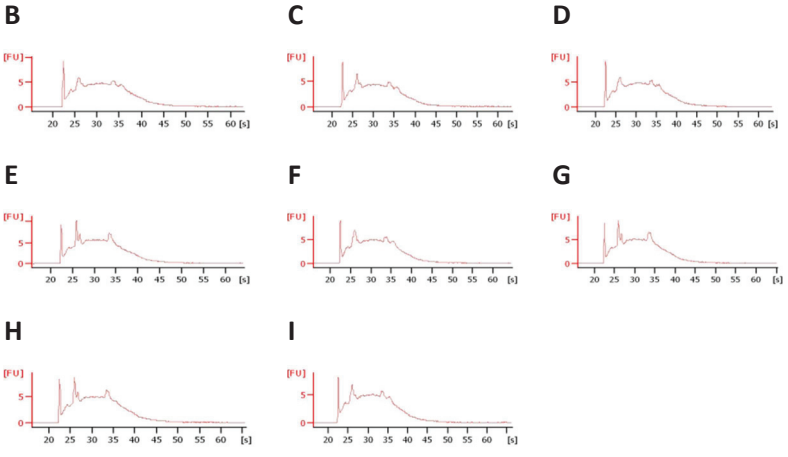


Figure II-25 Dosages et caractérisations des ADNc après amplification WTO Nano à partir des ARN extraits des réplicats techniques. (A) Dosages Nanodrop des ADNc des échantillons. (B-I) Profils de tailles Bioanalyzer des ADNc des échantillons. Le pic à 22 s est le marqueur interne correspondant à 25 nt.

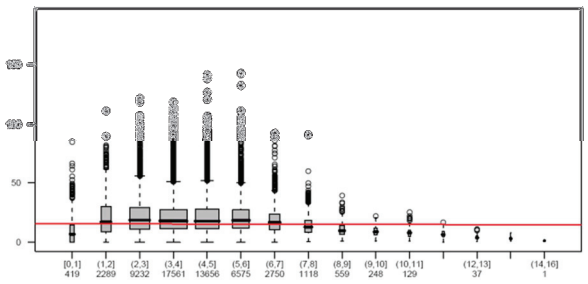
Les rendements d'amplification sont groupés entre 2,5 µg et 3,5 µg, ce qui supérieur d'un facteur 2 à 3,5 par rapport à ce qui avait été obtenu lors de l'étude de faisabilité générale. Cette

différence de rendement peut être imputable à deux causes : les échantillons biologiques de départ sont différents, et les lots du kit d'amplification ont changé (ici les lots sont plus récents que sur la première étude). Le coefficient de variation cumulé (extraction et amplification) à cette étape est de 9,8 % et souligne l'incidence potentielle des lots de réactifs utilisés dans des expériences qui s'étalent dans le temps. Les profils de tailles sont très homogènes et correspondent à des profils classiques attendus en appliquant le protocole d'amplification WTO Nano en conditions standard (comme déjà observé sur l'échantillon 670). Comme indiqué dans le schéma récapitulatif de cette étude en tête de paragraphe, le couple de produits d'amplification WTO Nano a été homogénéisé par réplicat et redistribué en deux tubes égaux contenant chacun 2 µg d'ADNc.

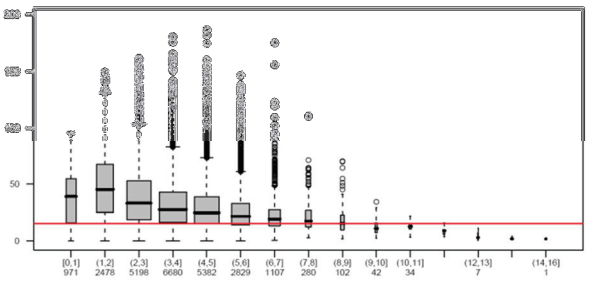
II.4.4.4. Etape d'hybridation

De là, les étapes de fragmentation, marquage et hybridation conduisent à la réalisation d'une puce HG-133-PLUS2 et d'une puce HERV-V2 par réplicat. L'étape de fragmentation est reproductible (non illustrée) et correspond aux valeurs de tailles attendues (50 nucléotides) et la reproductibilité de l'étape de marquage n'a pas été évaluée. La variation d'une étape étant liée à celle des étapes qui la précèdent, la variation finale post hybridation est une somme de contraintes additives et par là même prend en compte des sources de variations antérieures qui auraient pu être négligées. Les CV globaux sont ainsi calculés par probesets, et leurs valeurs distributives représentées en fonction de l'intensité des signaux (Figure II-26).

A



B



C

	[0,1]	(1,2]	(2,3]	(3,4]	(4,5]	(5,6]	(6,7]	(7,8]	(8,9]	(9,10]	(10,11]	(11,12]	(12,13]	(13,14]	(14,16]
CV median	7	17	19	18	18	18	17	13	10	9	8	6	4	3	1
# probesets	419	228	923	1756	1365	657	275	111	559	248	129	81	37	20	1
CV median	39	45	33	27	25	22	19	18	15	11	13	9	3	2	2
# probesets	971	247	519	6680	5382	282	110	280	102	42	34	20	7	11	1

Figure II-26 Coefficients de variation globaux (CV) et nombre de probesets associés. (A) Distribution des valeurs de CV par intervalles de tailles pour la puce HG-U133-PLUS2. (B) Distribution des valeurs de CV par intervalles de tailles pour la puce HERV-V2. La barre horizontale rouge indique la valeur de CV = 15 %. (C) Valeurs des CV et nombre de probesets par intervalles de tailles pour les deux types de puces.

Les coefficients de variation passent sous la barre des 15 % à partir de valeurs d'intensités de 128 (2^7) pour la puce HG-U133-PLUS2 et de 256 (2^8) pour la puce HERV-V2, permettant une détection reproductible sur l'ensemble du protocole d'environ 200 à 300 probesets de la puce HERV-V2 et de 2000 à 2500 probesets de la puce HG-U133-PLUS2. Dans ce contexte, la reproductibilité, donc la fiabilité, des signaux de détection des puces est déplacée d'une à deux puissances de 2 au-dessus du bruit de fond d'une puce réalisée en suivant les protocoles de référence. La reproductibilité un peu plus faible associée à la puce HERV-V2 peut venir du plus faible nombre de probesets ayant des signaux d'expression au-dessus du bruit de fond.

II.4.4.5. Conclusions de l'étude de reproductibilité

En utilisant PicoPure et le mode de traitement B des urines, l'ensemble du protocole de transcriptomique, de l'extraction à la détection, confirme les tendances qualitatives apportées par l'étude de faisabilité et montre une assez bonne reproductibilité générale de l'enchainement des étapes. L'extraction confirme que l'on peut atteindre 0,3 µg d'ARN ainsi qu'un profil-type de tailles présentant un léger épaulement sur l'intervalle de 200 à 3 000 nucléotides. Le CV de cette étape est de 7,2 % et il s'agit probablement de l'étape qui crée le plus d'écarts entre deux réplicats. L'amplification WTO Nano assure des rendements qui sont compatibles avec le protocole Affymetrix standard (quantité totale requise = 2 µg) et les profils de tailles, identiques d'un réplicat à l'autre, indiquent une bonne reproductibilité. Le CV jusqu'à cette étape (extraction et amplification) est de 9,8 % (Tableau II-7).

Etapes	CV cumulés HG-U133-PLUS2	CV cumulés HERV-V2
Extraction	≈7 %	
Amplification	<10 %	
Expression sur puce Affymetrix	< 13 % pour valeurs > 2^7	< 18 % pour valeurs > 2^7

Tableau II-7 Coefficients de variations cumulés des étapes pré-analytiques et de détection de l'étude de reproductibilité.

Enfin, la lecture des puces donne un ordre de grandeur du nombre de probesets présentant des signaux d'expression exploitables : il y en a environ 200 à 300 pour le répertoire HERV et presque 10 fois plus (2 000 à 2 500) pour le répertoire des gènes conventionnels. Les CV passent sous la barre des 15 % pour toutes les valeurs d'expression au-delà d'un seuil d'intensités de 2^8 (256), soit 2 fois le niveau du bruit de fond théorique.

II.4.5. Choix de systèmes référents de RT-PCR quantitative

La faisabilité et la reproductibilité d'une approche transcriptomique sur puce ADN à partir d'ARN extraits d'urine établies, nous avons cherché à mettre en évidence, de manière quantitative, la présence de quelques transcrits d'intérêt par une méthode de référence. Cette approche a été motivée par plusieurs limites inhérentes aux expériences sur puces à ADN qui venaient d'être faites : (i) l'amplification WTO Nano à partir des échantillons d'ARN dégradés utilisés dans la phase de faisabilité, puis l'hybridation d'un mélange de produits d'amplification suivie d'une normalisation RMA sur un lot de puces ayant un faible niveau général d'expression réduisent considérablement la pertinence quantitative des signaux observés, (ii) le répertoire des gènes conventionnels de la puce HERV-V2 est limité et biaisé, il contient notamment peu de gènes de ménage et, exception faite de KLK3, aucun gène à tropisme prostatique, (iii) les puces HG-U133-PLUS2 de l'étude de reproductibilité, si elles sont par définition réalisées dans des conditions homogènes et permettent d'étudier l'ensemble des gènes conventionnels, ne peuvent à elles seules constituer un référent à la méthodologie : il faut impérativement un ancrage des résultats par une technique de détection qui soit indépendante des puces à ADN. Ceci doit permettre, en premier lieu, de faire définitivement la preuve de concept de la présence de transcrits spécifiques de l'organe ou liés à un état pathologique. Mais au-delà de cet enjeu de validation, chercher à étudier quels transcrits ubiquitaires sont atteignables dans les urines peut asseoir une base de stratégie de normalisation, étant compris qu'une étude clinique impliquant une large cohorte de patients mettra en jeu une variabilité inter individuelle très forte, dès l'échantillon de départ. Pour ce faire, 21 systèmes de PCR Taqman ont été utilisés sur de l'ADNc généré à partir des ARN de l'étude de reproductibilité (Tableau II-8).

Gènes fortement exprimés (1000-10000 molécules par cellule)		
18S	Eucaryotic 18S rRNA	Hs03003631_g1
β actin	Human ACTB (beta actin)	4333762F
GAPDH	Human GAPD (GAPDH)	4333764F
GUS β	Beta glucuronidase	4333767F
TPB	Human TATA-box binding protein	4333769F
RPLP0	Ribosomal Protein, large, P0	Hs99999902_m1
PPIA	Peptidylprolyl isomerase A (Cyclophilin A)	Hs99999904_m1
Gènes faiblement exprimés (1-100 molécules par cellule)		
RARA	Retinoic acid receptor, alpha	Hs00940446_m1
RARB	Retinoic acid receptor, beta	Hs00233407_m1
RXRB	Retinoic X receptor, beta	Hs00232774_m1
Gènes associés à la prostate		
KLK3 (PSA)	Kallikrein-related peptidase 3 (prostate specific antigen)	Hs02576345_m1
ACPP	Acid phosphatase prostate	Hs00173475_m1
PCA3	Prostate cancer antigen 3	Hs01371938_m1
		Hs01371939_g1

Gènes associés à des états cancéreux		
MUC1	Mucin 1, cell surface associated	Hs00159357_m1
TRIM29	Tripartite Motif-containing 29	Hs00232590_m1
EZR	Ezrin	Hs00931653_m1
GOLM1	Golgi membrane protein 1	Hs00213061_m1
GSTP1	Glutathione S-transferase pi 1	Hs00943351_g1
ANXA3	Annexin A3	Hs00192983_m1
		Hs00971411_m1

Tableau II-8 Gènes candidats à la détection par RT-PCR Taqman. De gauche à droite : le symbole du gène, son nom complet et les références du système Taqman correspondant.

Les gènes de ménage testés (ARN ribosomaux ou gènes du métabolisme pour l'essentiel) sont choisis dans une large gamme de niveaux d'expression dans le but de faire apparaitre des limites de détection hautes et basses. Par ailleurs trois gènes spécifiques de la prostate sont évalués : le PSA (KLK3), l'acide phosphatase (une tyrosine-phosphatase androgéno-dépendante synthétisée dans les cellules épithéliales des glandes prostatiques) et le transcrit non-codant PCA3. Différents gènes associés à des états cancéreux dans des contextes variés sont également inclus. La moyenne de détection de quatre réplicats est donnée pour chaque gène (Tableau II-9).

Gène	**18S**	**RPLP0**	**PPIA**	**EZR**	**MUC1**	**GAPDH**	**ACTB**
Mean Ct	16,2	27,4	28,4	28,5	28,6	29,8	31,0
Gene	**KLK3**	**ACPP**	**PCA3_39**	**GSTP1**	**GOLM1**	**ANXA3_11**	**GUS**
Mean Ct	31,1	31,2	31,6	31,8	31,9	32,8	32,5
Gene	**ANXA3_1**	**RARA**	**RXRB**	**RARB**	**TRIM29**	**TPB**	**PCA3_38**
Mean Ct	33,4	33,6	34,5	34,7	36,3	36,6	38,1

Tableau II-9 Moyennes de détection des 21 systèmes de RT-PCR Taqman testés sur quatre réplicats techniques.

Compte-tenu des dilutions d'ADNc utilisées pour cette expérience de PCR, nous considérons qu'il est aventureux d'analyser des résultats au-delà de 32 cycles d'amplification. Sous ce seuil, sont détectés les gènes de ménage 18S, RPLP0, PPIA, GAPDH et ACTB, ainsi que deux gènes associés à des états cancéreux, EZR et MUC1, et les trois gènes prostatiques KLK3, ACPP et PCA3. Notez que l'absence de détection du second système PCA3 est vraisemblablement due à des variants d'épissage. En cohérence avec leur faible niveau d'expression, les gènes de ménage RARA, RARB et RXRB ont des signaux non analysables (au-delà de 32 cycles). A l'inverse, deux gènes de ménage supposés avoir une expression assez forte (GUSβ et TPB) ne passent pas la barre du bruit de fond, ce qui peut refléter le niveau de dégradation de l'échantillon de départ et les problèmes de sensibilité que celui induit. Dans une perspective d'utilisation de la RT-PCR pour normaliser des échantillons cliniques, nous avons donc exclu les systèmes dont la détection est trop faible, ainsi les gènes pour lesquels il est connu que l'expression varie en fonction de l'état de différenciation. Par ailleurs, le

système 18S, s'il affiche les signaux de détection les plus élevés, a été écarté car il peut croiser avec l'ADN. Ainsi sont retenus, et seront systématiquement utilisés sur les échantillons d'ARN à venir, les gènes de ménages RPLP0, PPIA, GAPDH et ACTB. Nous y ajoutons KLK3 qui reste la meilleure signature ARN de l'origine prostatique, bien que la variation de son expression cellulaire en contexte physiopathologique puisse être sujette à discussion.

II.4.6. Etude clinique pilote sur 45 patients

II.4.6.1. Enjeux et stratégie de l'étude

Trois classes de 15 patients ont été définies en concertation avec les praticiens hospitaliers dans le but d'identifier des marqueurs diagnostiques précoces du cancer de la prostate ou utiles en pronostic afin de différencier un cancer indolent (surveillance active) d'un cancer évolutif (traitement). Le recrutement a été positionné sur la zone grise du PSA, étendue entre 2 ng/µL et 13 ng/µL. Les biopsies ne donnant qu'une vision partielle de l'état de l'organe et sous-estimant régulièrement les foyers cancéreux, seuls les patients ayant eu au moins 2 séries de biopsies négatives sont utilisés en contrôle. Parallèlement, la prostatectomie radicale a été le critère absolu de lecture de l'état des tissus des patients atteints de cancer. Dans ce dernier cas, une valeur pronostique, fonction du score de Gleason, du grade TNM (voir paragraphe I.4.2.3 pour ces notions) et de la présence de marges chirurgicales a séparé deux populations de patients, par une dichotomie essentiellement associée à une appréciation du risque de récidive post-opératoire. Ainsi, deux niveaux d'analyse des données pourront coexister : la question du diagnostic par la comparaison des 15 patients contrôles aux 30 patients ayant un cancer, ou, à un niveau plus fin, l'opposition entre les deux groupes pronostiques.

II.4.6.2. Définition des classes de patients

II.4.6.2.1. Groupe des ponctions-biopsies prostatiques négatives : PBP NEG

Ce groupe contrôle regroupe des patients et qui ont eu au moins deux biopsies consécutives négatives durant leur suivi médical, faisant ainsi baisser le risque de faux négatifs de 20 % (cas d'une biopsie unique) à 5 % (Tableau II-10). La ponction-biopsie prostatique (PBP) s'effectue par voie transrectale sous contrôle d'un échographe endorectal, permettant des prélèvements étagés dans les deux lobes et pouvant être dirigés vers un nodule.

ID	Urine	Age	n PBP	Vol. prostate (cm^3)	Gleason PBP	PSA total (ng/µL)	Copies PCA3/mL	Copies PSA/ml	Score PCA3
191	109	62	2	57	Pas de Cancer	6,6	707	136493	5
703	209	66	2	50	Pas de Cancer	8,0	19625	528408	37
378	243	55	2	98	Pas de Cancer	12,1	44351	2855258	16
185	401	69	2	62	Pas de Cancer	3,7	31548	1665269	19
788	445	66	2	30	Pas de Cancer	3,9	8976	818144	11
891	557	54	3	77	Pas de Cancer	13,0	19505	638257	31
735	643	66	3	42	Pas de Cancer	3,7	21621	1131280	19
968	654	64	2	67	Pas de Cancer	8,9	531106	9923428	54
1087	782	69	2	28	Pas de Cancer	4,0	146402	5556978	26
220	894	60	3	34	Pas de Cancer	4,7	2461	422933	6
700	897	53	2	40	Pas de Cancer	4,5	3341	361948	9
461	924	62	3	48	Pas de Cancer	2,3	19858	302379	66
762	927	64	2	60	Pas de Cancer	11,1	3353	56339	60
90	943	54	3	50	Pas de Cancer	11,0	3135	293631	11
438	957	59	2	68	Pas de Cancer	4,8	56132	2894883	19

Tableau II-10 Les 15 patients du groupe PBP NEG. ID et Urine : double référencement des patients. Age : âge du patient au moment du prélèvement des urines. n PBP : nombre total de biospies réalisées sur le patient au cours de son suivi médical, en date du prélèvement. Le score de Gleason est évalué sur biopsies (PBP). Le volume de la prostate est évalué par échographie. Les dosages de PCA3 et du PSA sont fournis par le service d'urologie de l'hôpital de Lyon-Sud. Le score PCA3 correspond au rapport du nombre de copies PCA3 sur le nombre de copies PSA x 1 000.

II.4.6.2.2. Groupe de bon pronostic : GP

Ce groupe inclut des patients dont un diagnostic de cancer a été établi sur biopsies et ayant par la suite subi une prostatectomie radicale (Tableau II-11). Le score de Gleason sur PR est inférieur ou égal à 7 (3+4) et le grade de la tumeur est inférieur à pT3 avec absence de marge chirurgicale. Les cancers répondant à ces critères sont considérés plutôt comme à bon pronostic d'évolution.

ID	Urine	Age	Vol. prostate (cm^3)	PSA total (ng/µL)	Copies PCA3/mL	Copies PSA/mL	Score PCA3	Gleason PBP	Gleason PR	pTNM	Marge
141	127	60	36	2,8	35007	455496	8	6 (3+3)	7 (3+4)	2aNxM0	0
681	128	47	30	3,8	55645	188295	30	6 (3+3)	6 (3+3)	2cNxM0	0
241	330	56	30	2,9	40696	175680	23	6 (3+3)	6 (3+3)	2aNxM0	0
389	336	52	29	4,5	1461	63680	23	6 (3+3)	7 (3+4)	2cNxM0	0
124	339	51	20	4,0	1174	16779	70	6 (3+3)	6 (3+3)	2cNxM0	0
74	394	55	48	7,5	9977	81957	122	6 (3+3)	6 (3+3)	2cNxM0	0
375	421	53	30	2,1	21832	121495	180	6 (3+3)	6 (3+3)	2cNxM0	0
327	433	63	ND	7,8	32670	698778	47	6 (3+3)	6 (3+3)	2cNxM0	0
786	475	63	22	2,0	3424	192881	18	6 (3+3)	7 (3+4)	2cN0M0	0
795	600	53	26	3,4	28783	616164	47	6 (3+3)	6 (3+3)	2cNxM0	0
900	612	68	56	5,5	43811	257786	170	6 (3+3)	7 (3+4)	2cN0M0	0
102	717	58	14	3,6	4090	290553	14	6 (3+3)	7 (3+4)	2aN0M0	0
108	783	70	20	3,2	73313	287433	26	7 (3+4)	7 (3+4)	2cN0M0	0
998	785	50	17	5,0	44015	792353	56	6 (3+3)	7 (3+4)	2cN0M0	0
113	834	58	55	9,5	53143	176969	30	6 (3+3)	6 (3+3)	2cN0M0	0

Tableau II-11 (Légende de la page précédente) Les 15 patients du groupe GP. ID et Urine : double référencement des patients. Age : âge du patient au moment du prélèvement des urines. Le volume de la prostate est évalué par échographie. Les dosages de PCA3 et du PSA sont fournis par le service d'urologie de l'hôpital de Lyon-Sud. Le score PCA3 correspond au rapport du nombre de copies PCA3 sur le nombre de copies PSA x 1 000. Les scores de Gleason sont évalués à la fois sur les biopsies (PBP) puis plus tard sur les pièces de prostatectomies radicales (PR). pTNM : Classification TNM internationale après examen pathologique. Marge : 0 indique une marge chirurgicale négative, 1 indique une marge chirurgicale positive.

II.4.6.2.3. Groupe de mauvais pronostic : PP

Plusieurs critères d'inclusion ont été discutés et finalement retenus pour constituer ce groupe de patients dont le pronostic d'évolution est considéré comme défavorable (Tableau II-12). La tumeur doit être supérieure ou égale à pT3a, **ou** le score de Gleason lu sur PR supérieur ou égal à 7 (4+3), **ou** le score de Gleason lu sur PR égal à 7 (3+4) avec une tumeur ≥ pT2 et une marge positive.

ID	Urine	Age	Vol. prostate (cm^3)	PSA total (ng/µL)	Copies PCA3/mL	Copies PSA/mL	Score PCA3	Gleason PBP	Gleason PR	pTNM	Marge
605	79	65	17	7,8	70981	823068	86	7 (3+4)	7 (4+3)	3aN0M0	1
464	101	68	43	4,1	90157	2129477	42	6 (3+3)	7 (4+3)	2cN0M0	0
39	290	62	22	6,2	903219	4283713	211	7 (3+4)	7 (4+3)	3bN0M0	0
264	314	52	25	5,4	6786	275740	25	8 (4+4)	8 (4+4)	2aN0M0	0
544	434	61	35	7,5	402	9018	45	7 (3+4)	7 (4+3)	2cNxM0	0
784	485	61	22	5,6	84558	1157540	73	7 (4+3)	7 (4+3)	2cN0M0	1
915	583	70	47	3,8	16870	434476	39	8 (4+4)	8 (4+4)	2cN0M0	1
969	655	69	18	6,3	80678	1642637	49	7 (4+3)	7 (4+3)	3aN0M0	0
975	662	64	30	3,5	384962	2674899	144	6 (3+3)	7 (3+4)	3aNxM0	0
100	694	58	32	3,8	38121	1598809	24	8 (4+4)	8 (4+4)	3bN1M0	0
101	702	68	66	7,9	3711	361792	10	6 (3+3)	7 (3+4)	3aN0M0	1
102	716	60	40	4,6	1356	157681	9	6 (3+3)	7 (3+4)	2cN0M0	1
105	746	70	66	3,8	1155995	3617458	320	7 (3+4)	7 (3+4)	2cNxM0	1
107	763	65	27	5,1	107941	816143	132	9 (4+5)	9 (5+4)	3bN1M0	1
111	812	58	30	5,6	86390	894230	97	8 (4+4)	8 (3+5)	3aN0M0	0

Tableau II-12 Les 15 patients du groupe PP. ID et Urine : double référencement des patients. Age : âge du patient au moment du prélèvement des urines. Le volume de la prostate est évalué par échographie. Les dosages de PCA3 et du PSA sont fournis par le service d'urologie de l'hôpital de Lyon-Sud. Le score PCA3 correspond au rapport du nombre de copies PCA3 sur le nombre de copies PSA x 1 000. Les scores de Gleason sont évalués à la fois sur les biopsies (PBP) puis plus tard sur les pièces de prostatectomies radicales (PR). pTNM : Classification TNM internationale après examen pathologique. Marge : 0 indique une marge chirurgicale négative, 1 indique une marge chirurgicale positive.

II.4.6.3. Procédure de réalisation des puces HG-U133-PLUS2 et HERV-V2 sur les échantillons de l'étude clinique

De l'extraction des ARN à partir des culots urinaires jusqu'à l'hybridation sur les puces Affymetrix, un plan d'expérience pensé dans le but d'éviter des effets confondants techniques (dates, lots de produits, expérimentateur etc.) ou biologiques (état cancéreux, valeur du score de Gleason etc.) a été scrupuleusement suivi. L'extraction et l'amplification utilisent les procédés décrits plus haut (cf. II.3.1). Cependant, en fonction de l'échantillon, les amplifications WTO Nano ont été

réalisées en nombre variable (de 2 à 5), en parallèle, dans le but d'atteindre systématiquement la quantité de 4 µg (2 x 2 µg) d'ADNc nécessaire pour réaliser une hybridation sur les deux puces Affymetrix HG-U133-PLUS2 et HERV-V2 (cf. l'annexe III pour une estimation de l'impact d'un mélange de produits d'amplification). Afin d'étudier exactement la même cible sur ces deux répertoires nucléiques, les produits des amplifications parallèles d'un échantillon ont été mélangés et homogénéisés avant l'étape d'hybridation.

II.4.6.3.1. Extractions des ARN et caractérisations des 45 échantillons

Comme cela a été montré lors des étapes de faisabilité, les dosages des ARN, à la limite de détection des systèmes, sont loin d'être pertinents. En revanche la détection, sur les profils de tailles des produits d'extraction, d'un épaulement de la ligne de base, est un indicateur associé au succès des étapes pré-analytiques.

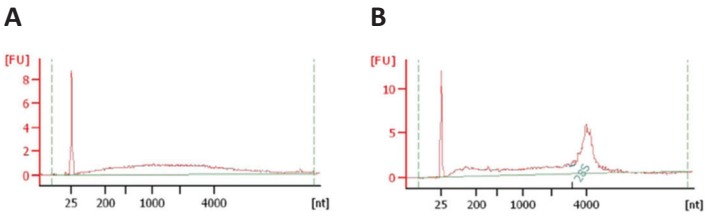


Figure II-27 Profils de tailles représentatifs de l'ensemble des ARN extraits dans le cadre de l'étude clinique. (A) Patient U330 : épaulement entre 25 nt et 4000 nt. (B) Patient U401 : un pic est, en plus, détectable dans la région des ARN ribosomaux 28S. Le pic à 25 nt est le marqueur interne.

Au terme des 45 extractions, les profils de tailles obtenus ont pu être regroupés en deux catégories (Figure II-27). A celle, attendue, faisant apparaître un épaulement entre 25 et 4000 nucléotides s'est ajoutée, pour environ un tiers des échantillons, la présence de populations d'ARN dans la zone de tailles des ARN ribosomaux. Ainsi, tout en présentant un certain niveau d'homogénéité, cette série d'extractions met en lumière un niveau de variabilité de la qualité des ARN, intrinsèque aux échantillons de départ.

II.4.6.3.2. Amplifications WTO Nano et caractérisations des ADNc

Pour chaque échantillon d'ARN, des amplifications WTO Nano sont réalisées en parallèle jusqu'à obtenir la quantité cumulée d'ADNc nécessaire à la réalisation de deux puces ADN, en l'occurrence : deux fois 2 µg. Les rendements d'amplification pouvant varier du simple au quadruple d'un échantillon à l'autre, certains cas ont nécessité jusqu'à cinq amplifications parallèles. La reproductibilité de l'amplification sur un même échantillon est telle que les profils de tailles qui en découlent sont très similaires, comme cela a été vu lors de l'étude de reproductibilité (cf. II.4.4.3).

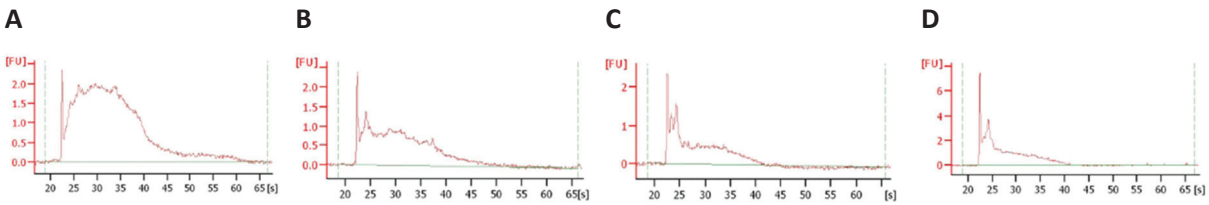


Figure II-28 Profils de tailles représentatifs de l'ensemble des produits d'amplification de l'étude clinique. (A) Patient U109 : profil-type attendu. (B) Patient U897 : profil intermédiaire, les rendements d'amplification sont plus faibles mais la distribution reste assez préservée. (C) Patient U717 : rendements d'amplification faibles et perte des tailles au-delà de 1 000 nt (40 s). (D) Patient U475 : état de fragmentation important. Le pic à 22 s est le marqueur interne correspondant à 25 nt.

Les profils de tailles des ADNc obtenus après amplification peuvent schématiquement être séparés en quatre catégories (Figure II-28), allant d'un profil-type de bonne qualité (cas A : 25 % des échantillons) à une situation de faibles rendements associés à un niveau de dégradation important (cas D : 33% des effectifs). Les échantillons entre ces deux extrêmes se répartissent de manière équilibrée. Tous les échantillons pour lesquels une population d'ARN ribosomaux était détectable au terme de l'extraction donnent des profils d'amplification de type A ou B, bien que des produits d'échantillons d'ARN qui ne présentaient qu'un léger épaulement se retrouvent également dans ces deux groupes d'amplicons.

II.4.6.4. Contrôle de l'amplification par RT-PCR et définition d'un score d'amplificabilité

Comme nous l'avons vu, travailler à volume constant (« à l'aveugle ») en entrée de l'étape d'amplification a pour conséquence qu'il reste très aléatoire de prédire la qualité des produits d'amplification qui seront obtenus. Au-delà de l'impact sur la planification des expériences et le coût de la mise en œuvre associée (amplifier 45 échantillons 5 fois ou lieu de 2 fois est loin d'être anodin), cette incertitude peut avoir, *de facto*, des répercussions fondamentales sur le recrutement des échantillons et la gestion qui en est faite dans le déploiement d'études cliniques. Plus pragmatiquement, l'intérêt de pouvoir, sinon anticiper un résultat, du moins gagner un certain niveau de contrôle sur la qualité des échantillons est un atout qui se révèlera utile dans une problématique de transposition d'un résultat de recherche en produit. Ainsi avons-nous entrepris, comme mentionné plus haut (cf. II.4.5), de réaliser systématiquement une détection par RT-PCR des quatre gènes de ménage RPLP0, PPIA, GAPDH et ACTB sur les ARN extraits. L'hypothèse est que le niveau de détection associé à ces transcrits ubiquitaires sera d'une manière ou d'une autre un indicateur qualitatif de l'échantillon. La mise en œuvre de cette stratégie a d'abord consisté à vérifier qu'aucune différence significative n'était observée dans l'expression de ces quatre gènes entre les classes cliniques (Test de Mann-Whitney, pour chaque gène, respectivement, p = 0,26 ; p = 0,42 ; p =

0,28 et p = 0,24). Les quantités relatives[13] ont ensuite servies à un calcul de stabilité[14] visant à identifier le couple de gènes dont l'expression est la moins variante. GAPDH et PPIA ont ainsi été identifiés comme le meilleur couple normalisateur. L'incidence de l'ajout des deux autres gènes de ménage à ce couple a été évaluée par un indicateur approprié[15] et a révélé que les quatre gènes pouvaient *in fine* constituer une combinaison robuste de normalisateurs. Nous avons ainsi défini un score RT-PCR dit d'amplificabilité, qui est la moyenne géométrique de l'expression des quatre systèmes référents.

$$\text{Score RT-PCR} = \text{Log}_{10}\left(\sqrt[4]{Q_{\beta\text{ actin}} \cdot Q_{GAPDH} \cdot Q_{RPLPO} \cdot Q_{PPIA}}\right)$$

La valeur de ce score, liée au niveau d'expression de quatre gènes constitutifs, reflète donc indirectement l'état de dégradation de l'échantillon ARN. Nous avons alors pu évaluer la corrélation entre les rendements d'amplification WTO Nano et le score d'amplificabilité pour les 45 échantillons de l'étude clinique.

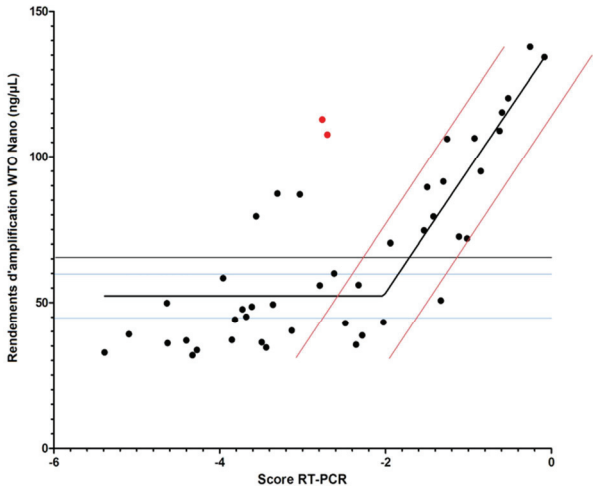


Figure II-29 Valeurs des rendements d'amplifications WTO Nano en fonction des scores de RT-PCR. Chaque point correspond aux valeurs (Score ; WTO) d'un échantillon. La courbe de tendance noire épaisse est obtenue par régression linéaire à deux pentes avec intervalles de confiance à 95 % (lignes bleues et rouges). Les deux échantillons rouges sont des outliers dans ce modèle. Le seuil à 66 ng/µL (ligne noire fine) indique la concentration requise pour une hybridation sur puce Affymetrix.

La régression linéaire à deux pentes appliquée au nuage de points (Figure II-29) semble indiquer qu'au-delà d'une valeur X_0 (-2 ± 0,5), il existe une relation croissante entre le score

[13] La quantité relative Q_x d'un gène x d'un échantillon est donnée par $Q_x = 2^{\Delta Ct}$ où $\Delta Ct = Ct_{min(x)} - Ct_{ech(x)}$.

[14] Le calcul de stabilité correspond à la moyenne arithmétique des variations d'expression des gènes deux à deux. Il nécessite le calcul de la quantité relative. Une valeur inférieure à 1,5 est considérée comme décrivant un état de stabilité d'expression dans un jeu de données (Vandesompele et al. 2002).

[15] L'indicateur de variation évalue l'impact d'une normalisation à *n* ou *n+1* gènes de ménage, par incrémentations successives des gènes les plus proches du couple de référence. Cette méthode permet d'identifier la meilleure combinaison à *n* gènes pour la normalisation d'un jeu de données de RT-PCR quantitative (Vandesompele et al. 2002).

d'amplificabilité et les rendements d'amplification WTO Nano. Le périmètre d'une telle corrélation reste toutefois limité, ce qui exclut d'attribuer au résultat du score RT-PCR une valeur prédictive des rendements d'amplification. Il est ainsi intéressant de faire le constat qu'amplification WTO Nano et amplificabilité RT-PCR ne sont probablement pas le reflet de la même propriété de l'échantillon. Cette distinction, nous le verrons, sera importante dans les étapes d'analyse des puces à ADN de l'étude clinique.

II.4.6.5. Analyse des données des puces HG-U133-PLUS2 et HERV-V2

II.4.6.5.1. Mise en évidence de l'existence de facteurs confondants dans les jeux de données

Les deux jeux des 45 puces à ADN ont été prétraités par la méthode RMA (avec adaptation de l'étape de correction du bruit de fond par l'utilisation du 15^{ième} percentile des sondes tryptophanes), puis la procédure COMBAT[16] a été appliquée pour compenser les effets de lots d'expériences. Avant d'entreprendre une recherche d'expression différentielle entre les classes cliniques, une étude des facteurs confondants a été réalisée dans le cadre du contrôle qualité des puces. Cette étude a pour objectif de révéler des « facteurs cachés » qui influencent la structure des jeux de données. Ces facteurs peuvent être de nature technique ou liés aux données cliniques et risquent de fausser l'interprétation des résultats s'ils restent méconnus. L'analyse SVA[17] a ainsi fait apparaître quatre composantes de variation principales (PC) qui ont été projetées sur les facteurs d'influence étudiés pour en faire ressortir les contributions respectives (Figure II-30).

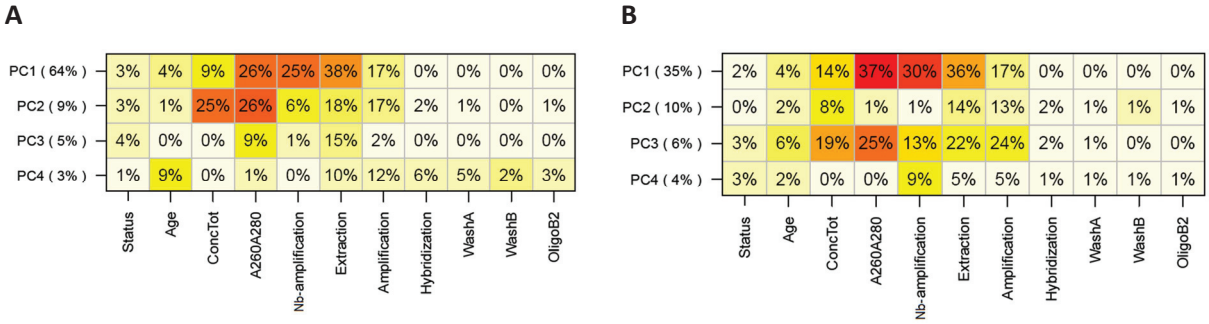


Figure II-30 Analyses en composantes principales des puces de l'étude clinique. Régression des 4 premières composantes principales (PC) des 45 puces (A) HG-U133-PLUS2 et (B) HERV-V2. Facteurs considérés : Status (PBP Neg, GP, PP) et Age sont des données cliniques. ConcTot et A260A280 revoient aux données quantitatives et qualitatives des ARN extraits. Nb-amplification est le nombre d'amplifications réalisées par échantillon. Extraction, Amplification et Hybridization font référence aux dates des expériences. WashA, WashB et OligoB2 sont des données liées aux étapes d'hybridation des puces. Le gradient de couleurs, de blanc à rouge, souligne le poids des facteurs dans la structure des jeux de données.

[16] COMbining BATches. Voir (Johnson et al. 2007).

[17] Surrogate Variable Analysis. Voir (Leek and Storey 2007).

Cette analyse met en évidence qu'une forte composante de variation est associée, pour les deux types de puces, à tous les facteurs liés à l'échantillon ARN (ConcTot, A260A280, Nb-amplification et Extraction). Autrement dit, la nature de l'échantillon utilisé entre en première ligne des effets confondants des jeux de données.

II.4.6.5.2. Recherche de GDE par la méthode SAM avec contrôle du FDR sur le jeu de puces complet

La recherche de gènes différentiellement exprimés (GDE) entre les classes cliniques a été abordée sous l'angle de deux problématiques : une question diagnostique, qui compare les 15 puces du groupe PBP NEG aux 30 puces des deux groupes GP et PP réunis, et une question pronostique, opposant le groupe GP au groupe PP. La méthode SAM avec contrôle du FDR[18] a été appliqué aux jeux de données des puces HG-U133-PLUS2 et HERV-V2 pour ces deux problématiques (Tableau II-13).

A

Seuil Δ	HG-U133-PLUS2		HERV-V2	
	FDR	Nb Probesets	FDR	Nb Probesets
0,1	0,38	4658	1,00	7
0,2	0,38	4658	0,26	2
0,3	0,38	4658	0,26	2
0,4	0,38	4658	0,26	2
0,5	0,1	29	0,00	0
0,6	0,21	11	0,00	0
0,7	0,00	0	0,00	0

B

Seuil Δ	HG-U133-PLUS2		HERV-V2	
	FDR	Nb Probesets	FDR	Nb Probesets
0,1	0,00	0	0,00	0
0,2	0,00	0	0,00	0
0,3	0,00	0	0,00	0
0,4	0,00	0	0,00	0
0,5	0,00	0	0,00	0
0,6	0,00	0	0,00	0
0,7	0,00	0	0,00	0

Tableau II-13 Recherche de probesets présentant une expression différentielle par la méthode SAM avec contrôle du FDR. (A) Question diagnostique. (B) Question pronostique.

Aucune expression différentielle n'a pu être mise en évidence pour la question pronostique, et un nombre très faible de probesets ressort de l'étude diagnostique à des niveaux de faux positifs élevés (11 probesets sous un FDR de 21 % pour la puce HG-U133-PLUS2 et 2 probesets sous un FDR de 26 % pour la puce HERV-V2). En particulier, la forme de la courbe de distribution des *p*-values (non illustrée) est inhabituelle dans ces quatre analyses, ce qui reflète vraisemblablement un niveau d'hétérogénéité dans les données et peut se traduire par des aberrations dans les calculs de FDR. Ces résultats nous ont amené à considérer l'impact des facteurs confondants dans la mise en œuvre de la recherche de GDE, et à le corriger.

[18] Significance Analysis of Microarray et False Discovery Rate. Voir (Tusher et al. 2001) et (Storey and Tibshirani 2003). Egalement plus de détails en annexe II.

II.4.6.5.3. Identification du facteur confondant prépondérant à partir du jeu de puces U133-PLUS2 et définition d'un sous-ensemble de puces analysables

Nous avons pris le parti de résoudre le problème des facteurs confondants par la compréhension de l'hétérogénéité des niveaux d'expression des puces. Cette démarche semblait en effet rationnelle au regard des difficultés d'application des tests statistiques classiques dans un environnement de données irrégulières. Restait à concrétiser cette notion d' « irrégularité » par une vision chiffrée du phénomène. Les puces HG-U133-PLUS2, en tant que référent expérimental, on servi à cela.

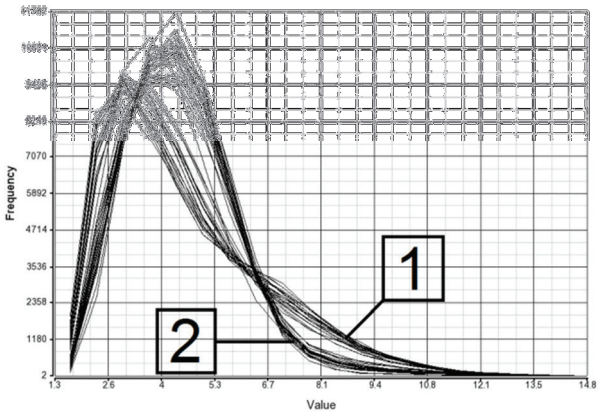


Figure II-31 Courbes de distribution des intensités de signaux des 45 puces HG-U133-PLUS2. Chaque courbe représente la distribution de signal d'une puce. Les valeurs en abscisse indiquent les niveaux d'intensités des probesets en $\log_2$ et l'effectif de probesets associés à ces valeurs est lu en ordonnés.

La distribution des intensités de signaux des 45 puces HG-U133-PLUS2 a fait apparaître l'existence de deux ensembles de puces (Figure II-31). Le premier (profil 1) illustre un bruit de fond centré sur 2^3 puis une décroissance progressive des effectifs jusqu'aux intensités de 2^{12}. Le second ensemble de puces (profil 2) figure un niveau de bruit de fond plus élevé, entre 2^4 et 2^5, suivi d'une décroissance brutale des effectifs entre les intensités 2^5 et 2^8. Pour ce second profil, aucun probeset n'affiche d'intensité de signal supérieure à 2^9. On peut donc dire que les puces du profil 1 donnent des signaux de meilleure qualité que les puces se rattachant au profil 2, puisqu'elles rendent un plus grand nombre de mesures au-delà du bruit de fond. Nous avons alors cherché à rattacher cette observation à un facteur confondant (Figure II-32, page suivante).

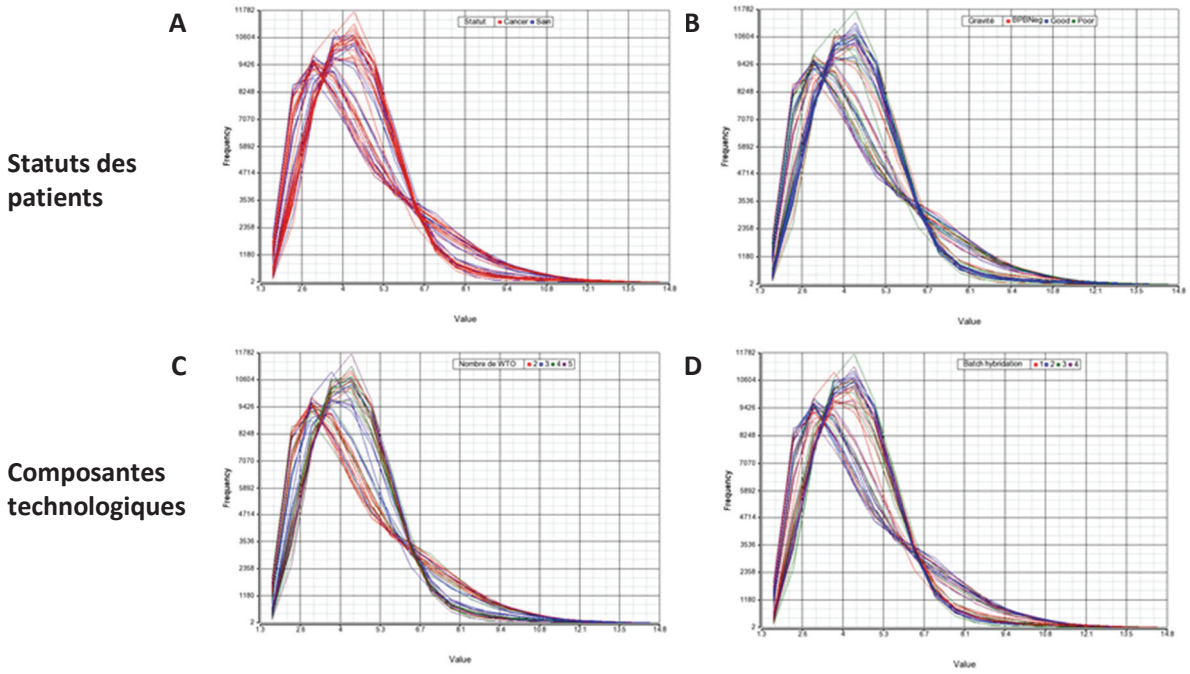


Figure II-32 Courbes de distribution des intensités de signaux des 45 puces HG-U133-PLUS2 colorées par facteurs confondants. (A) Statut diagnostique (PBP NEG : bleu ; GP et GP : rouge). (B) Statut diagnostique et pronostique (PBP NEG : rouge ; GP : bleu ; PP : vert). (C) Nombre d'amplifications WTO Nano (2 : rouge ; 3 : bleu ; 4 : vert ; 5 : violet). (D) Lots d'hybridations (premier lot : rouge ; deuxième lot : bleu ; troisième lot : vert ; quatrième lot : violet).

Aucun des facteurs biologiques (statut diagnostique ou pronostique) ou des facteurs techniques liés au procédé expérimental de réalisation de puces à ADN (nombre d'amplification WTO Nano, lots d'hybridation) n'ont pu corréler à l'observation de deux profils d'intensité, confirmant essentiellement l'existence d'une cause plus fondamentale. L'analyse en composantes principales nous ayant précédemment indiqué que l'origine de cette cause devait avoir trait à l'échantillon ARN d'une part, et un score de RT-PCR, par définition indépendant de la technologie des puces, ayant été mis au point d'autre part, nous avons finalement réussi à identifier un facteur confondant prépondérant au sein du jeu de données (Figure II-33).

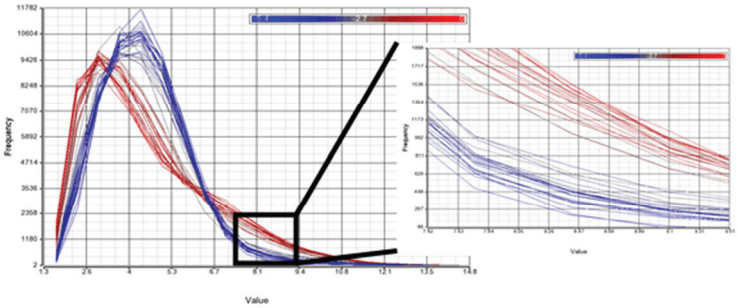


Figure II-33 Courbes de distribution des intensités de signaux des 45 puces HG-U133-PLUS2 colorées par valeurs du score RT-PCR d'amplificabilité. Gradient de couleur de rouge (score = 0) à bleu (score = -5,4).

Les deux profils se séparent nettement en fonction de la valeur du score de RT-PCR. Ainsi, le succès dans l'obtention de signaux au-dessus du bruit de fond est directement lié à l'amplificabilité

de l'ARN, c'est-à-dire à la qualité de l'échantillon de départ évaluée par un ensemble de gènes de ménage par une méthode de détection de référence. Toutes les puces ayant un score compris entre 0 et -2,6 appartiennent au profil 1, et rien que ces puces. Le corollaire est que l'on peut désormais utiliser un score de RT-PCR supérieur à -2,6 comme prérequis à la réussite d'une amplification et d'une détection sur puces à ADN, ce qui permet de disqualifier des échantillons de mauvaise qualité très tôt dans le protocole. Il s'agit donc là autant d'une solution technique apportée au problème des facteurs confondants de cette l'étude que d'un procédé alternatif de caractérisation d'échantillons complexes.

II.4.6.5.4. Ré analyse d'un sous-groupe et mise en évidence d'un échantillon atypique

L'identification de deux sous-ensembles de puces a posé la question de l'unification du jeu de données par une méthode corrective de biais appropriée. Cependant, le constat d'une séparation aussi tranchée en deux catégories d'expression nous a fait renoncer à l'application d'un second niveau de correction COMBAT, jugeant le risque de nivellement, voire de perte, d'éventuels signaux d'intérêt trop important. Ainsi, les puces appartenant au profil d'intensité numéro 2 ont été exclues et une analyse a été refaite sur le nouvel ensemble de 21 puces (8 PBPNEG, 6 GP et 7 PP). Compte-tenu des effectifs à présent réduits au sein des groupes, et au regard des tendances de résultats apportées par l'approche SAM-FDR sur la problématique du pronostique (cf. II.4.6.5.2), seule la question diagnostique sera dorénavant envisagée. La procédure de prétraitement des données reste identique à celle appliquée précédemment.

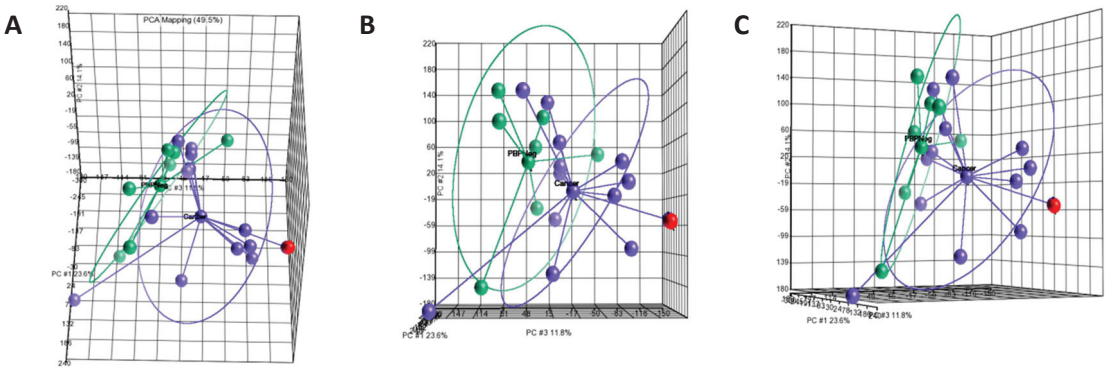


Figure II-34 Représentation 3D de l'analyse en composantes principales des 21 puces HG-U133-PLUS2 du jeu réduit. L'espace de représentation (*x*,*y*,*z*) est défini par les trois plus fortes composantes principales du jeu de données, PC#1 (23,8 %) PC#2 (14,1 %) et PC#3 (11,8 %), respectivement. Chaque sphère représente le positionnement d'une puce dans ses coordonnées PC. Les puces du groupe PBPNEG sont colorées en vert, les puces du groupe cancer (GP + PP) sont colorées en violet. Pour ces deux groupes, une centroïde (objet virtuel matérialisant le centre du groupe et auquel se rattachent ses puces) et une ellipsoïde (ellipse minimisant les distances entre les puces d'un même groupe) sont ajoutées. La puce rouge correspond au patient 461 et n'est volontairement rattachée à aucun des deux groupes dans cette représentation (NB : ses coordonnées superposent quasiment celles d'une puce cancer, en violet, derrière). (A) (B) et (C) : trois vues de l'espace. De (A) vers (B) : rotation selon *z*. De (B) vers (C) : rotation selon *y*.

Le positionnement des 21 puces dans leur espace PC représente un comportement global des signaux et à ce titre est un premier indicateur du niveau d'expression différentielle que l'on peut attendre d'une analyse gène à gène (Figure II-34, page précédente). Ici, les groupes de puces PBPNEG et cancer (GP + PP) s'organisent en deux territoires partiellement chevauchants mais dont certains axes semblent pouvoir discriminer les classes cliniques. Cette situation est assez commune dans les approches transcriptomiques qui s'intéressent à un grand nombre de gènes à la fois. On comprend en effet que, au sein d'un même organisme, deux types cellulaires différents aient davantage de gènes transcrits en commun que d'expressions tissu-spécifiques. Toutefois, cette représentation en composantes principales a clairement positionné une puce PBPNEG dans le territoire des puces cancer. Ce patient (numéro 461, en rouge sur la figure ci-dessus) a, comme tous les autres patients inclus dans ce groupe, eu deux séries de biopsies négatives à l'hôpital Lyon-Sud avant d'être recruté dans cette étude. Nous rappelons que le risque de faux négatifs associé à deux séries de biopsies négatives est évalué empiriquement par les médecins à 5 %.

II.4.6.5.5. Identification de probesets présentant une expression différentielle associée à la question diagnostique

La méthodologie SAM-FDR de recherche d'expression différentielle s'est heurtée au problème des faibles effectifs et n'a pas réussi à identifier d'expression différentielle sur ce jeu réduit. Rappelons en effet que l'on ne dispose à présent que de peu d'échantillons par groupe et que par conséquent les incertitudes associées aux méthodes de contrôle du taux de faux positifs sont importantes. Nous avons donc réalisé une recherche de GDE par la méthode statistique plus traditionnelle de l'ANOVA[19], appliquée entre les deux groupes PBPNEG et cancer. Un double critère de sélection *p*-value inférieure à 0,05 et ratio des médianes inter-groupes supérieur à ± 2 a été appliqué pour obtenir une liste de probesets d'intérêt.

	HG-U133-PLUS2	**HERV-V2**
21 puces (incl. 461)	86	0
20 puces (excl. 461)	440	19

Tableau II-14 Nombre de probesets différentiellement exprimés pour la question diagnostique. Le double critère de la *p*-value < 0,05 et du ratio des médianes entre les groupes PBPNEG et cancer > ± 2 est appliqué aux deux répertoires nucléiques HG-U133-PLUS2 et HERV-V2 après un test statistique prenant ou non en compte le patient 461.

Deux analyses parallèles, prenant ou non en compte le patient atypique 461 (c'est à dire : soit en l'intégrant dans le groupe PBPNEG, soit en le retirant complètement de l'analyse statistique), ont été réalisées (Tableau II-14). Le nombre de probesets différentiellement exprimés qui en découle

[19] ANalysis Of VAriance. Dans le cas de la comparaison de deux groupes, le *F*-test de l'ANOVA devient équivalent à un *t*-test par la relation $F = t^2$.

s'en trouve substantiellement affecté. Notablement, retirer le patient atypique de l'analyse permet l'identification de 19 probesets HERV venant au crédit d'un effectif resté nul autrement. Vis-à-vis du répertoire des gènes conventionnels, exclure le patient 461 multiplie par cinq le nombre de probesets identifiés, en gonflant l'effectif de 86 à 440. Il est à noter que ces 86 probesets sont inclusifs des 440 et qu'il s'agit donc d'un gain strict. Un tel écart dans l'identification de GDE n'a été observé pour aucune autre des combinaisons de retrait d'échantillons évaluées en contrôle. Par conséquent, l'impact de l'échantillon atypique dans la recherche de signaux différentiels associés à la question diagnostique induit une perte d'information et pose la question de la validité du statut clinique de ce patient. Dans l'attente d'une troisième série de biopsies, nous avons donc décidé, au moment de cette analyse, de poursuivre la démarche d'identification et de validation à partir des résultats statistiques obtenus sur 20 puces. L'échantillon atypique sera malgré tout conservé dans le jeu d'expérience et les données qui lui sont associées présentées avec les autres.

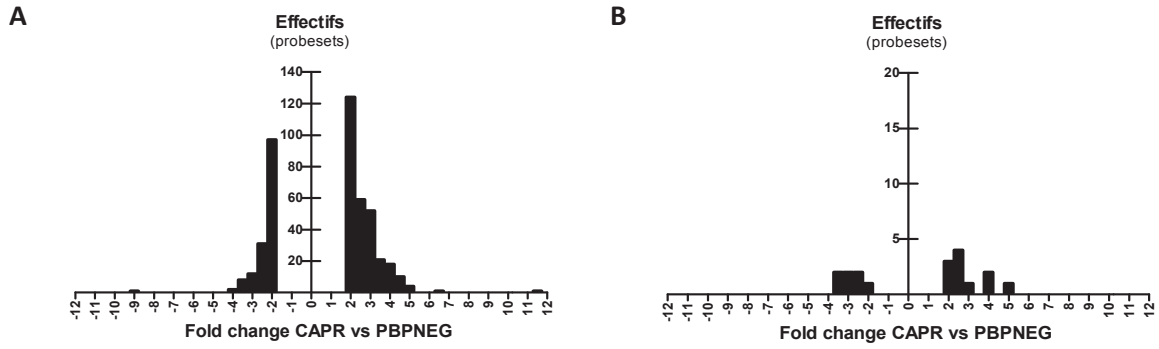


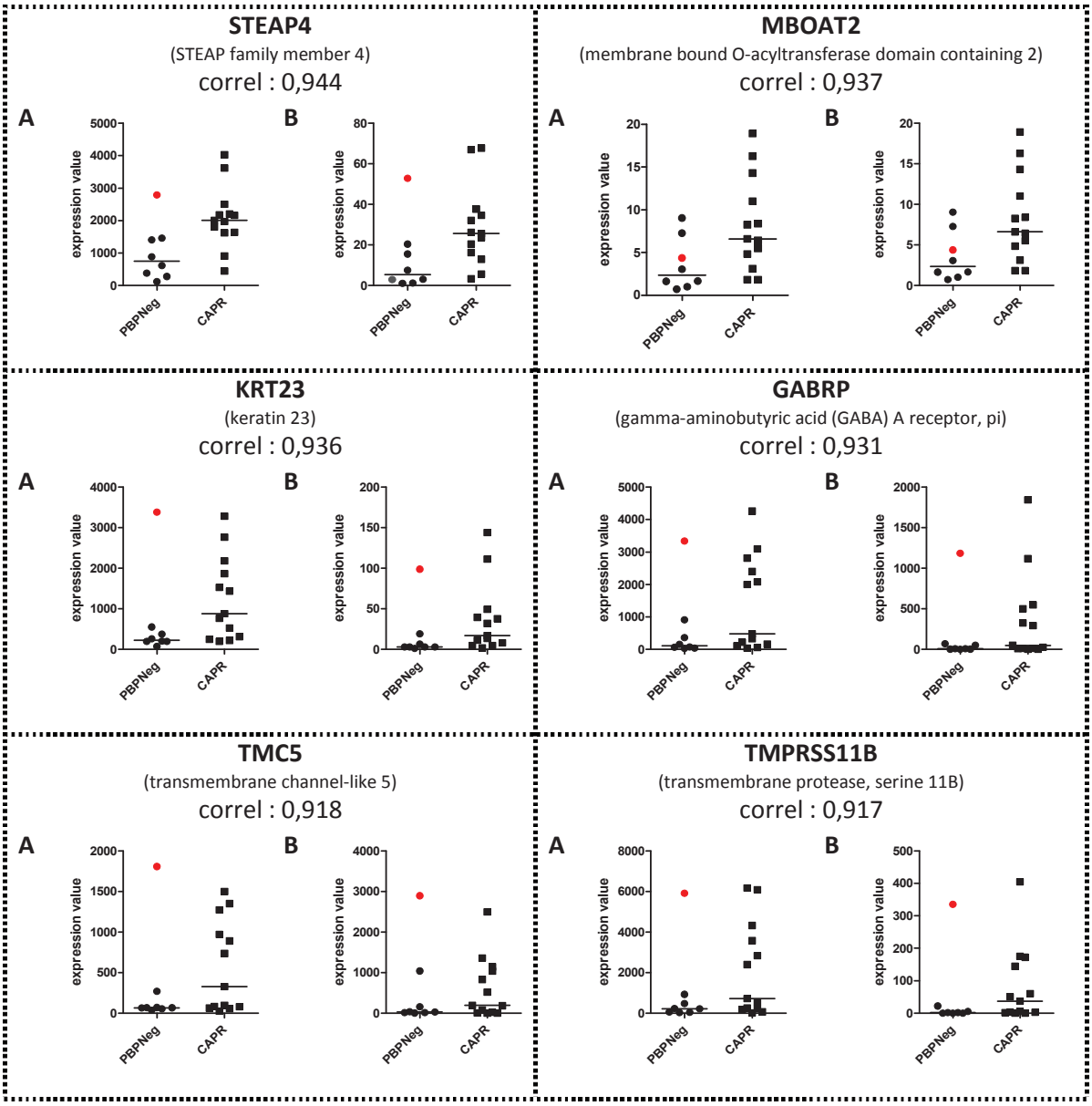
Figure II-35 Effectifs de GDE et ratios d'expression. Les probesets identifiés par l'analyse statistique sur 20 puces et répondant au double critère de la *p*-value < 0,05 et du ratio des médianes entre les groupes PBPNEG et cancer > ± 2 sont dénombrés par intervalles de ratios d'expression (pas : 0,5). (A) Les 440 probesets identifiés sur la puce HG-U133-PLUS2. (B) Les 19 probesets identifiés sur la puce HERV-V2.

La représentation des effectifs de probesets en fonction des ratios d'expression (Figure II-35) indique que, sauf cas uniques isolés, les variations observées entre les classes se jouent sur des facteurs de 2 à 5. Cette représentation, basée sur les médianes d'expression entre les classes cliniques, ne tient cependant pas compte de la variation qui peut exister entre les échantillons au sein des groupes.

II.4.6.5.6. Représentation de l'expression différentielle pour une sélection de probesets, validation des résultats par RT-PCR et description des séquences

Un ensemble de gènes associés à une expression différentielle a été choisi à partir de ces résultats sur puces, en privilégiant essentiellement les écarts de moyennes d'expression entre les classes et une faible variabilité intra-classes. D'autres critères comme la signification biologique des gènes ou l'accessibilité à des systèmes de validation RT-PCR commerciaux ont pu entrer en considération pour une part de la sélection. En effet, seules les expressions différentielles confirmées

en RT-PCR sont définitivement admises comme résultat compte tenu du faible niveau de recul dont nous disposons sur une étude transcriptomique de cette nature. Ainsi, la validation des résultats de la puce HG-U133-PLUS2 s'est faite à l'aide de systèmes de PCR Taqman, et la validation des séquences HERV a reposé sur la conception et l'utilisation de primers spécifiques par la méthode de RT-PCR HRM (cf. II.2.2). Les résultats des puces et des expériences de RT-PCR sont présentés en vis-à-vis, accompagnés d'un coefficient de corrélation (Figure II-36).



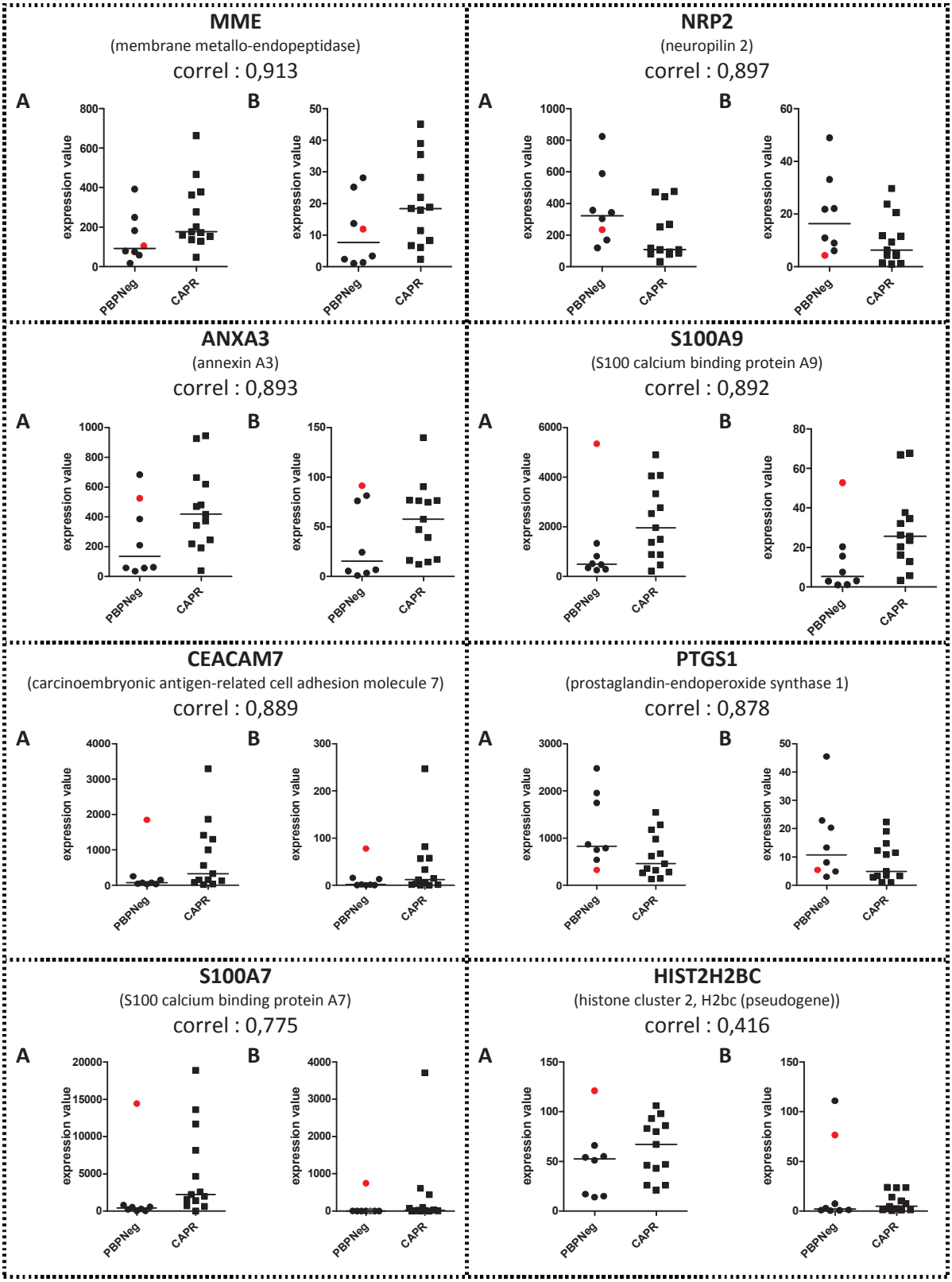


Figure II-36 Niveaux d'expression inter individus sur puces et en RT-PCR pour 14 gènes associés à la question diagnostique. La valeur d'expression associée à chaque patient est représentée par un cercle noir pour le groupe PBPNEG et par un carré noir pour le groupe cancer. Le cercle rouge représente le patient atypique 461. Les barres horizontales indiquent les valeurs moyennes des groupes. Correl = $Cov_{puce;RT\text{-}PCR}$ / (σ_{puce} x $\sigma_{RT\text{-}PCR}$). (A) Intensités normalisées de la puce HG-U133-PLUS2. (B) Intensités normalisées de RT-PCR.

Tous les probesets représentés et que l'on a cherché à valider en RT-PCR répondent au double critère d'une *p*-value < 5 % et d'un ratio d'expression entre les groupes PBPNEG et cancer supérieur à ± 2. Cependant la représentation des valeurs d'expression au niveau des échantillons individuels fait apparaître des distributions intra-classes souvent chevauchantes (MME, ANXA3, MBOAT2 par exemple). Ainsi, dans les cas de surexpression, aucun gène n'est à lui seul parfaitement discriminant de l'état clinique. En revanche, si l'on met de côté de patient atypique (en rouge), il est des gènes dont le niveau d'expression au-dessus d'un certain seuil accompagne spécifiquement les états cancéreux (KRT23, GABRP, S100A9, CEACAM7 entre autres). Une brève description des 14 gènes est fournie (Tableau II-15).

Symbole	Localisation cellulaire	Catégorie	Fonction / description
ANXA3	Cytoplasme / Exosome	Enzyme	Inhibiteur de la phospholipase A2 et rôle anticoagulant
CEACAM7	Membrane plasmique	Autre	Antigène CGM2
GABRP	Membrane plasmique	Canal ionique	Inhibition rapide de la transmission synaptique
HIST2H2BC	Noyau	Autre	Cluster d'histones, composant central du nucléosome
KRT23	Cytoplasme	Autre	Filament de cytokératine
MBOAT2	Inconnu	Enzyme	Acyltransferase de lysophospholipides
MME	Membrane plasmique	Peptidase	Clivage des peptides sur les résidus hydrophobes
NRP2	Membrane plasmique	Kinase	Récepteur de VEGF-165, VEGF-145 et PLGF-2
PTGS1	Cytoplasme	Enzyme	Cyclooxygénase, synthèse de prostaglandine
S100A7	Cytoplasme	Autre	Liaison au calcium
S100A9	Cytoplasme	Autre	Liaison au calcium et activité antimicrobienne
STEAP4	Membrane plasmique	Enzyme	Metalloréductase et antigène prostatique
TMC5	Inconnu	Autre	Transporteur ionique potentiel
TMPRSS11B	Membrane	Enzyme	Sérine protéase

Tableau II-15 Descriptif des 14 gènes dont l'expression différentielle a été validée en RT-PCR. La localisation cellulaire et la catégorie sont adaptées à partir de Ingenuity Pathway Analysis. Les fonctions principales sont recensées par GeneCards.

Le même travail de représentation et de validation a été réalisée pour 5 probesets HERV, mesurant l'expression de régions appartenant à 3 loci distincts : 2000045_h, 1900007_h et 100555_e (Figure II-37, page suivante). Des validations ont également été tentées sur le reste de l'effectif des séquences HERV présentant un différentiel d'expression mais ont donné des corrélations trop faibles (inférieures à 0,4) pour constituer un résultat fiable.

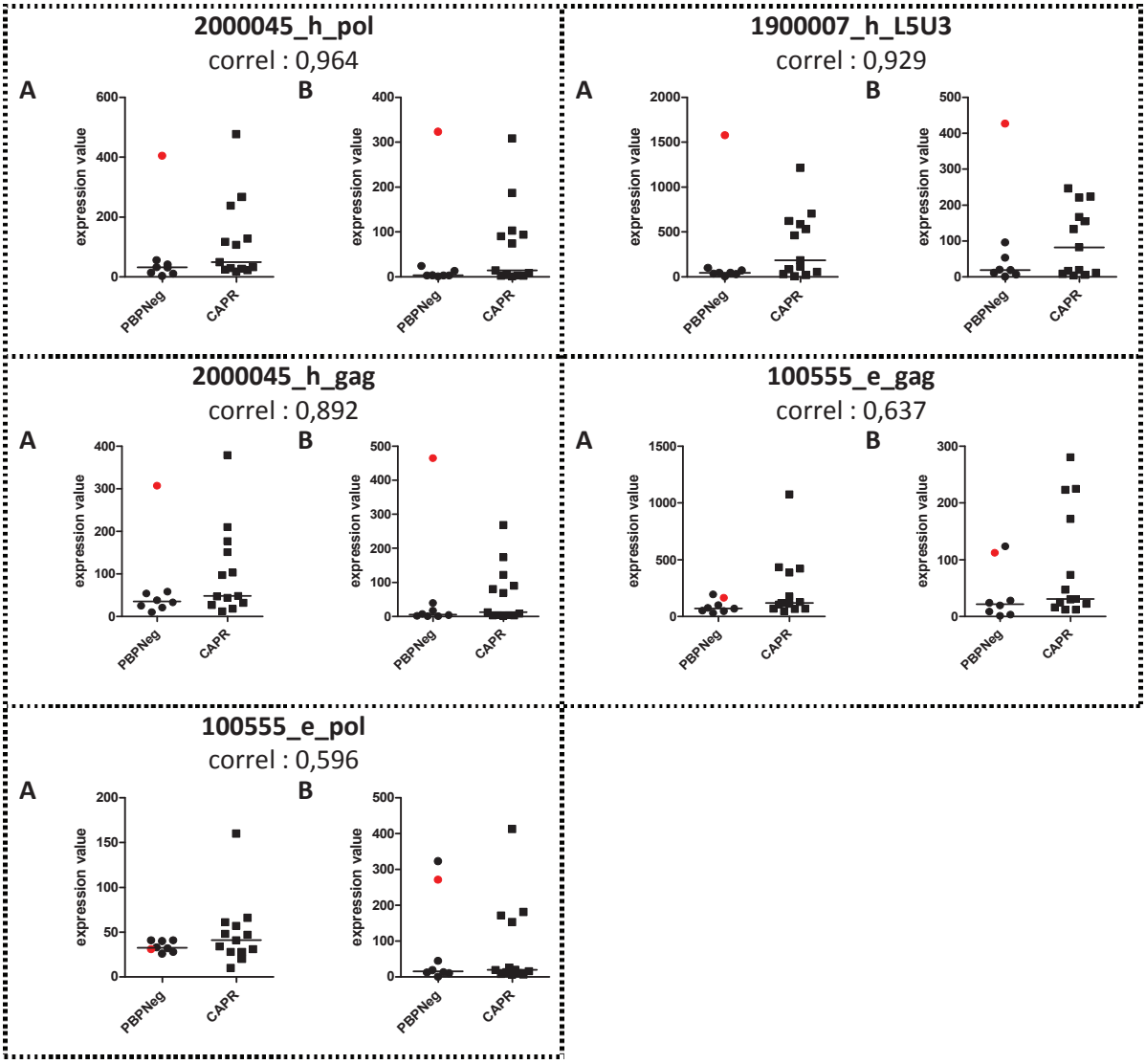


Figure II-37 Niveaux d'expression inter individus sur puces et en RT-PCR pour 5 probesets HERV associés à la question diagnostique. La valeur d'expression associée à chaque patient est représentée par un cercle noir pour le groupe PBPNEG et par un carré noir pour le groupe cancer. Le cercle rouge représente le patient atypique 461. Les barres horizontales indiquent les valeurs moyennes des groupes. Correl = $Cov_{puce;RT\text{-}PCR}$ / (σ_{puce} x $\sigma_{RT\text{-}PCR}$). (A) Intensités normalisées de la puce HERV-V2. (B) Intensités normalisées de RT-PCR.

Le constat réalisé sur l'expression des gènes de la puce HG-U133-PLUS2 s'applique également aux quelques séquences HERV identifiées et validées ci-dessus. Le patient atypique présente ici aussi des niveaux d'expression qui sont assez systématiquement plus proches de ceux du groupe cancer que du groupe PBPNEG et, si aucune des séquences HERV n'offre à elle seule une discrimination parfaite des classes cliniques, certaines semblent pouvoir s'inscrire dans une logique de seuil de positivité lié à un état cancéreux. Les 3 loci HERV sont présentés, ainsi que leur environnement génomique (Tableau II-16, page suivante).

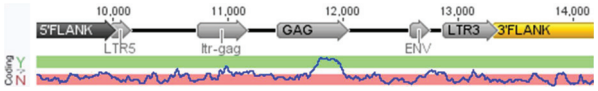
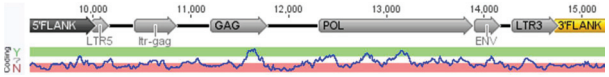
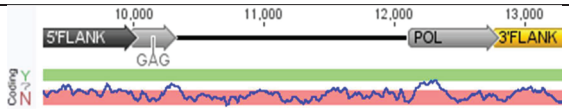
Nom	Loc. génomique	Description copie	Envir. génomique
1900007_h	(+) chr 19 start : 5797895 end : 5801213		[FUT3]* FUT5 [+15,6 kb]* FUT6 [-7,2 kb]* NRTN [-18,6 kb] NDUFA11 [+41,1 kb]*
2000045_h	(-) chr 20 start :19680691 end : 19685420		ø
100555_e	(-) chr 1 start : 204451670 end : 204454441		[C1orf186] CTSE [-29,6 kb]*

Tableau II-16 Descriptif et environnement génomique des 3 copies HERV pour lesquelles l'expression différentielle a été validée en RT-PCR. Nom : identifiant unique du locus HERV. Loc. génomique : coordonnées NCBI36/hg18. Description copie : les régions du locus HERV sont représentées en gris clair, entre leurs régions flanquantes 5' et 3'. La numérotation (bp) du locus HERV commence arbitrairement à 10 000. Une prédiction *in silico* des capacités codantes est indiquée sous l'élément HERV (Y : yes ; N : no). Envir. génomique : obtenu à partir des annotations de la table RefGene (UCSC) sur une distance de ± 50 kb à partir des bornes du locus HERV. Si le locus HERV est intronique, le gène de son environnement est indiqué entre crochets (exemple : FUT3). Sinon, la distance qui sépare le locus HERV du gène, ainsi que le positionnement du locus HERV par rapport au gène, sont indiqués (+ : gène en 3' du locus HERV. - : gène en 5' du locus HERV. * : gène en orientation antisens par rapport au locus HERV).

Ces trois copies HERV présentent des structures provirales partiellement conservées. Des insertions d'éléments de type SINE ont pu être décrites entre les séquences *gag* et *pol* du locus 1900007_h ; la région *pol* du locus 2000045_h a subi une duplication ; les séquences LTR, quand elles existent, n'ont pas conservé une intégrité suffisante pour permettre une définition de probesets qui apporte une lecture fonctionnelle de la transcription. Notre étude du transcriptome HERV sur un panel de tissus (cf. II.3) a précédemment montré que ces trois copies sont actives dans différents contextes cellulaires. Le locus 1900007_h a été trouvé surexprimé dans le cancer de l'ovaire. Le locus 2000045-h a un tropisme d'expression fortement associé au cancer du côlon, et le locus 100555_e est entré dans la composition d'un groupe d'expression hétérogène relié à la prostate (tissu sain et tumoral), l'ovaire (sain et tumoral) et l'utérus (sain). Indépendamment des limitations inhérentes à la nature et au nombre des échantillons de tissus utilisés pour l'étude du transcriptome HERV, il est intéressant de constater que les éléments HERV qui sont surexprimés dans les urines de patients ayant un cancer de la prostate ont tous les trois été associés avec des états cancéreux.

II.4.6.6. Recherche de liens fonctionnels entre les gènes et les séquences HERV identifiés dans l'étude clinique

Pour aller plus loin dans la signification biologique des transcrits détectés dans les urines, nous avons cherché à établir des relations fonctionnelles entre les séquences HERV et leurs gènes environnants d'une part, et entre les différents gènes cellulaires entre eux d'autre part. Comme nous l'avons vu, les LTR des 3 loci HERV n'ont pas permis de mettre en évidence une fonction d'initiation ou de terminaison de la transcription. Nous avons donc basé la recherche d'interactions HERV-gènes

sur l'étude des niveaux d'expression respectifs du locus 1900007_h (puce HERV-V2) et de ses gènes environnants FUT3, FUT6 et NDUFA11 (puce HG-U133-PLUS2) (Figure II-38). Les niveaux d'expression de FUT5 et NRTN (pour 1900007_h) ainsi que C1orf186 et CTSE (pour 100555_e), sous la barre du bruit du fond, ne nous ont pas donné la possibilité d'étendre cette démarche.

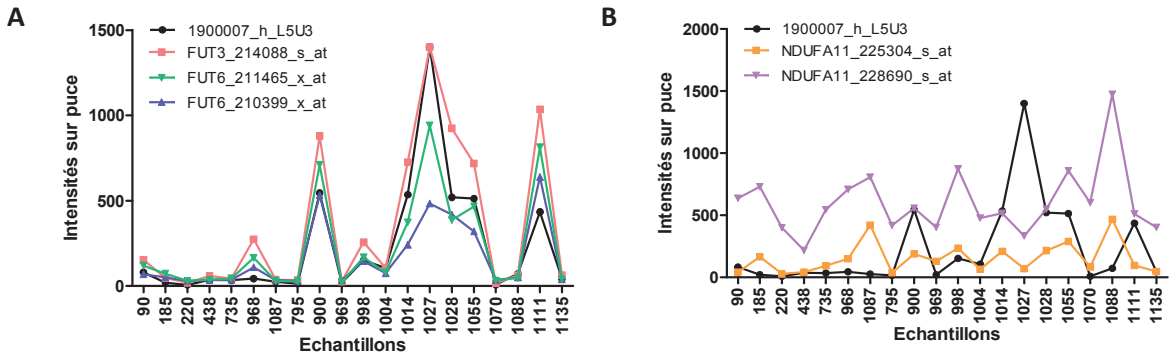


Figure II-38 Corrélations d'expressions entre le locus HERV 1900007_h et ses gènes environnants. Les valeurs d'expression du locus 1900007_h lues sur la puce HERV-V2 et celles des probesets des gènes FUT3, FUT6 et NDUFA11 lues sur la puce HG-U133-PLUS2 sont indiquées pour les 20 patients de l'étude clinique (hors patient 461). (A) Valeurs corrélées. (B) Absence de corrélation entre la séquence HERV et les gènes environnants.

L'existence d'une corrélation d'expression assez forte entre le locus 1900007_h et les probesets de FUT3 et FUT6 ($correl_{HERV\text{-}FUT3}$ = 0,931 ; $correl_{HERV\text{-}FUT6}$ = 0,897 et 0,790) par rapport aux probesets de NDUFA11 ($correl_{HERV\text{-}NDUFA11}$ = 0,021 et -0,170) peut laisser penser que ces éléments sont co-régulés. La relative proximité à ± 50 kb de toutes les séquences codantes est un argument pour exclure l'hypothèse d'une zone euchromatinienne permissive ou d'un embarquement transcriptionnel général conséquence d'une ouverture de brin locale. Il semble donc possible qu'une interaction de transcription spécifique ait lieu entre le locus 1900007_h et un ou des gènes FUT de son environnement. Les membres de la famille des fucosyltransferases (FUT) ont pour fonction l'ajout de fucose aux polysaccharides sanguins avant qu'ils ne soient absorbés par les hématies (ils entrent dans la composition du groupe sanguin de Lewis). Plusieurs variants d'épissage de FUT3 ont été précédemment détectés (Cameron et al. 1995), ce qui peut laisser penser que la présence intronique du locus HERV 1900007_h participe à un phénomène d'épissage alternatif, par un mécanisme qui reste à élucider.

La question des relations fonctionnelles entre les gènes cellulaires a ensuite été adressée par la construction de réseaux d'activités en association à des voies de signalisation canoniques. En se basant sur un niveau de démonstration expérimentale, un ensemble de liens fonctionnels a été établi entre trois des gènes identifiés précédemment (MME, S100A7 et S100A9) et des intermédiaires jouant un rôle, notamment, dans la division cellulaire (Figure II-39, page suivante). Il en va ainsi des antagonistes PTEN/Akt, qui, par une voie de transduction intracellulaire tyrosine-kinase dépendante, participent à la régulation de l'apoptose, de la croissance cellulaire et de l'angiogenèse. Plus

particulièrement, la phosphorylation du récepteur à androgène (AR) par la voie Akt a été associée à la survie de cellules de prostate cancéreuses et ce mécanisme pourrait être central dans l'établissement des phases hormono-indépendantes de la maladie (Lin et al. 2003). D'autres facteurs de transcription, impliqués dans des voies de signalisation ubiquitaires (Frizzle/Wnt pour CTNNB1, cytokines pour NFκB), ont également pu être reliés au réseau.

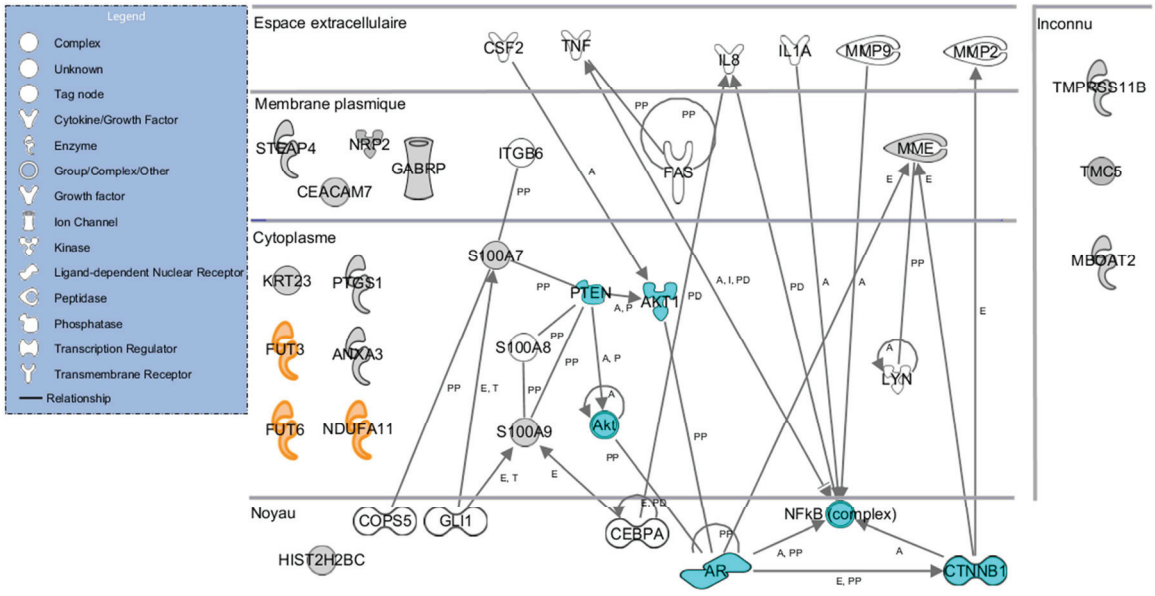


Figure II-39 Liens fonctionnels entre les gènes identifiés. A partir des 14 produits des gènes validés (gris) et des 3 produits des gènes de l'environnement des loci HERV (orange), 3 réseaux fonctionnels, impliquant un ensemble de molécules plus large (blanc), ont été identifiés puis fusionnés. Le résultat des interactions directes est présenté dans un référentiel cellulaire. Les éléments du réseau faisant partie de la voie de signalisation du cancer de la prostate sont colorés en cyan. A : activation ; E : expression ; I : inhibition ; P : phosphorylation ; PD : Interaction protéine-DNA ; PP : interaction protéine-protéine ; T : transcription.

Ainsi, les liens de fonctions qui ressortent du petit nombre de gènes d'intérêt de l'étude clinique tournent autour des voies de signalisation impliquées dans des mécanismes génériques de régulation de l'expression, de division cellulaire et de réponse aux androgènes, ce qui pourrait indiquer que, parmi les gènes détectés dans les urines, nous avons identifié des candidats en association à des processus de cancérisation.

II.4.7. Bilan de l'axe clinique

Au terme de cet axe de recherche de marqueurs transcriptomiques du cancer de la prostate dans les urines, nous avons donc (i) mis au point un protocole de recueil et des procédés de traitement des échantillons compatibles avec la technologie des puces à ADN (ii) développé un outil de qualification des ARN basé sur un score de RT-PCR, alternative aux méthodes classiques (mais inadaptées ici) de dosages et de caractérisation des acides nucléiques (iii) mis en évidence et validé une expression différentielle entre deux groupes de patients pour un ensemble de gènes et de

séquences HERV (iv) révélé un comportement atypique pour un patient du groupe contrôle, par une vision exclusivement moléculaire[20] et (v) avancé une hypothèse de co-régulation HERV-gènes et brièvement étayé la vision d'un processus biologique par la signification des gènes et de leurs interactions. Ce dernier point, très anecdotique dans le cadre de cette étude clinique, est l'objet d'une approche exploratoire sur lignées cellulaires qui conclura la partie expérimentale de la thèse.

II.5. Mise en œuvre d'une démarche exploratoire dans la recherche d'interactions des transcriptomes HERV et HG-U133-PLUS2

II.5.1. Choix d'un modèle cellulaire de transformation et d'invasion tumorale

Les modèles cellulaires offrent une source abondante de matériel d'étude biologique. Dans le but d'aller plus loin dans l'analyse de la dérégulation des rétrovirus endogènes humains en contexte pathologique et des interactions avec les voies métaboliques cellulaires, mais aussi comme une extension à l'approche clinique pour la recherche de marqueurs moléculaires, nous avons choisi d'exploiter le modèle de Webber (Webber et al. 2001) qui mime les étapes de progression du cancer de la prostate. A partir de la lignée humaine non-néoplasique RWPE1, quatre lignées tumorigènes indépendantes ont été obtenues par exposition au N-methyl-N-nitrosourea (MNU). Les propriétés d'invasion de chacune des cellules ont été caractérisées par transplantation chez la souris par les auteurs du modèle. Les cellules RWPE1 n'induisent pas de tumeur, alors que les quatre lignées exposées à l'agent chimique MNU se multiplient *in vivo* et développent une tumeur avec chacune un potentiel d'invasion propre. Ces lignées filles indépendantes peuvent alors être classées par leur caractère malin croissant comme suit : WPE1-NA22, WPE1-NB14, WPE1-NB11 et WPE1-NB26. La lignée WPE1-NA22 (la moins invasive) forme des petites tumeurs bien différenciées, alors qu'à l'autre bout du spectre la lignée WPE1-NB26 forme de grosses tumeurs peu différenciées et très invasives. En plus de ces cinq types cellulaires, une lignée tumorigène, appelée RWPE2 et obtenue par transformation de RWPE1 par l'oncogène v-Ki-ras, a également été retenue dans notre étude pour ses propriétés d'invasion *in vitro* (Bello et al. 1997). Ainsi, un ensemble d'évènements de transformation indépendants et de plusieurs natures (chimique ou oncogénique) permet d'aborder

[20] Au moment de la rédaction de ce manuscrit il s'est écoulé environ une année depuis l'identification du patient atypique. Une troisième série de biopsies vient tout juste d'être réalisée sur ce patient, révélant la présence d'un foyer cancéreux de Gleason 6. Ce point est repris dans la discussion générale.

la question de la dérégulation des répertoires HERV et HG-U133-PLUS2 dans un modèle d'établissement et de progression tumorale.

II.5.2. Mise en évidence d'une expression différentielle HERV et HG-U133-PLUS2 en association à des voies métaboliques

Nous avons mis en culture ces six lignées, puis extrait leur ARN et réalisé en parallèle une puce HERV-V2 et une puce HG-U133-PLUS2 pour chacune d'entre elles (voir paragraphe II.3.1 pour rappels méthodologiques sur le procédé expérimental de réalisation de puces à ADN utilisé dans ce travail). Une recherche de gènes différentiellement exprimés (GDE) a été réalisée en comparant tour à tour la lignée saine RWPE1 à chacune des lignées transformées (soit un total de 5 comparaisons effectuées pour les deux types de puces). Des critères analytiques différents ont dû être appliqués d'un répertoire nucléique à l'autre pour réussir à mettre en évidence des GDE : un FDR de 20 % a été retenu pour les puces HERV-V2 alors qu'un FDR de 5 % a été suffisant pour identifier de nombreux évènements d'expression différentielle à partir des puces HG-U133-PLUS2 (Tableau II-17) (se référer à l'annexe II pour un complément méthodologique sur le contrôle du FDR). Ces différences reflètent intrinsèquement des niveaux de variations plus faibles au sein du répertoire des rétrovirus endogènes. Puis, à partir des listes des gènes cellulaires, une recherche des voies métaboliques les plus présentes a été réalisée en utilisant l'ontologie KEGG implémentée dans l'outil d'annotation fonctionnelle du programme bioinformatique DAVID[21]. Pour chaque couple de comparaison, le nombre de gènes par voie est indiqué dans le tableau II-17.

	RWPE1 vs NA22	**RWPE1 vs NB14**	**RWPE1 vs NB11**	**RWPE1 vs NB26**	**RWPE1 vs RWPE2**
Nombre d'éléments différentiellement exprimés					
HERV (HERV-V2) – FDR 20 %	32	67	53	81	214
Gènes cellulaires (HG-U133-PLUS2) – FDR 5 %	623	625	1957	421	195
Principales voies métaboliques KEGG associées					
Adherens junction				7	
Alanine, aspartate and glutamate metabolism					3
Arginine and proline metabolism			12		
Biosynthesis of unsaturated fatty acids		4			3
Bladder cancer		5			
Butanoate metabolism			8		
Cell adhesion molecules (CAMs)				10	
Citrate cycle (TCA cycle)			8		
Colorectal cancer	8				
Cytokine-cytokine receptor interaction				12	
ECM-receptor interaction	8		17		

[21] Database for Annotation, Visualization and Integrated Discovery : http://david.abcc.ncifcrf.gov

Focal adhesion			33		
Hematopoietic cell lineage				7	5
Lysine degradation			10		
MAPK signaling pathway				14	
p53 signaling pathway	8	9	19		
Pancreatic cancer	7				
Pathways in cancer	27	24	60	18	
Prion diseases		5	8	5	
Pyrimidine metabolism			16		
RNA degradation	6		12		
Small cell lung cancer	8		18		
Steroid biosynthesis	6		7		
Synthesis and degradation of ketone bodies	3	3	4		
TGF-beta signaling pathway	9				5
Valine, leucine and isoleucine biosynthesis	4				
Valine, leucine and isoleucine degradation	6	5	11		

Tableau II-17 Expression différentielle HERV et HG-U133-PLUS2 et voies métaboliques principales identifiées à partir des modèles de cultures cellulaires. Pour chaque couple de comparaison, le nombre de loci HERV et de gènes cellulaires présentant une expression différentielle est indiqué. A partir des identifiants de la puce HG-U133-PLUS2, une recherche des voies métaboliques principales a été réalisée en utilisant l'ontologie KEGG. Les chiffres inscrits dans les cases indiquent le nombre de gènes cellulaires différentiellement exprimés ayant été associés à la voie métabolique considérée.

D'un point de vue moléculaire, le modèle de Webber ne semble pas s'inscrire dans la progression continue qui avait été décrite phénotypiquement chez la souris. En effet, le nombre de GDE augmente jusqu'à l'état WPE1-NB11, en conservant une relative cohérence d'un point de vue métabolique vis-à-vis des états moins invasifs (par exemple : les voies *p53 signaling pathway* et *Pathways in cancer* sont communes et incluent un nombre globalement croissant de gènes), puis une rupture semble marquer le passage à l'état WPE1-NB26, pour lequel le nombre de GDE identifiés chute d'un facteur 4,5 fois et où un nouveau profil métabolique semble s'installer (de nouvelles voies apparaissent parmi lesquelles *Cell adhesion molecules*, *Cytokine-cytokine receptor interaction*, *MAPK signaling pathway* ou encore *Hematopoietic cell lineage*, et des voies précédemment prépondérantes se réduisent ou ne sont plus identifiées du tout). D'autre part, il est intéressant de constater que la transformation oncogénique aboutissant à la lignée RWPE2 donne le seul exemple d'un nombre de séquences HERV réactivées plus élevé que le nombre de GDE cellulaire (il faut toutefois garder à l'esprit que les critères du FDR ne sont pas les mêmes), et les voies métaboliques impliquées ne se rattachent pas clairement à l'une ou l'autre des phases du modèle de Webber, possiblement en raison du plus faible nombre de gènes cellulaires inclus.

II.5.3. Principe d'association fonctionnelle des répertoires nucléiques

Ainsi, compte-tenu de la rupture dans les phases du développement tumoral caractérisée par la transition de RWPE1 à WPE1-NB26, nous avons choisi d'illustrer, par cet exemple, la mise en place d'une approche exploratoire pour la recherche d'interactions entre les transcriptomes HERV et HG-U133-PLUS2. Notre approche a consisté à co-localiser les évènements d'expression différentielle des deux répertoires nucléiques et à y superposer la lecture fonctionnelle des voies métaboliques. L'outil de représentation Circos[22] a été utilisé pour cela (Figure II-40).

Figure II-40 Représentation génomique circulaire des différentiels d'expressions HERV et HG-U133-PLUS2 et des liens fonctionnels KEGG identifiés dans la comparaison RWPE1 vs WPE1-NB26. Le caryotype humain est représenté en cercle avec indication des cytobandes, tout en préservant les proportions relatives des chromosomes. A l'extérieur du cercle génomique, deux niveaux de représentations co-localisent les valeurs de différentiel d'expression de la puce HERV-V2 (face la plus extérieure) et de la puce HG-U133-PLUS2 (la plus proche des chromosomes). Voir la figure II-41 pour détails. Au centre de l'image, les liens d'une même couleur relient entre eux les localisations des gènes qui appartiennent à une même voie KEGG : Pathways in cancer (gris) ; Cell adhesion molecules (rouge) ; Prion diseases (vert) ; Adherens junction (bleu) ; MAPK signaling pathway (violet) ; Hematopoietic cell lineage (jaune) et Cytokine-cytokine receptor interaction (orange).

[22] http://circos.ca

Une des manières d'explorer cette représentation complexe est de chercher à associer une co-localisation d'évènements de réactivations HERV et HG-U133-PLUS2 à une voie métabolique d'intérêt. Nous illustrerons cette méthodologie par un exemple sur le chromosome 7 (Figure II-41), en lien avec la voie 'Cell adhesion molecules' spécifiquement associée à l'état invasif WPE1-NB26.

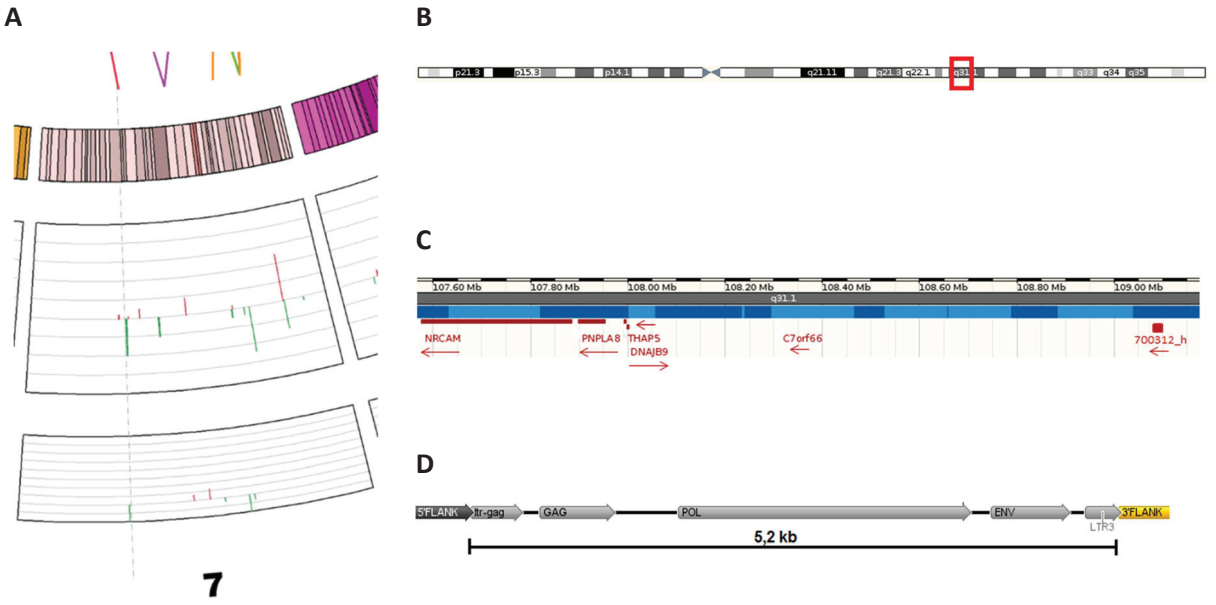


Figure II-41 Exemple d'une recherche de co-localisation fonctionnelle. (A) Grossissement de la figure II-40 sur le chromosome 7 et les valeurs d'expression associées aux puces HERV-V2 (en bas) et HG-U133-PLUS2 (au centre). Les valeurs des différentiels d'expressions entre les lignées RWPE1 et WPE1-NB26 sont représentées sous forme de bâtons. Vert : surexpression dans la lignée WPE1-NB26 ; rouge : sous-expression dans la lignée WPE1-NB26. Le trait pointillé gris aligne une voie KEGG (ici : Cell adhesion molecules) avec un gène cellulaire surexprimé appartenant à cette voie (NRCAM) et une séquence HERV surexprimée co-localisée (700312_h). (B) Représentation Ensembl du chromosome 7. La boite rouge indique la région contenant NRCAM et la séquence HERV 700312_h. (C) Description Ensembl de la région NRCAM-HERV. L'orientation des flèches indique le sens de la transcription. (D) Représentation de la séquence HERV 700312_h, d'après les bases de données du laboratoire.

Sur la base de la proximité génomique, la réactivation de NRCAM (*neuronal cell adhesion molecule*), qui est un gène impliqué dans la voie KEGG 'Cell adhesion molecules', a pu être visualisée aux côtés d'une séquence HERV elle aussi réactivée et appelée 700312_h (Figure II-41 A). L'environnement génomique, étudié localement, a précisé que 700312_h se situe en amont du gène NRCAM à une distance d'environ 1,2 Mb (Figure II-41 C). Cette configuration est incompatible avec le fait que 700312_h puisse jouer un rôle de promoteur alternatif (les promoteurs alternatifs HERV les plus lointains ayant été identifiés se situent à quelques dizaines de kb du gène), mais il n'est pas exclu que cette séquence HERV puisse recruter des facteurs de transcription et des protéines activatrices pour finalement contribuer à la régulation génique en tant qu'enhancer. Il n'a malheureusement pas été possible d'aller plus loin durant ce travail, et la caractérisation d'un tel mécanisme, ainsi que la mise en évidence systématique d'associations fonctionnelles HERV-gène, devra faire l'objet d'études plus approfondies. La base méthodologique amorcée ici par l'outil de représentation Circos semble pouvoir aider à progresser dans ce sens.

III Discussion générale

III.1. Apports à la compréhension de la biologie des HERV

Les rétrovirus endogènes constituent 8 % de l'ADN humain et représentent quelques 200 000 séquences indépendantes, dispersées sur le génome. Leurs capacités à provoquer des phénomènes de recombinaison chromosomique et à moduler l'expression des gènes cellulaires par l'entremise des séquences LTRs les désignent souvent comme des éléments parasites dont l'expression est réprimée par l'hôte en contexte physiologique. La domestication d'une séquence HERV au profil du développement de l'organisme est un évènement remarquable et s'inscrit dans une logique de contrôle de l'expression à plusieurs niveaux, comme cela a été présenté au travers de l'exemple de la Syncytine-1 en introduction. La génomique comparative, permise progressivement par le séquençage de différentes espèces animales, indique que des domestications d'enveloppes rétrovirales ont également jalonné l'évolution de plusieurs branches de mammifères euthériens et que ces acquisitions indépendantes concourent à la réalisation d'une même fonction[23]. Si l'expression physiologique des HERV est également documentée dans les cellules de la lignée germinale en lien avec des hypothèses de reprogrammation épigénétique et dans certaines glandes sensibles à la régulation hormonale, c'est en contexte pathologique que l'activité des rétrovirus endogènes devient foisonnante. De nombreux cancers (tumeurs solides et lymphomes), mais également certaines maladies auto-immunes et des affections du système nerveux fournissent autant d'exemples de réactivations, au niveau transcriptionnel, protéique, ou allant parfois jusqu'à l'observation de particules. L'étude du transcriptome HERV est cependant essentiellement basée sur le comportement de groupes d'éléments apparentés et n'adresse pas la question de l'expression des loci individuels à grande échelle. Cet état de l'art, s'il découle en partie d'un manque de technologies matures, repose aussi souvent sur un *a priori* théorique qui voudrait que des éléments appartenant à un même groupe phylogénétique présentent, peu ou prou, la même susceptibilité vis-à-vis des facteurs de transcription cellulaires. Or cette vision n'intègre pas la complexité de la dérive génétique, dont les effets fonctionnels s'évaluent difficilement, ni les particularismes des environnements génomiques d'insertion des séquences provirales à l'origine de contextes d'expression singuliers. L'approche développée au cours ce travail de thèse s'est voulue sans *a priori*, en étudiant un peu plus de 5 500 loci HERV distincts, abordés comme autant d'évènements indépendants, et en particulier en ne préjugeant pas des fonctions exercées aujourd'hui par les LTRs au regard de ce qu'elles furent au moment de leur entrée dans le génome humain.

[23] Un portrait comparé des différentes Syncytines connues à ce jour est proposé au lecteur, sous forme de revue, en annexe I.

III.1.1. Portrait d'un nouveau paysage transcriptomique

III.1.1.1. Niveaux d'expression

A l'aide de la puce haute-densité HERV-V2, l'expression des rétrovirus endogènes humains a été étudiée sur un panel d'ARN provenant de tissus adultes somatiques, de tissus de la lignée germinale et d'organes impliqués dans la fonction de reproduction. L'association, quand cela a été possible, de l'état cellulaire sain avec une contrepartie cancéreuse, a été recherchée. Ainsi la double question du tropisme d'expression des HERV dans les tissus humains et de leur réactivation en contexte pathologique a pu être adressée. Ce travail a permis d'identifier 1 718 loci HERV actifs dans au moins un des états cellulaires étudiés. Cet effectif, composé aux trois-quarts de LTRs, représente un tiers du contenu de la puce et suggère, par extrapolation, qu'une proportion similaire des séquences HERV du génome possède un potentiel transcriptionnel. De plus, l'âge de la famille ne semble pas être le facteur déterminant du niveau d'activité des séquences HERV (Figure III-1). En effet, la famille HERV-H, la plus ancienne des 6 familles étudiées par la puce HERV-V2, est celle ayant la plus importante fraction d'éléments actifs (41 %), et les deux familles HERV-E et HERV-FRD, intégrées à plusieurs millions d'années d'intervalle, ont des niveaux d'activité très comparables (22 % et 20 %, respectivement). Ceci laisse penser que l'activité des HERV est devenue en partie indépendante de la conservation des séquences provirales d'origine, probablement par l'acquisition de sites de fixation de facteurs de transcription.

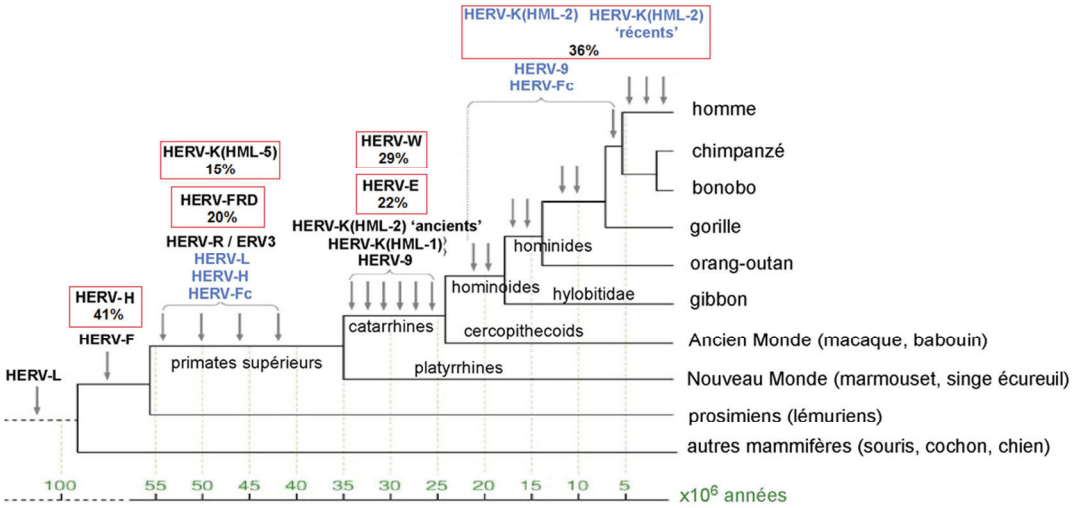


Figure III-1 Entrées dans les génomes des rétrovirus et niveaux d'activité contemporaine chez l'homme. En noir : les infections initiales ; en bleu : les vagues d'amplification importantes de certaines familles. Les 6 familles encadrées en rouge sont celles étudiées par la puce HERV-V2. Le pourcentage indiqué est le rapport du nombre de loci transcrits sur le nombre de loci du génome (projection à partir du contenu de la base HERV-gDB3). Figure adaptée de (Bannert and Kurth 2006) et (Christensen 2005).

III.1.1.2. Tropisme, phylogénie et facteurs de transcription

Neuf profils d'expression ont par la suite été mis en évidence et définissent des tropismes tissulaires. Le fait qu'une famille puisse être surreprésentée dans certains de ces groupes, comme nous l'avons observé pour la famille HERV-H dans le groupe cancer du côlon ou encore pour la famille HERV-E dans l'ensemble de tissus prostate (saine et tumorale) utérus (sain et tumoral) et ovaire tumoral, peut interroger la notion de dérive génétique et de comportements individuels des loci HERV. Pour aller plus loin sur ce point, nous avons tenté des regroupements phylogénétiques des séquences HERV sur la base de leur(s) activité(s) (ou inactivité). En particulier, les séquences LTRs actives et silencieuses de famille HERV-W identifiées dans ce travail ont été regroupées dans un arbre (Figure III-2).

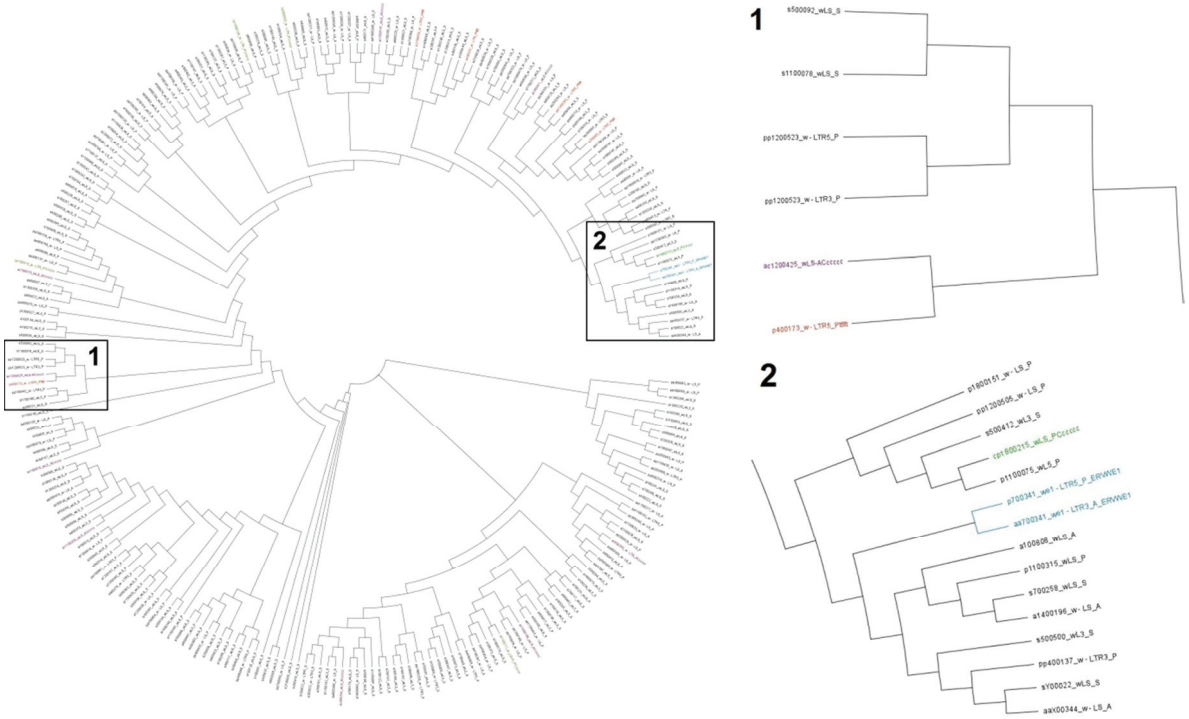
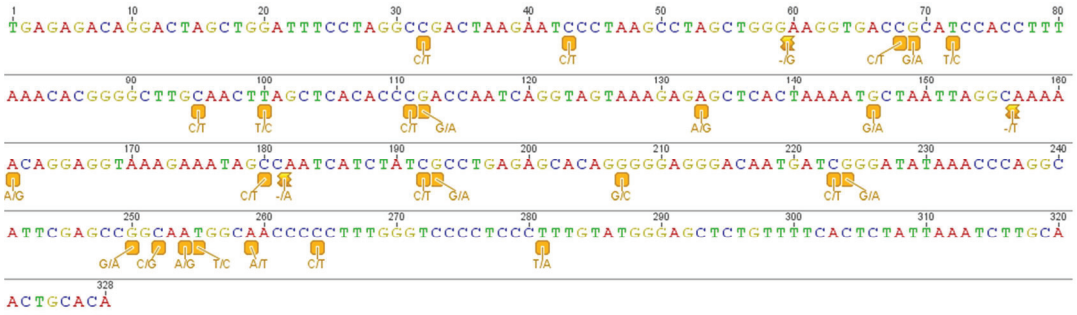


Figure III-2 Phylogénie des séquences LTRs HERV-W actives et silencieuses. Un arbre phylogénétique sans racine a été construit à partir des séquences des LTRs de la famille HERV-W pour lesquelles une fonction a pu être attribuée dans l'étude sur tissus. Les préfixes/suffixes de la nomenclature des séquences indiquent leur nature transcriptionnelle : promoteur (p/P), polyA (a/A), silencieux (s/S) ou expression constitutive (c/Ccccc). Les promoteurs constitutifs sont représentés en vert et les polyA constitutifs en violet. En rouge : séquence promotrice dans le testicule tumoral exclusivement ; en bleu : les LTRs 5' et 3' du locus ERVWE1/Syncytine-1.

Ainsi, au niveau de l'ensemble des LTRs de la famille HERW-W comme au niveau de sous-groupes de l'arbre, les séquences promotrices, polyA, silencieuses ou ayant une expression constitutive (noms de séquences commençant respectivement par p, a, s ou c) sont mélangées, ce qui semble indiquer qu'il n'existe pas de groupe phylogénétique clair au sein des LTRs de la famille HERV-W en lien avec un type d'activité transcriptionnelle. Ceci confirme que les mécanismes d'activation de l'expression des HERV sont complexes et, encore une fois, renvoie à un besoin d'analyse plus fine (même théorique) de la fixation de facteurs de transcription. Dans une tentative

pour approcher cette question, nous avons réalisé un premier niveau d'analyse du polymorphisme des séquences U3-R des LTRs de la famille HERV-W. Un alignement des LTRs HERV-W actives et inactives a mis en évidence 28 SNPs (Figure III-3 A), définis par un minimum de 25 % de séquences polymorphiques sur la base considérée. Nous avons ensuite étudié la composantes de deux d'entre eux au regard de l'activité fonctionnelle des LTRs (Figure III-3 B), l'un en position 146 de type G/A et l'autre en position 142 (CAAT box) correspondant à un polymorphisme d'insertion.

A



B

	Base 146			Base 182		
	A	G	Autre	-	A	Autre
Consensus des LTRs	43,7%	55,2%	1,1%	63,6%	36,4%	0,0%
LTRs promotrices	33,3%	65,4%	1,2%	72,8%	27,2%	0,0%
LTRs polyA	51,2%	44,2%	4,7%	55,8%	44,2%	0,0%
LTRs silencieuses	47,4%	52,6%	0,0%	60,6%	39,4%	0,0%

Figure III-3 Polymorphisme de la région U3-R des LTRs HERV-W. (A) Séquence consensus de la région U3-R des LTRs HERV-W actives et inactives. Les sites polymorphiques sont indiqués par un carré orange. On mentionnera également la CAAT box (position 181) et la TATAA box (position 228). (B) Etude du polymorphisme des bases 146 et 182 pour les LTRs actives (promotrices et polyA) et silencieuses.

Cette approche a permis de mettre en évidence des variations quantitatives dans la composition des SNPs, en association à un type d'activité fonctionnelle. La base 146 est un G pour 65 % des LTRs promotrices contre 44 % pour les LTRs polyA, ce qui représente, respectivement, 10 points de plus et de moins que la valeur du consensus obtenue à partir de l'ensemble des LTRs actives et inactives. Par ailleurs, l'insertion d'une base A en position 182, dans la CAAT box, semble plus fortement associée aux LTRs polyA (44 % d'entre elles) qu'aux LTRs promotrices (27 %). Si l'on peut donc conclure qu'il n'est pas possible d'identifier de manière caricaturale des sites de fixation de facteurs de transcription qui découleraient de règles de conservation de séquences au sein des catégories de fonctions, il semble en revanche exister des tendances de comportement à l'échelle des bases individuelles. Une telle observation suggère que l'étude de la combinatoire des SNPs aidera à définir des critères structurels prédictifs de l'activité des HERV.

III.1.1.3. Expression constitutive

Les évènements de transcription sont principalement des réactivations et ne décrivent en conséquence qu'une fraction de l'activité transcriptionnelle totale. En allant plus loin et comme évoqué plus haut, nous avons défini un groupe de 463 loci présentant une forte expression quelque

soit le contexte biologique (le 10ième profil dans l'article PLos ONE). Si une expression basale de familles HERV a précédemment été détectée dans plusieurs tissus sains et notamment les cellules de la lignée germinale (Ahn and Kim 2009; Buzdin et al. 2006a; Seifarth et al. 2005), il reste communément admis que l'activité des HERV en contexte physiologique est réprimée par l'hôte. En montrant une expression constitutive des HERV, pour la première fois à ce niveau de description, nous pouvons remettre en perspective la pensée dominante avec des résultats amenés progressivement par la génomique fonctionnelle. Le projet ENCODE par exemple, dont le but est l'identification de tous les éléments fonctionnels du génome humain, a apporté en 2007 un regard nouveau sur l'ADN non codant après avoir accompli sa phase pilote sur 1% du génome (ENCODE Project 2007). La conclusion la plus retentissante de ce travail a probablement été que le génome est transcrit de manière permissive, c'est-à-dire que des phénomènes de transcription ont lieu à peu près en tout point, et non pas seulement dans les régions codantes, à un moment ou à un autre. Cela conforte nos observations sur les HERV mais, plus encore, les inscrit dans une perspective fonctionnelle quand ENCODE identifie de nouveaux sites d'initiation de la transcription dans des régions supposées jusque-là silencieuses et ajoute en point d'orgue qu'une masse d'éléments fonctionnels serait entretenue à distance des gènes par les contraintes de pression de sélection.

III.1.2. De nouveaux éléments fonctionnels en attente de nouvelles fonctions

De ce fait, la suggestion qu'il y ait besoin de définir autrement la notion de promoteur cellulaire entraine des changements d'appréciation de la contribution de l'ADN non codant à la régulation génique. Dans ce cadre, les séquences LTRs, en tant que promoteurs naturels des rétrovirus, ont une place à occuper. En utilisant la dichotomie des signaux U3/U5 de la puce HERV-V2, nous avons pu identifier 326 LTRs promotrices, 209 LTRs polyA et 672 LTRs silencieuses. En particulier, nous avons montré que (i) une LTR active est une LTR spécialisée dans une unique fonction (ii) la densité en gènes cellulaires est 1,5 fois plus élevée dans l'environnement des LTRs actives que dans l'environnement des LTRs silencieuses et (iii) la configuration de l'environnement cellulaire diffère entre les LTRs actives et silencieuses sur la zone de 8 kb en amont des LTRs promotrices. Ces descriptions s'accordent bien avec l'idée que les LTRs puissent être des éléments fonctionnellement intégrés à un équilibre génétique. La vision exclusivement parasitique et délétère des HERV peut donc être repensée sous un angle plus proche de l'idée de symbiose, mais dont les tenants et aboutissants à l'échelle d'un génome entier restent encore largement à préciser, quoi qu'ils semblent d'ores et déjà ne pas pouvoir s'inscrire facilement dans les schémas classiques de connaissance. Il faut par exemple, une fois des séquences promotrices et polyA identifiées, caractériser les transcrits complets qui leur sont associés. Sur ce point nous avons pu constater, sur la

base de nos expériences, qu'il n'existe que très peu de cas d'association d'une LTR fonctionnelle ayant dans son environnement immédiat et orienté comme attendu un gène cellulaire qui pourrait tenir lieu de partenaire dans un transcrit hybride. Qu'initient donc (ou que terminent donc) l'écrasante majorité des LTRs qui ont une autonomie transcriptionnelle ?

III.1.2.1. Interférence ARN

Un élément de réponse peut venir des mécanismes d'interférence transcriptionnelle. L'interférence ARN opère une régulation négative et module ainsi l'expression génique. Elle peut faire intervenir plusieurs types de petits ARN, qui se définissent par rapport à leur voie de biogenèse. Chez la souris, il a été montré qu'une fraction d'ARN interférent d'origine cellulaire correspond à des éléments transposables de types LINE, SINE et ERV et que l'ajout, à un ARN messager synthétique, d'une séquence 3'UTR dérivée de ces rétrotransposons, rend instable l'ARNm une fois injecté dans des cellules en culture (Watanabe et al. 2006). Chez l'homme, deux LTRs HERV-K HML-2 sont localisées dans les introns des gènes SLC4A8 et IFT172, respectivement, mais en orientation antisens par rapport aux gènes. Il a été montré que l'activité promotrice de ces LTRs génère des transcrits ARN antisens dans les séquences exoniques des gènes, et que ce mécanisme conduit à une baisse d'expression des gènes correspondants (Gogvadze et al. 2009). Bien qu'aucun travail n'ait été mené sur ce point, on peut ainsi penser que l'autonomie d'expression de LTRs qui se trouveraient loin des gènes mais à proximité, par exemple, de zones riches en miRNA, participerait indirectement à la régulation de l'expression des gènes par des mécanismes d'interférence.

III.1.2.2. Transcription antisens

Nous avons abordé la fonctionnalité des LTRs selon une lecture directionnelle classique d'initiation dans la région U3 et de prolongement de la transcription dans la région U5 et au-delà. Cependant, la puce HERV-V2 du laboratoire a été conçue en intégrant systématiquement des sondes de capture de polarité positive et négative, et le procédé d'amplification WTO Nano utilisé conserve théoriquement la polarité des ARNm initiaux. Il a donc été possible d'étiqueter 'sens' ou 'antisens' les signaux observés dans notre étude du transcriptome HERV mais ce point n'a finalement pas été mentionné dans les résultats tant l'interprétation biologique a été jugée périlleuse. Nous avons en effet observé, sur l'ensemble des milliers de probesets présentant une détection positive dans un tissu ou un autre, que les signaux associés à une transcription antisens représentent 44 % des occurrences (56 % pour les signaux de transcription sens). Un biais de la méthode d'amplification associée au protocole de réalisation de puces à ADN a été envisagé, mais il semble pourtant clair que le déséquilibre d'effectif sens/antisens observé, sur un échantillonnage de cette taille, ne peut découler d'un effet purement aléatoire (p-value < 10^{-10}), ce qui eut dû être la conséquence d'un

procédé d'amplification qui ne conserverait pas la polarité des brins d'ARNm. Tout de même, depuis quelques années, l'existence de transcrits de polarité négative initiés à partir de la LTR 3' des rétrovirus HIV (Michael et al. 1994) et HTLV (Gaudray et al. 2002) a été rapportée. Si les mécanismes précis de cette transcription et de son contrôle restent à approfondir, il a par la suite été précisé que des phénomènes d'épissages ont lieu à partir de la LTR 3' (Cavanagh et al. 2006), que les facteurs de transcription JunD et Sp1 sont nécessaires à la fonction promotrice (Gazon et al. 2012) et que l'ARN antisens ainsi produit pourrait constituer un nouveau mécanisme de contrôle de la réplication virale (Kobayashi-Ishihara et al. 2012). De manière intéressante, l'existence de transcrits de polarité négative ne semble pas limitée aux rétrovirus complexes ni à leur LTR 3' provirale. L'induction de l'expression de MLV dans des lymphocytes B et T a été associée à des transcrits de polarité négative initiés dans la LTR 5' et poursuivis dans les proto-oncogènes cellulaires Jdp2 et Bach2, sous forme de transcrits chimériques (Rasmussen et al. 2010). De telles observations supportent l'idée d'un potentiel de transcription bidirectionnelle par des LTRs beta- et gammarétrovirales endogènes. Quelques travaux ont également documenté la transcription antisens HERV, pour la famille HERV-K (Domansky et al. 2000; Mack et al. 2004), HERV-H (Feuchter et al. 1992) et HERV-W (Lee et al. 2003), mais dans des proportions qui n'atteignent pas les observations que nous avons faites. Une démarche pour aller plus loin, à partir de nos données d'expression, dans la caractérisation directionnelle de la transcription HERV et de ses implications fonctionnelles, pourrait être d'étudier systématiquement l'environnement génomique des LTRs qui génèrent supposément des transcrits antisens de manière constitutive. Puis, une fois des hypothèses de régulation HERV-gène posées sur un modèle, l'étude d'un ensemble d'interactions transcriptionnelles théoriques pouvant exister entre la LTR et ses gènes voisins (par la conception de systèmes de PCR *ad hoc*) pourrait permettre de confirmer ou d'infirmer l'existence, par exemple, de transcrits alternatifs ou chimériques.

III.1.2.3. Unifier les regards portés à différents niveaux

Il est enfin d'autres contributions des LTRs à la régulation de l'activité génique qui pourront constituer autant d'axes complémentaires pour prolonger ce travail de thèse. Les propriétés d'enhancer, par exemple (qui parfois s'additionnent à celles de promoteurs, comme nous l'avons vu sur l'exemple du locus contrôle de la beta globine (Ling et al. 2002)), sont à rechercher par d'autres approches que la seule proximité génomique des LTRs et des gènes. L'étude des modifications épigénétiques à grande échelle et notamment l'analyse des marques histones, remise en perspective avec des réseaux fonctionnels de gènes, peut faire apparaitre des ensembles d'éléments co-régulés et aider à attribuer des fonctions d'enhancer à des LTRs actives. L'étude de la génétique, quant à elle, peut permettre l'identification de sites de recombinaisons ou de translocations, à l'image de la fusion ETV1 (chromosome 7) et HERV-K17 (chromosome 17) dans le cancer de la prostate (Hermans et al.

2008). Plus que des séquences embarquées, les LTRs, en tant qu'éléments transposables, ont le potentiel de provoquer des réarrangements chromosomiques et leur expression peut tout à fait être un prérequis à la recombinaison, comme cela est montré pour les transposons RAG1 et RAG2 impliqués dans la recombinaison VDJ des immunoglobulines (Kapitonov and Jurka 2005). Il ressort de ces suggestions que les avancées à venir dans la compréhension des fonctions des LTRs viendront probablement de l'intégration de différents niveaux de recherche, dont le transcriptome peut être une voie d'entrée. Travailler dans ce sens pourra, *in fine*, conduire à multiplier les descriptions qui combinent plusieurs niveaux de connaissance et fournissent des modèles rapidement transférables en clinique. La réactivation transcriptionnelle, dans les lymphocytes B, d'une LTR, suite à une levée de méthylation, et qui conduit à un transcrit de fusion impliquant l'oncogène cellulaire CSF1R (Lamprecht et al. 2010) est, pour les lymphomes de Hodgkin, un exemple abouti de ce qui peut être ambitionné pour différentes pathologies (Figure III-4).

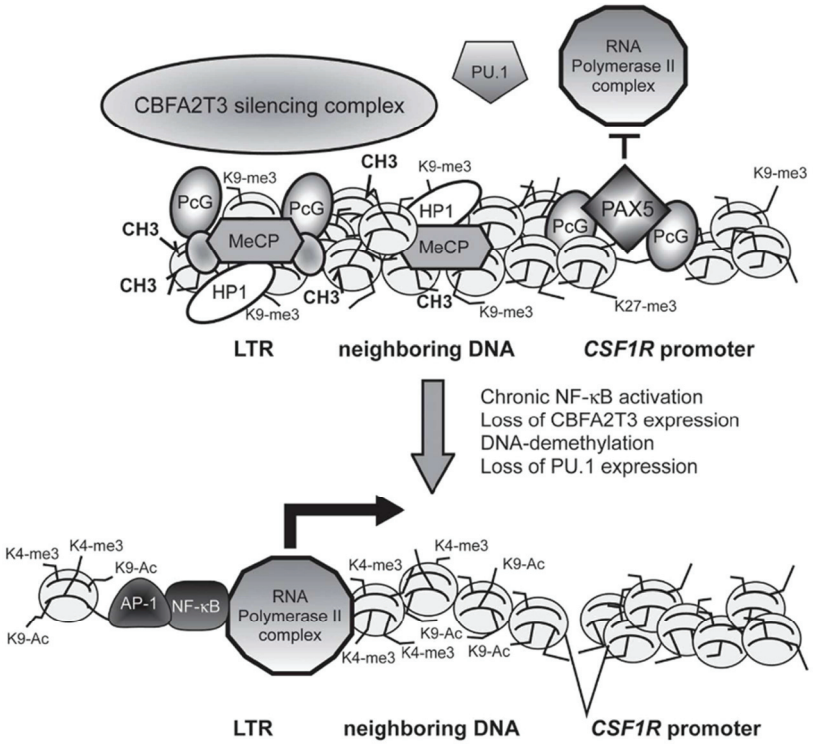


Figure III-4 Modèle de la réactivation aberrante d'une LTR endogène dans le lymphome de Hodgkin. Partie supérieure : la région amont du locus CSF1R, incluant une LTR, est organisée en hétérochromatine compacte caractérisée par une méthylation des sites CpG et accompagnée de marques histones inactives telles que la triméthylation de H3K9. Cet état engage le recrutement de protéines de fixation aux groupements méthyles (MeCP), des complexes polycomb (PcG) et de la protéine HP1. L'état hétérochromatinien de la LTR est maintenu par le complexe répressif CBFA2T3 et le promoteur naturel de CSF1R est réprimé par le facteur de transcription PAX5, bloquant ainsi le recrutement de l'ARN polymérase II médié par PU.1. Partie inférieure : l'expression de PU.1 et CBFA2T3 disparait, l'état chromatinien silencieux n'est plus maintenu et l'ADN est déméthylé. De plus, des facteurs de transcription inductibles comme NFκB et AP-1 sont activés de façon chronique et se lient à la LTR. La chromatine est remodelée, conduisant au recrutement de l'ARN polymérase II sur la séquence LTR et à l'activation d'une transcription qui by-pass le promoteur naturel. D'après (Lamprecht et al. 2010).

III.2. Le délicat passage en clinique

L'étude du transcriptome que nous avons réalisée n'a pas atteint le niveau de maturité qui vient d'être évoqué, mais a cependant fourni une liste d'éléments HERV réactivés, de manière autonome au non, dans différents états cancéreux. Ainsi l'expression d'une centaine de probesets HERV a été associée au cancer de la prostate, plus de 300 probesets sont ressortis dans le cancer du côlon, plus de 600 dans le cancer de l'ovaire et la barre des 1 000 est franchie pour le séminome. Ces chiffres découlent d'une analyse statistique et il est évidemment très difficile d'évaluer la proportion d'information ayant une réelle pertinence biologique. On peut par exemple redouter l'existence d'une forte proportion de signaux assimilables à du bruit de fond, mais également craindre d'avoir observé des variations qui ne sont pas liées à l'état pathologique ou bien encore que ce qui a été montré sur les échantillons recrutés en nombre limité pour une étude donnée ne soit pas représentatif d'une population plus large. Néanmoins l'ensemble des séquences candidates dont nous disposons constitue une base de connaissance pour renforcer la démarche d'identification de marqueurs de cancers. Par ailleurs, le recours à des procédures dédiées (Ptolemy and Rifai 2010), incluant notamment une étude de faisabilité et une phase pilote avant d'engager de grandes cohortes, permet de pondérer les risques inhérents aux enjeux d'une recherche clinique. Mais le meilleur garde-fou à des études inutilement couteuses reste probablement l'effort fourni en amont pour positionner un sujet dans une thématique de santé publique opportune. Nous avons pour cela développé un partenariat de recherche avec les Hospices Civils de Lyon dans le but de définir et de répondre aux besoins de marqueurs du cancer de la prostate.

III.2.1. La transcriptomique dans un environnement complexe

III.2.1.1. Gestion des référents méthodologiques

III.2.1.1.1. Maitrise des échantillons

L'accès aux prélèvements biologiques d'un hôpital et la possibilité de redéfinir des pratiques de recrutement en fonction de besoins de recherche spécifiques est un atout considérable dans les initiatives de recherches cliniques de biomarqueurs. En travaillant avec les services d'urologie et d'anatomopathologie de l'hôpital, nous avons mis au point un protocole de recrutement des urines post massage prostatique qui nous a permis de conduire des expériences de transcriptomique à haut débit. En particulier, une standardisation des gestes médicaux et des mouvements de transfert des échantillons au sein des services de l'hôpital a été effectuée pour assurer une constance dans les

modalités de recrutement au cours du temps. Aboutir à de telles habitudes n'a rien d'automatique mais découle d'ajustements et de mises au point prises à l'initiative des différents acteurs du projet. L'exemple de l'ajout de RNA Cell Protect aux urines fraichement centrifugées, initialement exclu pour préserver une double approche de protéomique et de transcriptomique dans des projets du laboratoire, est représentatif de ces démarches. Au-delà de l'aspect technique, le regard du praticien est décisif s'agissant des critères à appliquer pour constituer des classes cliniques susceptibles de répondre à une question médicale. Il a ainsi été décidé de ne retenir que des patients ayant subi au moins deux séries de biopsies négatives pour constituer le groupe contrôle, et seuls des prélèvements dont le pronostic a été rendu sur pièces de prostatectomies radicales ont été retenus pour délimiter des classes de gravité de la maladie. Sur ce dernier point, les critères sont essentiellement fondés sur le score de Gleason. Un groupe de bon pronostic a été défini par un score de Gleason inférieur ou égal à 7 (3+4) et le groupe de mauvais pronostic par les scores à partir de 7 (4+3) ou exceptionnellement en deçà dans le cas où la chirurgie n'a pas pu éliminer toute la présence cancéreuse (marges chirurgicales). Il est évident que, les patients les plus avancés dans l'évolution de la maladie n'étant pas opérés, nous ne disposons pas des échantillons à plus mauvais pronostic, ce qui peut constituer une critique au recrutement et éventuellement expliquer la difficulté à mettre en évidence de l'expression différentielle entre les groupes Gp et Pp. Nous défendons néanmoins les choix retenus comme le résultat d'un compromis entre une vitesse de recrutement à assurer (les effectifs des différents groupes doivent être équilibrés), et les chances de succès que l'on peut raisonnablement espérer d'une catégorisation de cette nature.

III.2.1.1.2. De l'échantillon clinique au processus analytique

Un suivi méthodologique à double niveau a été mis en place pour encadrer l'étude clinique sur cellules urinaires. D'une part, des puces HG-U133-PLUS2 ont été systématiquement réalisées en parallèle des puces HERV-V2 dans le but de disposer d'une méthode de référence en transcriptomique. D'autre part, l'expression de quelques gènes, choisis pour leur association à l'organe ou leur pouvoir normalisateur, a été quantifiée en RT-PCR afin d'ancrer les niveaux d'expression. L'association qui découle, d'un score empirique de RT-PCR aux profils d'expression des puces HG-133-PLUS2, a permis de résoudre le principal facteur confondant de l'expérience et de définir un critère objectif de validation d'une expérience transcriptomique menée à partir de très faibles quantités de matériel nucléique. Dans ce cadre, le profil d'expression d'un échantillon (numéro 461) est ressorti de manière atypique, en apparaissant plus proche du groupe cancer que du groupe des patients négatifs dans lequel il avait été classé. En appliquant aux puces HERV-V2 l'indicateur qualité basé sur la RT-PCR et en tenant compte de l'observation faite sur le patient atypique, la recherche d'expression différentielle a abouti à l'identification de quelques séquences

discriminant les classes cliniques. Le récent suivi médical du patient 461 a établi le diagnostic d'un foyer cancéreux de Gleason 6, ce qui semble donner raison *a posteriori* aux développement méthodologiques sur lesquels nous nous sommes appuyés. Ce cancer, soit qu'il ait été manqué par les ponctions de biopsie précédentes (risque estimé à 5 %) soit qu'il se soit développé dans l'intervalle, a bien été « vu » par ses caractéristiques transcriptomiques dans notre étude. Ceci invite à réfléchir à ce que pourraient être des approches de re-stratification de patients, indépendantes des classifications établies par les médecins, visant à identifier des signatures moléculaires qui puissent apporter un niveau de pertinence supplémentaire, voire supplétif, au jugement diagnostique, comme cela a déjà été réalisé dans le cadre d'un Gleason moléculaire (True et al. 2006).

III.2.1.1.3. La statistique à l'épreuve du réel

Plus largement, le recours profitable à des référents méthodologiques, conjugué à des niveaux d'observations élémentaires des données, questionne la validité des méthodes bioinformatiques analytiques appliquées à la recherche de signatures transcriptomiques. Il est communément admis que le déséquilibre entre le nombre de données d'une puce (plusieurs dizaines de milliers de gènes) et le nombre d'observations faites (quelques dizaines d'échantillons par étude) entraine un risque statistique d'erreur, et en particulier un risque de faux positifs. Des méthodes correctives développées pour contrôler le FDR existent (Storey and Tibshirani 2003) mais, en conséquence, réduisent les effectifs retournés. Comme nous l'avons vu par le travail sur tissus, un contrôle du FDR est insuffisant pour se prémunir des signaux à la limite du bruit du fond ; en ce qui concerne l'étude clinique sur cellules urinaires, s'affranchir du contrôle du FDR a été la condition *sine qua non* pour identifier une expression différentielle par la suite validée en RT-PCR. La question des erreurs commises par les tests statistiques semble donc plutôt liée à une querelle entre valeur statistique et grandeur physique d'un phénomène. S'il est vrai qu'un test de Student peut identifier, à tort, des expressions différentielles lorsqu'il est appliqué à la limite d'un bruit de fond, il devient très difficile de défendre que l'observation d'un phénomène tangible, comme un fort niveau d'expression, et fût-il de un pour dix mille, puisse découler d'une erreur théorique liée à une loi des grands nombres. En face d'une manifestation physique, des causes sont à discuter (techniques, biologiques etc.) sans commencer par nier la réalité de l'observation. Il s'agit pourtant d'un amalgame qui semble faire loi dans la recherche de signatures moléculaires aujourd'hui. A titre d'exemple, une étude récente a évalué la valeur pronostique de 47 signatures d'expression dans le cadre du cancer du sein. Ces signatures impliquent un nombre de gènes allant de deux à plus de mille. Pour chacune d'entre elles, mille signatures de même taille ont été générées aléatoirement et leur valeur prédictive a été comparée à celles des signatures publiées. Le constat est accablant : plus de 90 % des signatures aléatoires de plus de cent gènes ont une valeur pronostique comparable aux

combinaisons de gènes identifiées par des approches statistiques (Venet et al. 2011). Faut-il conclure que toute tentative d'identification de signatures moléculaires par les outils statistiques actuels est vouée à l'échec ? Certainement pas, et cette étude provocante démontre également une réelle valeur ajoutée pour les tests Oncotype de l'entreprise Genomic Health (Paik et al. 2004), pour la signature de Van de Vijver dont découle le test MammaPrint de la société Agendia (van de Vijver et al. 2002) et pour le Genomic Grade de Sotiriou exploité par Ipsogen (Sotiriou et al. 2006). Il faut en revanche certainement réviser l'optimisme sans réserve qui a été prêté jusqu'à présent à la puissance de l'informatique dans le traitement des données issues des approches à haut débit. Cela passe à coup sûr par une réappropriation des questionnements biologiques fondamentaux, mais peut également s'accompagner d'une remise en question des tests statistiques utilisés de manière routinière. Il a par exemple été proposé de remplacer l'utilisation du test de Student par une procédure appelée 'limma' pour l'analyse de puces à ADN dans le cas où le nombre d'échantillons biologiques est limité (Jeanmougin et al. 2010), et différents outils, parmi lesquels HTSelf, ont été développés pour appréhender des effectifs de petites tailles tout en apportant des niveaux d'interprétations fonctionnelles (Koide et al. 2006; Silva et al. 2010; Vencio et al. 2006; Vencio and Koide 2005).

III.2.1.2. Niveaux de difficultés dans la validation d'un résultat de recherche

III.2.1.2.1. Variabilité individuelle, variabilité populationnelle

Quelle que soit la méthode conduisant à identifier des marqueurs candidats de pathologies, le passage à une échelle de validation impliquant un grand nombre d'individus est incontournable pour établir l'intérêt clinique d'un résultat de recherche. Dans le cadre du partenariat avec les Hospices Civils de Lyon, nous avons initié un protocole de recherche de marqueurs diagnostiques du cancer de la prostate dans les urines qui, compte tenu des risques, s'articule en plusieurs phases (Ptolemy and Rifai 2010). La faisabilité technique ainsi qu'une première étude clinique ont été réalisées durant ce travail de thèse, ce qui a conduit à l'identification de plusieurs gènes cellulaires et de quelques séquences HERV présentant une expression différentielle entre individus sains et malades. En particulier, l'expression différentielle de trois loci HERV, 2000045_h, 1900007_h et 100555_e, a été confirmée en RT-PCR. De forts niveaux de variation sont cependant observés d'un échantillon à l'autre, ce qui pose la question de la variabilité inter individuelle et, dans le cas particulier de la détection dans les urines, de la normalisation de la fraction d'ARN prostatique. Dans cette suite, une seconde étude clinique, utilisant des échantillons indépendants, est en cours de réalisation au laboratoire pour (i) augmenter le nombre d'échantillons répondant au critère qualitatif mis au point précédemment et basé sur la RT-PCR (ii) valider ou invalider les candidats actuels, qu'il

s'agisse de gènes cellulaires ou de séquences HERV (iii) tester une méthode d'amplification plus sensible pour essayer de mettre en évidence de nouveaux candidats et (iv) répondre à la problématique de la normalisation de quantité d'ARN d'origine prostatique.

Si l'étude clinique sur cellules urinaires du projet cancer de la prostate progresse selon une logique propre dans le sens de l'identification de marqueurs, le travail réalisé sur des tumeurs de côlon a offert un cadre d'étude à l'évaluation de la variabilité inter individuelle de séquences HERV candidates. Sur un recrutement de 28 tumeurs, nous avons pu associer la réactivation du locus X00041_h (Xp22.3) a environ 50 % des échantillons, ce qui rejoint les observations faites par d'autres équipes (Alves et al. 2008; Wentzensen et al. 2007). Un fort déséquilibre en faveur des échantillons provenant de Shanghai a cependant été constaté, ce qui fait écho aux travaux d'une équipe chinoise qui étudie ce locus HERV depuis plusieurs années (Liang et al. 2009a; Liang et al. 2012; Liang et al. 2007) et peut suggérer une prédisposition populationnelle en lien, par exemple, avec des changements d'habitudes alimentaires (Chiu et al. 2003; You et al. 2002). A l'inverse, nous avons identifié puis étudié deux nouvelles séquences HERV, 2000045_h et 1400177_h, dont l'expression semble absente des échantillons de cancer de côlon en provenance de Shanghai alors que leur réactivation est associée à un tiers du reste du recrutement. En combinant ces trois loci HERV, la couverture de détection des cancers du côlon à partir de biopsies pourrait atteindre les 80 % et être complémentaire des différents tests FOB d'identification de sang dans les selles dont les valeurs de sensibilité restent en général faibles pour les cancers colorectaux et les adénomes avancés (entre 11 % et 64 % de sensibilité par exemple pour les tests gFOB) (van Dam et al. 2010).

III.2.1.2.2. Hétérogénéité tumorale

Comme un lien entre les cancers, il est frappant de constater que le locus 2000045_h, associé aux tumeurs du côlon dans notre étude du transcriptome sur un panel de tissus, est une des rares séquences HERV dont l'expression différentielle a été mise en évidence dans les urines de patients ayant un cancer de la prostate. Une explication peut venir du nombre limité d'échantillons de tissus utilisés pour représenter chaque organe, et ainsi rejoindre un problème de variabilité inter individuelle. On peut d'ailleurs rappeler qu'une détection positive de cette séquence a également été obtenue en RT-PCR sur un mélange d'ARN de tumeur de l'ovaire (article PLoS ONE, supplementary figure S2), ce qui semblerait indiquer que la réactivation de ce locus est liée à certains états cancéreux. Alternativement, la question de l'hétérogénéité tumorale a peut-être insuffisamment été prise en considération. Il a récemment été rappelé dans une étude de grande envergure que l'hétérogénéité d'une tumeur peut conduire à sous-estimer les paysages génomique et transcriptomique obtenus à partir d'une biopsie unique, ce qui représente un défi majeur pour le développement d'une médecine personnalisée basée sur l'utilisation de biomarqueurs (Gerlinger et

al. 2012). Il n'est donc pas exclu que la connaissance acquise à partir de tissus de prostate macro disséqués ne soit pas directement transférable à la fraction de cellules relarguée dans les urines. En particulier, les influences réciproques des cellules luminales, basales et stromales restent mal connues dans les mécanismes de différenciation cellulaire. Utiliser la microdissection pour établir une cartographie de l'expression des rétrovirus endogènes humains dans chacun de ces compartiments cellulaires devrait permettre de renforcer la connaissance biologique associée aux détections observées, comme cela a pu être réalisé pour le répertoire des gènes conventionnels (Oudes et al. 2006; Pascal et al. 2009a).

III.2.2. Quel positionnement pour les rétrovirus endogènes humains en cancérologie et avec quelles méthodes ?

III.2.2.1. Place des HERV dans la thématique du cancer de la prostate

Dans la course aux marqueurs moléculaires, le répertoire HERV que nous avons choisi d'exploiter présente sur le papier un certain nombre d'avantages. Il est d'abord mal maitrisé par une concurrence qui focalise principalement ses recherches sur les régions codantes de l'ADN. Rappelons sur ce point que le seul test diagnostique alternatif au PSA disponible en France détecte dans les urines un transcrit non codant, spécifique du cancer de la prostate, et dénommé PCA3 (Bussemakers et al. 1999; Vlaeminck-Guillem et al. 2010). Un nouvel ARN non codant qui régulerait spécifiquement la prolifération des cellules prostatiques vient d'ailleurs tout récemment d'être caractérisé et désigné sous le terme PCAT-1 (Prensner et al. 2011). Mais plus largement, les séquences LTRs des HERV peuvent participer aux phénomènes d'expression dans la prostate en lien avec des levées de méthylation (Goering et al. 2011), être associées à des transcrits de fusion issus de recombinaisons (Hermans et al. 2008; Tomlins et al. 2007a) ou encore, comme cela vient d'être démontré pour le PSA, moduler l'expression de gènes tissu-spécifiques *via* des balances hormonales (Lawrence et al. 2012). Etre capable de caractériser un nouveau transcrit HERV ou d'étendre la connaissance des mécanismes de régulation d'un gène cellulaire par l'effet d'un élément HERV sont, en outre, autant de façons qui permettent de sortir du périmètre de la propriété industrielle actuelle (Murphy and Watson 2012). La prostate n'est toutefois pas connue pour être un contexte privilégié d'expression des HERV, bien que, par exemple, une déméthylation globale d'éléments transposables de type LINE-1 ait pu être associée à l'évolution de la maladie (Florl et al. 2004; Netto et al. 2008). Nous avons pourtant identifié, à partir du travail sur prostatectomies de cette thèse, un profil d'expression impliquant un petit nombre de séquences HERV, et notamment de séquences HERV-E, ce qui corrobore en partie des observations antérieures (Bai et al. 2007; Molinaro et al. 2006; Pichon et al.

2006; Wang-Johanning et al. 2003a). Sur notre base expérimentale, il semble néanmoins clair que la réactivation des rétrovirus endogènes humains est bien plus limitée dans une prostate tumorale qu'elle ne l'est, par exemple, dans des cellules de la lignée germinale ou de côlon, ce qui renvoie une fois de plus à une sensibilité à des facteurs de transcription tissulaires acquise par certaines séquences HERV au fil de l'évolution. L'âge des familles, cependant, reste insuffisant pour prédire la spécialisation d'une LTR, comme le rappelle la fonction d'enhancer du locus du PSA assurée par une LTR de la famille HERV-L (Lawrence et al. 2012), qui est la plus ancienne famille du génome humain. De plus, le niveau général d'expression des HERV est globalement plus faible que celui des gènes codants. Ainsi, en abordant la problématique d'une détection de transcrits prostatiques dans les urines, nous avons été confrontés à une somme de risques et d'incertitudes qu'il a fallu pondérer par plusieurs niveaux de maitrise, et en particulier arriver à mesurer des variations de niveaux d'expression qui ne sont pas aussi tranchées que les cas de réactivations 'on-off' traditionnellement décrits. On peut alors légitimement se demander si la technologie des puces à ADN est à ce jour la plus adaptée à une telle problématique de détection.

III.2.2.2. Apports et limites des nouvelles technologies de séquençage (NGS)

Les technologies NGS sont actuellement utilisées pour faire le séquençage *de novo* ou le re-séquençage de génomes complets, pour étudier la diversité des génomes par des projets de métagénomique, mais aussi pour gagner un regard sur les contrôles épigénétiques à grande échelle, découvrir des ARN non-codants ainsi que de nouveaux sites de fixation de protéines et, par le RNA-Seq, établir des profils d'expression. Parmi les machines les plus commercialisées dans le monde, citons le séquenceur 454 de Roche, les technologies Solexa d'Illumina et les marques SOLiD issues de Life Technologies. Les coûts d'investissement (600 k€ pour le HiSeq Illumina) et de fonctionnement (20 k€ par run pour le HiSeq) restent souvent très élevés et les durées de séquençage généralement longues (8 jours pour un run sur la dernière génération SOLiD 5500 xi ; pour un comparatif complet voir (Glenn 2011)), ce qui fait que ces machines sont encore majoritairement présentes dans des grandes organisations académiques ou gouvernementales. Depuis 2007 cependant, la conception du Ion Torrent par Jonathan Rothberg (l'inventeur du 454) puis son rachat par Life Technologies offre un positionnement alternatif. De la taille d'un ordinateur de bureau, ce séquenceur surnommé PGM pour Personal Genome Machine réduit le temps de lecture à 2 heures en concédant 1 % de précision par rapport à ses concurrents, le tout conjugué à une capacité de séquençage (Mb/run) encore très inférieure mais en régulière progression (la version de décembre 2010 permet de lire 10 millions de bases). Cela se traduit par des coûts globaux divisés par dix et permet pour la première fois à tous les scientifiques, les hôpitaux ou les universités d'envisager un investissement. Un article paru en ligne sur le site GenomeWeb rappelle par exemple que l'utilisation du Ion Torrent a permis une

identification en trois jours de la souche E. Coli responsable de l'épidémie mortelle en Europe qui sévit au printemps 2011. Malgré tout, le Ion Torrent, s'il est tout à fait adapté pour le séquençage de génomes microbiens, ne l'est plus vraiment sur des génomes volumineux.

L'utilisation du RNA-Seq a fourni de nombreux résultats de recherches de marqueurs de cancers. En particulier, différentes méthodes d'identification de transcrits de fusion ont été développées pour exploiter avantageusement les données du séquençage à haut débit, parmi lesquelles TopHat-Fusion (Kim and Hahn 2011), Comrad (McPherson et al. 2011b), deFuse (McPherson et al. 2011a) nFuse (McPherson et al. 2012) ou d'autres encore (Kinsella et al. 2011; Pflueger et al. 2011; Ren et al. 2012). De plus, la faisabilité d'études transcriptomiques sur de très petites quantités d'acides nucléiques, allant jusqu'à établir le profil d'expression d'une unique cellule (Lao et al. 2009; Ramskold et al. 2012; Tang et al. 2009) a été maintes fois montrée. Pour ces raisons les NGS semblent pouvoir jouer un rôle prédominant dans la thématique de recherche de marqueurs du cancer de la prostate dans les urines, et finalement réussir à s'imposer dans le champ de la médecine personnalisée, comme cela été revendiqué dans le cadre d'une étude pilote clinique en oncologie (Roychowdhury et al. 2011). Pourtant, la limitation principale du RNA-Seq tient à la difficulté d'étudier les longues séquences répétées. Les éléments répétés en général, parmi lesquels les rétrovirus endogènes, se sont en effet toujours heurtés à des problèmes d'alignement et d'assemblage (Martin and Wang 2011), mais l'arrivée des NGS, qui reposent essentiellement sur des lectures de courtes tailles et impliquent d'importants volumes de données, ont accentué ces problèmes (Treangen et al. 2011). Du point de vue bioinformatique, les éléments répétés sont donc une source d'ambiguïtés qui produit des biais d'analyses conduisant à des interprétations erronées et, à l'heure actuelle, aucun algorithme, même sophistiqué (Treangen and Salzberg 2012), n'arrive à sauver de la poubelle ces pans entiers du génome humains.

III.2.2.3. Capitaliser sur un savoir-faire

Dans ce contexte, la maitrise du répertoire génomique des rétrovirus endogènes et d'un procédé de conception de sondes de 25 mer locus-spécifique est certainement un atout sur lequel assoir une stratégie compétitive. Durant ce travail, nous avons eu l'occasion d'évoquer l'insuffisance des annotations des bases de données publiques ainsi que les limites des approches basées sur les banques d'ESTs. Tirer toutes les conséquences de ces constats a amené le laboratoire à développer ses propres bases de connaissances et d'annotations, qui recouvrent aujourd'hui huit familles HERV sur la quarantaine de groupes définis. Parallèlement, le modèle initial de stabilité d'hybridation cible/sonde développé par Pozhitkov (Pozhitkov et al. 2006) a été appliqué expérimentalement au format Affymetrix et ses paramètres, affinés au cours du temps, fournissent un cadre méthodologique à la définition de sondes spécifiques. Si l'évolution du savoir-faire du laboratoire en

matière de définition de puces à ADN dédiées au répertoire HERV peut donc continuer à progresser dans ce sens, il se heurte néanmoins à la limite des génomes de référence utilisés comme source de séquences. Or une partie des séquences HERV-K HML-2 est absente de la population (Subramanian et al. 2011), et d'une manière plus générale le niveau de polymorphisme humain est estimé à 1 sur 0,31 kb pour les éléments répétés (contre 1 sur 1,8 kb à 2 kb pour les régions codantes) (Nickerson et al. 1998), ce qui indique que de nombreux SNPs doivent exister sur les séquences HERV d'un individu à l'autre. Abondant dans ce sens, une étude récente vient de comparer les variants génomiques identifiés par les projets 1 000 génomes (1000 Genomes Project Consortium 2010) et Complete Sequencing (Roach et al. 2010) et conclut sur l'impossibilité de prendre actuellement en compte l'ensemble des variations de séquences existant au sein de la population humaine (Rosenfeld et al. 2012). Il doit donc être possible d'imaginer à plus long terme une exploitation de la puce HERV comme un outil de capture, utilisé avec un niveau de spécificité maitrisé, et couplé en aval au séquençage à haut débit des cibles ainsi concentrées. Cette solution présenterait l'avantage de caractériser sans ambiguïté une fraction ciblée du génome humain tout en réussissant à observer les différents niveaux de variabilité nécessaires à la mise au point de tests de biologie moléculaire robustes.

III.3. Conclusion et perspectives

Les différents éléments de discussion qui viennent d'être évoqués suggèrent que les puces à ADN, et en particulier les puces HERV, peuvent garder une fenêtre d'applications encore quelques années. La question de leur champ d'utilisation est par conséquent posée. D'un point de vue méthodologique, il sera sans doute nécessaire d'arriver à couvrir l'ensemble du répertoire HERV en incluant toutes les familles du génome humain pour valider ou invalider les projections qui ont été proposées durant ce travail de thèse à partir de l'étude de 6 familles. Sur la base des observations faites à partir d'un ensemble de tissus, nous avons en particulier revendiqué qu'un tiers des séquences HERV sont actives, dont un nombre significatif le serait de manière constitutive. Ce résultat s'inscrit à l'heure actuelle largement en opposition de l'état des connaissances dont un des postulats est que l'activité des rétrovirus endogènes humains en contexte physiologique est réprimée par l'hôte, hors cas de domestication. Cette vision commence pourtant à être réévaluée avec les conclusions du projet ENCODE et plus largement par la prise de conscience de la place de l'ADN non codant dans les mécanismes de régulation de l'activité transcriptionnelle. La validité des observations que nous avons faites peut par ailleurs être défendue en faisant le constat que les expression constitutives concernent pour les trois quarts des séquences LTRs, et par conséquent ne rentrent pas dans le cadre des travaux menés par les équipes de référence qui basent principalement

leurs systèmes de détection sur les régions *pol* ou *env* des provirus. Il devient ainsi naturel de questionner la fonction de régulation des LTRs. Dans ce travail, 326 LTRs promotrices et 209 LTRs polyA ont été identifiées par la puce HERV-V2. Ceci a permis de montrer que la fonction d'une LTR ne change pas d'un tissu à l'autre, et a par conséquent fait faire un pas en avant à la définition d'une LTR endogène active en suggérant qu'une LTR autonome est un élément qui s'est fonctionnement spécialisé au cours du temps. A l'inverse, la collection de 672 LTRs silencieuses est une source de travail pour définir les caractéristiques d'une LTR inactive. Ainsi, un prolongement de ces résultats viendra, par exemple, de l'étude approfondie des sites de fixation de facteurs de transcriptions des LTRs et donc de leurs acquisitions (ou pertes) d'un point de vue évolutif.

Le but poursuivi est d'arriver à appréhender la contribution des HERV à la régulation de l'expression des gènes. En particulier, les séquences LTRs semblent très impliquées dans l'établissement de profils d'expression géniques durant les phases de développement, comme cela a été initialement montré chez la souris (Peaston et al. 2004). Les phénomènes de déméthylation qui ont lieu entre le moment de la fertilisation et le stade morula permettent de restaurer le caractère totipotent des cellules et offrent des modèles d'études excitants qui peuvent permettre de clarifier la contribution des rétrovirus endogènes au développement embryonnaire. Il a par exemple été récemment montré chez le poulet que l'expression d'une LTR est contrôlée par les voies métaboliques Gata et Nanog des cellules pluripotentes et contribuerait à initier la séparation des feuillets embryonnaires (Mey et al. 2012). Chez la souris, la déplétion du facteur de transcription Rex1, nécessaire au renouvellement des cellules ES, est associée à une variation de l'expression de plusieurs éléments ERV (Guallar et al. 2012). Plus précisément et chez la souris toujours, un travail vient d'isoler une population de cellules ES qui présente les caractéristiques transcriptomiques du stade embryonnaire à deux cellules, et pour laquelle des transcrits sont initiés par des séquences LTRs (Macfarlan et al. 2012). Les auteurs de ce dernier travail suggèrent, en conclusion, que la cooptation de séquences de rétrovirus endogènes pourrait assumer un rôle d'égale importance dans le développement embryonnaire précoce que celui joué durant la morphogenèse placentaire par les gènes (H)ERV domestiqués.

Alternativement à l'étude très réglementée des cellules ES humaines, il reste possible d'exploiter des modèles cellulaires qui répondent à une problématique ciblée. La théorie des cellules souches tumorales (CSC pour Cancer Stem Cell) offre par exemple un cadre conceptuel qui peut faire le lien entre biologie du développement et différenciation cellulaire. Bien que l'identification des CSC soit sujette à polémiques, comme cela peut être le cas pour le cancer de la prostate où deux études importantes publiées dans Science et Nature identifient respectivement une cellule basale (Goldstein et al. 2010) et luminale (Wang et al. 2009) comme origine de la tumeur, une caractérisation des propriétés transcriptomiques des cellules souches tumorales peut conduire à l'identification de marqueurs spécifiques, être informative des voies de signalisation du cancer et, *in fine*, fournir de

nouvelles pistes diagnostiques et thérapeutiques. Plus prosaïquement mais en lien avec les démarches d'accréditation de produits à usage pharmaceutique, le recours à des collections de cellules, comme le panel NCI-60 de cellules cancéreuses, permettrait d'amorcer un inventaire trans-pathologies de l'activité des rétrovirus endogènes. Un tel travail serait l'occasion d'intégrer plusieurs répertoires nucléiques (gènes cellulaires, HERV, petits ARN non codants) à une recherche transcriptomique intra- et extracellulaire, qui trouverait son pendant protéomique. Ce que nous enseignent finalement le mieux les rétrovirus endogènes humains est à quel point il peut être préjudiciable de se limiter à une fraction d'information au détriment de beaucoup d'autres. Aller vers des démarches globalisantes, dont le revers est inévitablement d'assumer la complexité de la biologie systémique, permettra de rester dans une course à la connaissance qui se durcit à chaque évolution technologique et s'autoalimente par la diffusion de plus en plus spontanée des informations scientifiques. A l'opposé, ce contexte de recherche doit réaffirmer la priorité des hypothèses biologiques sur les automatismes analytiques, pour permettre le confort d'une biologie de niche valorisable au long cours.

Abd-Elsalam KA (2003) Bioinformatic tools and guideline for PCR primer design. African Journal of Biotechnology 2: 91-95

Affymetrix (2003) Design and Performance of the GeneChip Human Genome U133 Plus 2.0 and Human Genome U133A 2.0 Arrays. Part No. 701483 Rev. 2

Ahn K, Kim HS (2009) Structural and quantitative expression analyses of HERV gene family in human tissues. Molecules and cells 28:99-103

Akopov SB, Nikolaev LG, Khil PP, Lebedev YB, Sverdlov ED (1998) Long terminal repeats of human endogenous retrovirus K family (HERV-K) specifically bind host cell nuclear proteins. FEBS Lett 421:229-233

Alves PM, Levy N, Stevenson BJ, Bouzourene H, Theiler G, Bricard G, Viatte S, Ayyoub M, Vuilleumier H, Givel JC, Rimoldi D, Speiser DE, Jongeneel CV, Romero PJ, Levy F (2008) Identification of tumor-associated antigens by large-scale analysis of genes expressed in human colorectal cancer. Cancer Immun 8:11

An DS, Xie Y, Chen IS (2001) Envelope gene of the human endogenous retrovirus HERV-W encodes a functional retrovirus envelope. J Virol 75:3488-9

Anderssen S, Sjottem E, Svineng G, Johansen T (1997) Comparative analyses of LTRs of the ERV-H family of primate-specific retrovirus-like elements isolated from marmoset, African green monkey, and man. Virology 234:14-30

Andersson AC, Merza M, Venables P, Ponten F, Sundstrom J, Cohen M, Larsson E (1996) Elevated levels of the endogenous retrovirus ERV3 in human sebaceous glands. J Invest Dermatol 106:125-128

Andersson AC, Venables PJ, Tonjes RR, Scherer J, Eriksson L, Larsson E (2002) Developmental expression of HERV-R (ERV3) and HERV-K in human tissue. Virology 297:220-5

Andersson AC, Yun Z, Sperber GO, Larsson E, Blomberg J (2005) ERV3 and related sequences in humans: structure and RNA expression. J Virol 79:9270-84

Andriole GL, Crawford ED, Grubb RL, III, Buys SS, Chia D, Church TR, Fouad MN, Isaacs C, Kvale PA, Reding DJ, Weissfeld JL, Yokochi LA, O'Brien B, Ragard LR, Clapp JD, Rathmell JM, Riley TL, Hsing AW, Izmirlian G, Pinsky PF, Kramer BS, Miller AB, Gohagan JK, Prorok PC (2012) Prostate cancer screening in the randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial: mortality results after 13 years of follow-up. Journal of the National Cancer Institute 104:125-132

Antony JM, Ellestad KK, Hammond R, Imaizumi K, Mallet F, Warren KG, Power C (2007) The human endogenous retrovirus envelope glycoprotein, syncytin-1, regulates neuroinflammation and its receptor expression in multiple sclerosis: a role for endoplasmic reticulum chaperones in astrocytes. J Immunol 179:1210-24

Antony JM, van Marle G, Opii W, Butterfield DA, Mallet F, Yong VW, Wallace JL, Deacon RM, Warren K, Power C (2004) Human endogenous retrovirus glycoprotein-mediated induction of redox reactants causes oligodendrocyte death and demyelination. Nat Neurosci 7:1088-95

Ariza ME, Williams MV (2011) A human endogenous retrovirus K dUTPase triggers a TH1, TH17 cytokine response: does it have a role in psoriasis? The Journal of investigative dermatology 131:2419-2427

Armbruester V, Sauter M, Krautkraemer E, Meese E, Kleiman A, Best B, Roemer K, Mueller-Lantzsch N (2002) A novel gene from the human endogenous retrovirus K expressed in transformed cells. Clin Cancer Res 8:1800-7

Armbruester V, Sauter M, Roemer K, Best B, Hahn S, Nty A, Schmid A, Philipp S, Mueller A, Mueller-Lantzsch N (2004) Np9 protein of human endogenous retrovirus K interacts with ligand of numb protein X. J Virol 78:10310-9

Arnaud F, Varela M, Spencer TE, Palmarini M (2008) Coevolution of endogenous betaretroviruses of sheep and their host. Cellular and molecular life sciences : CMLS 65:3422-3432

Arruda JT, Silva DM, Silva CC, Moura KK, da Cruz AD (2008) Homologous recombination between HERVs causes duplications in the AZFa region of men accidentally exposed to cesium-137 in Goiania. Genet Mol Res 7:1063-9

Azar GA, Thibodeau J (2002) Human endogenous retrovirus IDDMK(1,2)22 and mouse mammary tumor virus superantigens differ in their ability to stimulate murine T cell hybridomas. Immunol Lett 81:87-91

Baba K, Nakaya Y, Shojima T, Muroi Y, Kizaki K, Hashizume K, Imakawa K, Miyazawa T (2011) Identification of novel endogenous betaretroviruses which are transcribed in the bovine placenta. Journal of Virology 85:1237-1245

Baczyk D, Drewlo S, Proctor L, Dunk C, Lye S, Kingdom J (2009) Glial cell missing-1 transcription factor is required for the differentiation of the human trophoblast. Cell Death Differ 16:719-727

Badenhoop K, Donner H, Neumann J, Herwig J, Kurth R, Usadel KH, Tönjes RR (1999) IDDM patients neither show humoral reactivities against endogenous retroviral envelope protein nor do they differ in retroviral mRNA expression from healthy relatives or normal individuals. Diabetes 48:215-218

Bai VU, Kaseb A, Tejwani S, Divine GW, Barrack ER, Menon M, Pardee AB, Reddy GP (2007) Identification of prostate cancer mRNA markers by averaged differential expression and their detection in biopsies, blood, and urine. Proceedings of the National Academy of Sciences of the United States of America 104:2343-2348

Bainbridge SA, Minhas A, Whiteley KJ, Qu D, Sled JG, Kingdom JC, Adamson SL (2012) Effects of Reduced Gcm1 Expression on Trophoblast Morphology, Fetoplacental Vascularity, and Pregnancy Outcomes in Mice. Hypertension 59:732-739

Balaj L, Lessard R, Dai L, Cho YJ, Pomeroy SL, Breakefield XO, Skog J (2011) Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences. Nat Commun 2:180

Bannert N, Kurth R (2006) The evolutionary dynamics of human endogenous retroviral families. Annu Rev Genomics Hum Genet 7:149-73

Basyuk E, Cross JC, Corbin J, Nakayama H, Hunter P, Nait-Oumesmar B, Lazzarini RA (1999) Murine Gcm1 gene is expressed in a subset of placental trophoblast cells. Dev Dyn 214:303-311

Baust C, Seifarth W, Germaier H, Hehlmann R, Leib-Mosch C (2000) HERV-K-T47D-Related long terminal repeats mediate polyadenylation of cellular transcripts. Genomics 66:98-103

Bayani J, Selvarajah S, Maire G, Vukovic B, Al-Romaih K, Zielenska M, Squire JA (2007) Genomic mechanisms and measurement of structural and numerical instability in cancer cells. Seminars in cancer biology 17:5-18

Baylin SB, Jones PA (2011) A decade of exploring the cancer epigenome - biological and translational implications. Nature reviews Cancer 11:726-734

Bello D, Webber MM, Kleinman HK, Wartinger DD, Rhim JS (1997) Androgen responsive adult human prostatic epithelial cell lines immortalized by human papillomavirus 18. Carcinogenesis 18:1215-1223

Belshaw R, Dawson AL, Woolven-Allen J, Redding J, Burt A, Tristem M (2005) Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. J Virol 79:12507-12514

Belshaw R, Watson J, Katzourakis A, Howe A, Woolven-Allen J, Burt A, Tristem M (2007) Rate of recombinational deletion among human endogenous retroviruses. J Virol 81:9437-42

Bengtsson A, Blomberg J, Nived O, Pipkorn R, Toth L, Sturfelt G (1996) Selective antibody reactivity with peptides from human endogenous retroviruses and nonviral poly(amino acids) in patients with systemic lupus erythematosus. Arthritis Rheum 39:1654-63

Benit L, Dessen P, Heidmann T (2001) Identification, phylogeny, and evolution of retroviral elements based on their envelope genes. J Virol 75:11709-19

Bensalah K, Montorsi F, Shariat SF (2007) Challenges of cancer biomarker profiling. European urology 52:1601-1609

Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, Sharon E, Spector Y, Bentwich Z (2005) Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet 37:766-770

Bessis D, Moles JP, Basset-Seguin N, Tesniere A, Arpin C, Guilhou JJ (2004) Differential expression of a human endogenous retrovirus E transmembrane envelope glycoprotein in normal, psoriatic and atopic dermatitis human skin. Br J Dermatol 151:737-45

Beyer U, Moll-Rocek J, Moll UM, Dobbelstein M (2011) Endogenous retrovirus drives hitherto unknown proapoptotic p63 isoforms in the male germ line of humans and great apes. Proceedings of the National Academy of Sciences of the United States of America 108:3624-3629

Bi S, Gavrilova O, Gong DW, Mason MM, Reitman M (1997) Identification of a placental enhancer for the human leptin gene. J Biol Chem 272:30583-8

Bieche I, Laurent A, Laurendeau I, Duret L, Giovangrandi Y, Frendo JL, Olivi M, Fausser JL, Evain-Brion D, Vidaud M (2003) Placenta-specific INSL4 expression is mediated by a human endogenous retrovirus element. Biol Reprod 68:1422-9

Bieda K, Hoffmann A, Boller K (2001) Phenotypic heterogeneity of human endogenous retrovirus particles produced by teratocarcinoma cell lines. J Gen Virol 82:591-6

Biomarkers Definitions Working Group (2001) Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clinical pharmacology and therapeutics 69:89-95

Bjerregaard B, Holck S, Christensen IJ, Larsson LI (2006) Syncytin is involved in breast cancer-endothelial cell fusions. Cell Mol Life Sci 63:1906-1911

Bjerregaard B, Talts JF, Larsson LI (2011) The Endogenous Envelope Protein Syncytin Is Involved In Myoblast Fusion. In: Larsson LI (ed) Cell Fusions: Regulation and Control, Springer Netherlands, pp 267-275

Blaise S, de Parseval N, Benit L, Heidmann T (2003) Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. Proc Natl Acad Sci U S A 100:13013-8

Blaise S, de PN, Heidmann T (2005) Functional characterization of two newly identified Human Endogenous Retrovirus coding envelope genes. Retrovirology 2:19

Blanco P, Shlumukova M, Sargent CA, Jobling MA, Affara N, Hurles ME (2000) Divergent outcomes of intrachromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. J Med Genet 37:752-8

Blikstad V, Benachenhou F, Sperber GO, Blomberg J (2008) Evolution of human endogenous retroviral sequences: a conceptual account. Cell Mol Life Sci 65:3348-65

Blomberg J, Benachenhou F, Blikstad V, Sperber G, Mayer J (2009) Classification and nomenclature of endogenous retroviral sequences (ERVs): problems and recommendations. Gene 448:115-123

Blomberg J, Nived O, Pipkora R, Bengtsson A, Erlinge D, Sturfelt G (1994) Increased antiretroviral antibody reactivity in sera from a defined population of patients with sytemic lupus erythematosus. Arthr Rheum 37:57-66

Blond JL, Beseme F, Duret L, Bouton O, Bedin F, Perron H, Mandrand B, Mallet F (1999) Molecular characterization and placental expression of HERV-W, a new human endogenous retrovirus family. J Virol 73:1175-85

Blond JL, Lavillette D, Cheynet V, Bouton O, Oriol G, Chapel-Fernandes S, Mandrand B, Mallet F, Cosset FL (2000) An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. J Virol 74:3321-9

Boeke JD, Stoye JP (1997) Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In: Coffin JM, Hughes SH, Varmus HE (ed) Retroviruses, Cold Spring Harbor laboratory press, New-York, pp 343-435

Boese A, Sauter M, Galli U, Best B, Herbst H, Mayer J, Kremmer E, Roemer K, Mueller-Lantzsch N (2000) Human endogenous retrovirus protein cORF supports cell transformation and associates with the promyelocytic leukemia zinc finger protein. Oncogene 19:4328-36

Bogerd HP, Zhang F, Bieniasz PD, Cullen BR (2011) Human APOBEC3 proteins can inhibit xenotropic murine leukemia virus-related virus infectivity. Virology 410:234-239

Boller K, Frank H, Lower J, Lower R, Kurth R (1983) Structural organization of unique retrovirus-like particles budding from human teratocarcinoma cell lines. J Gen Virol 64 (Pt 12):2549-59

Boller K, Janssen O, Schuldes H, Tonjes RR, Kurth R (1997) Characterization of the antibody response specific for the human endogenous retrovirus HTDV/HERV-K. J Virol 71:4581-8

Boller K, Konig H, Sauter M, Mueller-Lantzsch N, Lower R, Lower J, Kurth R (1993) Evidence that HERV-K is the endogenous retrovirus sequence that codes for the human teratocarcinoma-derived retrovirus HTDV. Virology 196:349-53

Boller K, Schonfeld K, Lischer S, Fischer N, Hoffmann A, Kurth R, Tonjes RR (2008) Human endogenous retrovirus HERV-K113 is capable of producing intact viral particles. J Gen Virol 89:567-72

Bonnaud B, Beliaeff J, Bouton O, Oriol G, Duret L, Mallet F (2005) Natural history of the ERVWE1 endogenous retroviral locus. Retrovirology 2:57

Bonnaud B, Bouton O, Oriol G, Cheynet V, Duret L, Mallet F (2004) Evidence of selection on the domesticated ERVWE1 env retroviral element involved in placentation. Mol Biol Evol 21:1895-901

Brand A, Griffiths DJ, Herve C, Mallon E, Venables PJ (1999) Human retrovirus-5 in rheumatic disease. Journal of autoimmunity 13:149-154

Brodsky I, Foley B, Haines D, Johnston J, Cuddy K, Gillespie D (1993) Expression of HERV-K proviruses in human leukocytes. Blood 81:2369-74

Brosius J (2005) Echoes from the past-are we still in an RNP world? Cytogenetic and genome research 110:8-24

Brosius J, Gould SJ (1992) On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". Proceedings of the National Academy of Sciences of the United States of America 89:10706-10710

Burmeister T, Ebert AD, Pritze W, Loddenkemper C, Schwartz S, Thiel E (2004) Insertional polymorphisms of endogenous HERV-K113 and HERV-K115 retroviruses in breast cancer patients and age-matched controls. AIDS Res Hum Retroviruses 20:1223-9

Buscher K, Hahn S, Hofmann M, Trefzer U, Ozel M, Sterry W, Lower J, Lower R, Kurth R, Denner J (2006) Expression of the human endogenous retrovirus-K transmembrane envelope, Rec and Np9 proteins in melanomas and melanoma cell lines. Melanoma Res 16:223-34

Buscher K, Trefzer U, Hofmann M, Sterry W, Kurth R, Denner J (2005) Expression of human endogenous retrovirus K in melanomas and melanoma cell lines. Cancer Res 65:4172-80

Bussemakers MJ, van BA, Verhaegh GW, Smit FP, Karthaus HF, Schalken JA, Debruyne FM, Ru N, Isaacs WB (1999) DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. Cancer Research 59:5975-5979

Buzdin A, Kovalskaya-Alexandrova E, Gogvadze E, Sverdlov E (2006a) At least 50% of human-specific HERV-K (HML-2) long terminal repeats serve in vivo as active promoters for host nonrepetitive DNA transcription. J Virol 80:10752-62

Buzdin A, Kovalskaya-Alexandrova E, Gogvadze E, Sverdlov E (2006b) GREM, a technique for genome-wide isolation and quantitative analysis of promoter active repeats. Nucleic Acids Res 34:e67

Buzdin AA (2004) Retroelements and formation of chimeric retrogenes. Cellular and molecular life sciences : CMLS 61:2046-2059

Cameron HS, Szczepaniak D, Weston BW (1995) Expression of human chromosome 19p alpha(1,3)-fucosyltransferase genes in normal tissues. Alternative splicing, polyadenylation, and isoforms. The Journal of biological chemistry 270:20112-20122

Cassens S, Ulrich U, Beimling P, Simon D (1994) Inhibition of human T cell leukaemia virus type I long terminal repeat expression by DNA methylation: implications for latency. J Gen Virol 75 (Pt 11):3255-9

Cavanagh MH, Landry S, Audet B, Arpin-Andre C, Hivin P, Pare ME, Thete J, Wattel E, Marriott SJ, Mesnard JM, Barbeau B (2006) HTLV-I antisense transcripts initiating in the 3'LTR are alternatively spliced and polyadenylated. Retrovirology 3:15

Chaipan C, Dilley KA, Paprotka T, viks-Frankenberry KA, Venkatachari NJ, Hu WS, Pathak VK (2011) Severe restriction of xenotropic murine leukemia virus-related virus replication and spread in cultured human peripheral blood mononuclear cells. Journal of Virology 85:4888-4897

Chang CC, Lin PC, Lin CH, Yeh KT, Hung HY, Chang JG (2012) Rapid identification of CYP2C8 polymorphisms by high resolution melting analysis. Clinica chimica acta; international journal of clinical chemistry 413:298-302

Chang CW, Chang GD, Chen H (2011) A novel cyclic AMP/Epac1/CaMKI signaling cascade promotes GCM1 desumoylation and placental cell fusion. Mol Cell Biol 31:3820-3831

Chang CW, Chuang HC, Yu C, Yao TP, Chen H (2005) Stimulation of GCMa transcriptional activity by cyclic AMP/protein kinase A signaling is attributed to CBP-mediated acetylation of GCMa. Mol Cell Biol 25:8401-14

Chang WK, Yang KD, Shaio MF (1999) Effect of glutamine on Th1 and Th2 cytokine responses of human peripheral blood mononuclear cells. Clin Immunol 93:294-301

Charrel RN, de L, X (2010) Zoonotic aspects of arenavirus infections. Veterinary microbiology 140:213-220

Chen CP, Chen LF, Yang SR, Chen CY, Ko CC, Chang GD, Chen H (2008) Functional characterization of the human placental fusogenic membrane protein syncytin 2. Biol Reprod 79:815-23

Chen CP, Wang KG, Chen CY, Yu C, Chuang HC, Chen H (2006) Altered placental syncytin and its receptor ASCT2 expression in placental development and pre-eclampsia. Bjog 113:152-8

Chen HJ, Carr K, Jerome RE, Edenberg HJ (2002) A retroviral repetitive element confers tissue-specificity to the human alcohol dehydrogenase 1C (ADH1C) gene. DNA Cell Biol 21:793-801

Chen YX, Allars M, Maiti K, Angeli GL, bou-Seif C, Smith R, Nicholson RC (2011) Factors affecting cytotrophoblast cell viability and differentiation: Evidence of a link between syncytialisation and apoptosis. The international journal of biochemistry & cell biology 43:821-828

Cheng YH, Aronow BJ, Hossain S, Trapnell B, Kong S, Handwerger S (2004a) Critical role for transcription factor AP-2alpha in human trophoblast differentiation. Physiol Genomics 18:99-107

Cheng YH, Richardson BD, Hubert MA, Handwerger S (2004b) Isolation and characterization of the human syncytin gene promoter. Biol Reprod 70:694-701

Chew YC, West JT, Kratzer SJ, Ilvarsonn AM, Eissenberg JC, Dave BJ, Klinkebiel D, Christman JK, Zempleni J (2008) Biotinylation of histones represses transposable elements in human and mouse cells and cell lines and in Drosophila melanogaster. J Nutr 138:2316-22

Cheynet V, Oriol G, Mallet F (2006) Identification of the hASCT2-binding domain of the Env ERVWE1/syncytin-1 fusogenic glycoprotein. Retrovirology 3:41

Cheynet V, Ruggieri A, Oriol G, Blond JL, Boson B, Vachot L, Verrier B, Cosset FL, Mallet F (2005) Synthesis, assembly, and processing of the Env ERVWE1/syncytin human endogenous retroviral envelope. J Virol 79:5585-93

Chiu BC, Ji BT, Dai Q, Gridley G, McLaughlin JK, Gao YT, Fraumeni JF, Jr., Chow WH (2003) Dietary factors and risk of colon cancer in Shanghai, China. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 12:201-208

Christensen T (2005) Association of human endogenous retroviruses with multiple sclerosis and possible interactions with herpes viruses. Rev Med Virol 15:179-211

Christensen T (2010) HERVs in neuropathogenesis. Journal of neuroimmune pharmacology : the official journal of the Society on NeuroImmune Pharmacology 5:326-335

Christensen T, Sorensen PD, Hansen HJ, Moller-Larsen A (2003) Antibodies against a human endogenous retrovirus and the preponderance of env splice variants in multiple sclerosis patients. Mult Scler 9:6-15

Cianciolo GJ, Copeland TD, Oroszlan S, Snyderman R (1985) Inhibition of lymphocyte proliferation by a synthetic peptide homologous to retroviral envelope protein. Science 230:453-455

Coffin JM (1992) Genetic diversity and evolution of retroviruses. Curr Top Microbiol Immunol 176:143-164

Cohen CJ, Lock WM, Mager DL (2009) Endogenous retroviral LTRs as promoters for human genes: a critical assessment. Gene 448:105-114

Cohen CJ, Rebollo R, Babovic S, Dai EL, Robinson WP, Mager DL (2011) Placenta-specific expression of the interleukin-2 (IL-2) receptor beta subunit from an endogenous retroviral promoter. The Journal of biological chemistry 286:35543-35552

Cohen SX, Moulin M, Hashemolhosseini S, Kilian K, Wegner M, Muller CW (2003) Structure of the GCM domain-DNA complex: a DNA-binding domain with a novel fold and mode of target site recognition. Embo J 22:1835-1845

Conley A, Hinshelwood M (2001) Mammalian aromatases. Reproduction 121:685-95

Conley AB, Jordan IK (2010) Identification of transcription factor binding sites derived from transposable element sequences using ChIP-seq. Methods in molecular biology (Clifton, N J) 674:225-240

Conley AB, Piriyapongsa J, Jordan IK (2008) Retroviral promoters in the human genome. Bioinformatics 24:1563-7

Conrad B, Weidmann E, Trucco G, Rudert WA, Behboo R, Ricordi C, Rodriquez-Rilo H, Finegold D, Trucco M (1994) Evidence for superantigen involvement in insulin-dependent diabetes mellitus aetiology. Nature 371:351-355

Conrad B, Weissmahr RN, Boni J, Arcari R, Schupbach J, Mach B (1997) A human endogenous retroviral superantigen as candidate autoimmune gene in type I diabetes. Cell 90:303-13

Contreras-Galindo R, Kaplan MH, Contreras-Galindo AC, Gonzalez-Hernandez MJ, Ferlenghi I, Giusti F, Lorenzo E, Gitlin SD, Dosik MH, Yamamura Y, Markovitz DM (2012) Characterization of human endogenous retroviral elements in the blood of HIV-1-infected individuals. Journal of Virology 86:262-276

Contreras-Galindo R, Kaplan MH, Leissner P, Verjat T, Ferlenghi I, Bagnoli F, Giusti F, Dosik MH, Hayes DF, Gitlin SD, Markovitz DM (2008) Human endogenous retrovirus K (HML-2) elements in the plasma of people with lymphoma and breast cancer. J Virol 82:9329-36

Contreras-Galindo R, Kaplan MH, Markovitz DM, Lorenzo E, Yamamura Y (2006) Detection of HERV-K(HML-2) viral RNA in plasma of HIV type 1-infected individuals. AIDS Research and human retroviruses 22:979-984

Contreras-Galindo R, Lopez P, Velez R, Yamamura Y (2007) HIV-1 infection increases the expression of human endogenous retroviruses type K (HERV-K) in vitro. AIDS Research and human retroviruses 23:116-122

Crowell RC, Kiessling AA (2007) Endogenous retrovirus expression in testis and epididymis. Biochem Soc Trans 35:629-33

D'Arcy F, Foley R, Perry A, Marignol L, Lawler L, Gaffney E, Watson RGW, Fitzpatrick JM, Lynch TH (2008) No evidence of XMRV in Irish prostate cancer patients with the R462Q mutation. European Urology Supplements 7:271

de Parseval N, Alkabbani H, Heidmann T (1999) The long terminal repeats of the HERV-H human endogenous retrovirus contain binding sites for transcriptional regulation by the Myb protein. J Gen Virol 80 (Pt 4):841-5

de Parseval N, Lazar V, Casella JF, Benit L, Heidmann T (2003) Survey of human genes of retroviral origin: identification and transcriptome of the genes with coding capacity for complete envelope proteins. J Virol 77:10414-22

Denne M, Sauter M, Armbruester V, Licht JD, Roemer K, Mueller-Lantzsch N (2007) Physical and functional interactions of human endogenous retrovirus proteins Np9 and rec with the promyelocytic leukemia zinc finger protein. J Virol 81:5607-16

Depil S, Roche C, Dussart P, Prin L (2002) Expression of a human endogenous retrovirus, HERV-K, in the blood cells of leukemia patients. Leukemia 16:254-9

Dewannieux M, Harper F, Richaud A, Letzelter C, Ribet D, Pierron G, Heidmann T (2006) Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. Genome Res 16:1548-56

Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM (2001) Delineation of prognostic biomarkers in prostate cancer. Nature 412:822-826

Diamandis EP (2010) Cancer biomarkers: can we turn recent failures into success? Journal of the National Cancer Institute 102:1462-1467

Dickerson F, Rubalcaba E, Viscidi R, Yang S, Stallings C, Sullens A, Origoni A, Leister F, Yolken R (2008) Polymorphisms in human endogenous retrovirus K-18 and risk of type 2 diabetes in individuals with schizophrenia. Schizophr Res 104:121-6

Diem O, Schaffner M, Seifarth W, Leib-Mosch C (2012) Influence of antipsychotic drugs on human endogenous retrovirus (HERV) transcription in brain cells. PLoS ONE 7:e30054

Dolei A (2005) MSRV/HERV-W/syncytin and its linkage to multiple sclerosis: the usability and the hazard of a human endogenous retrovirus. J Neurovirol 11:232-5

Domansky AN, Kopantzev EP, Snezhkov EV, Lebedev YB, Leib-Mosch C, Sverdlov ED (2000) Solitary HERV-K LTRs possess bi-directional promoter activity and contain a negative regulatory element in the U5 region. FEBS Lett 472:191-5

Dong B, Kim S, Hong S, Das GJ, Malathi K, Klein EA, Ganem D, DeRisi JL, Chow SA, Silverman RH (2007) An infectious retrovirus susceptible to an IFN antiviral pathway from human prostate tumors. Proceedings of the National Academy of Sciences of the United States of America 104:1655-1660

Donner H, Tonjes RR, Bontrop RE, Kurth R, Usadel KH, Badenhoop K (1999) Intronic sequence motifs of HLA-DQB1 are shared between humans, apes and Old World monkeys, but a retroviral LTR element (DQLTR3) is human specific. Tissue Antigens 53:551-558

Drewlo S, Leyting S, Kokozidou M, Mallet F, Potgens AJ (2006) C-Terminal truncations of syncytin-1 (ERVWE1 envelope) that increase its fusogenicity. Biol Chem 387:1113-20

Dunk CE, Gellhaus A, Drewlo S, Baczyk D, Potgens AJ, Winterhager E, Kingdom JC, Lye SJ (2012) The Molecular Role of Connexin 43 in Human Trophoblast Cell Fusion. Biology of reproduction 86 115:1-10

Dunlap KA, Palmarini M, Varela M, Burghardt RC, Hayashi K, Farmer JL, Spencer TE (2006) Endogenous retroviruses regulate periimplantation placental growth and differentiation. Proc Natl Acad Sci U S A 103:14390-14395

Dunn CA, Medstrand P, Mager DL (2003) An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon. Proc Natl Acad Sci U S A 100:12841-6

Dupressoir A, Marceau G, Vernochet C, Benit L, Kanellopoulos C, Sapin V, Heidmann T (2005) Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc Natl Acad Sci U S A 102:725-30

Dupressoir A, Vernochet C, Harper F, Guegan J, Dessen P, Pierron G, Heidmann T (2011) A pair of co-opted retroviral envelope syncytin genes is required for formation of the two-layered murine placental syncytiotrophoblast. Proceedings of the National Academy of Sciences of the United States of America 108:E1164-E1173

Ehlhardt S, Seifert M, Schneider J, Ojak A, Zang KD, Mehraein Y (2006) Human endogenous retrovirus HERV-K(HML-2) Rec expression and transcriptional activities in normal and rheumatoid arthritis synovia. J Rheumatol 33:16-23

Ejtehadi HD, Freimanis GL, Ali HA, Bowman S, Alavi A, Axford J, Callaghan R, Nelson PN (2006) The potential role of human endogenous retrovirus K10 in the pathogenesis of rheumatoid arthritis: a preliminary study. Ann Rheum Dis 65:612-6

Ejthadi HD, Martin JH, Junying J, Roden DA, Lahiri M, Warren P, Murray PG, Nelson PN (2005) A novel multiplex RT-PCR system detects human endogenous retrovirus-K in breast cancer. Arch Virol 150:177-84

ENCODE Project (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799-816

Epstein JI, Allsbrook WC, Jr., Amin MB, Egevad LL (2005) The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. The American journal of surgical pathology 29:1228-1242

Erlander MG, Ma XJ, Kesty NC, Bao L, Salunga R, Schnabel CA (2011) Performance and clinical evaluation of the 92-gene real-time PCR assay for tumor classification. The Journal of molecular diagnostics : JMD 13:493-503

Esnault C, Priet S, Ribet D, Vernochet C, Bruls T, Lavialle C, Weissenbach J, Heidmann T (2008) A placenta-specific receptor for the fusogenic, endogenous retrovirus-derived, human syncytin-2. Proc Natl Acad Sci U S A 105:17532-7

Feuchter AE, Freeman JD, Mager DL (1992) Strategy for detecting cellular transcripts promoted by human endogenous long terminal repeats: identification of a novel gene (CDC4L) with homology to yeast CDC4. Genomics 13:1237-46

Fischer N, Hellwinkel O, Schulz C, Chun FK, Huland H, Aepfelbacher M, Schlomm T (2008) Prevalence of human gammaretrovirus XMRV in sporadic prostate cancer. Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology 43:277-283

Flockerzi A, Maydt J, Frank O, Ruggieri A, Maldener E, Seifarth W, Medstrand P, Lengauer T, Meyerhans A, Leib-Mosch C, Meese E, Mayer J (2007) Expression pattern analysis of transcribed HERV sequences is complicated by ex vivo recombination. Retrovirology 4:39

Flockerzi A, Ruggieri A, Frank O, Sauter M, Maldener E, Kopper B, Wullich B, Seifarth W, Muller-Lantzsch N, Leib-Mosch C, Meese E, Mayer J (2008) Expression patterns of transcribed human endogenous retrovirus HERV-K(HML-2) loci in human tissues and the need for a HERV Transcriptome Project. BMC Genomics 9:354

Florl AR, Lower R, Schmitz-Drager BJ, Schulz WA (1999) DNA methylation and expression of LINE-1 and HERV-K provirus sequences in urothelial and renal cell carcinomas. Br J Cancer 80:1312-21

Florl AR, Steinhoff C, Muller M, Seifert HH, Hader C, Engers R, Ackermann R, Schulz WA (2004) Coordinate hypermethylation at specific genes in prostate carcinoma precedes LINE-1 hypomethylation. Br J Cancer 91:985-994

Foerster J, Nolte I, Junge J, Bruinenberg M, Schweiger S, Spaar K, van der SG, Ehlert C, Mulder M, Kalscheuer V, Blumenthal-Barby E, Winter J, Seeman P, Stander M, Sterry W, Te MG (2005) Haplotype sharing analysis identifies a retroviral dUTPase as candidate susceptibility gene for psoriasis. J Invest Dermatol 124:99-102

Forsman A, Yun Z, Hu L, Uzhameckis D, Jern P, Blomberg J (2005) Development of broadly targeted human endogenous gammaretroviral pol-based real time PCRs Quantitation of RNA expression in human tissues. J Virol Methods 129:16-30

Frank O, Giehl M, Zheng C, Hehlmann R, Leib-Mosch C, Seifarth W (2005) Human endogenous retrovirus expression profiles in samples from brains of patients with schizophrenia and bipolar disorders. J Virol 79:10890-901

Frank O, Jones-Brando L, Leib-Mosch C, Yolken R, Seifarth W (2006) Altered transcriptional activity of human endogenous retroviruses in neuroepithelial cells after infection with Toxoplasma gondii. The Journal of infectious diseases 194:1447-1449

Frank O, Verbeke C, Schwarz N, Mayer J, Fabarius A, Hehlmann R, Leib-Mosch C, Seifarth W (2008) Variable transcriptional activity of endogenous retroviruses in human breast cancer. J Virol 82:1808-18

Freimanis G, Hooley P, Ejtehadi HD, Ali HA, Veitch A, Rylance PB, Alawi A, Axford J, Nevill A, Murray PG, Nelson PN (2010) A role for human endogenous retrovirus-K (HML-2) in rheumatoid arthritis: investigating mechanisms of pathogenesis. Clinical and experimental immunology 160:340-347

Frendo JL, Olivier D, Cheynet V, Blond JL, Bouton O, Vidaud M, Rabreau M, Evain-Brion D, Mallet F (2003) Direct involvement of HERV-W Env glycoprotein in human trophoblast cell fusion and differentiation. Mol Cell Biol 23:3566-3574

Fuchs NV, Kraft M, Tondera C, Hanschmann KM, Lower J, Lower R (2011) Expression of the human endogenous retrovirus (HERV) group HML-2/HERV-K does not depend on canonical promoter elements but is regulated by transcription factors Sp1 and Sp3. Journal of Virology 85:3436-3448

Furuta RA, Miyazawa T, Sugiyama T, Kuratsune H, Ikeda Y, Sato E, Misawa N, Nakatomi Y, Sakuma R, Yasui K, Yamaguti K, Hirayama F (2011) No association of xenotropic murine leukemia virus-related virus with prostate cancer or chronic fatigue syndrome in Japan. Retrovirology 8:20

Gabriel U, Steidler A, Trojan L, Michel MS, Seifarth W, Fabarius A (2010) Smoking increases transcription of human endogenous retroviruses in a newly established in vitro cell model and in normal urothelium. AIDS Research and human retroviruses 26:883-888

Galli UM, Sauter M, Lecher B, Maurer S, Herbst H, Roemer K, Mueller-Lantzsch N (2005) Human endogenous retrovirus rec interferes with germ cell development in mice and may cause carcinoma in situ, the predecessor lesion of germ cell tumors. Oncogene 24:3223-8

Ganopoulos I, Madesis P, Zambounis A, Tsaftaris A (2012) High-resolution melting analysis allowed fast and accurate closed-tube genotyping of Fusarium oxysporum formae speciales complex. FEMS microbiology letters 334:16-21

Gaudin P, Ijaz S, Tuke PW, Marcel F, Paraz A, Seigneurin JM, Mandrand B, Perron H, Garson JA (2000) Infrequency of detection of particle-associated MSRV/HERV-W RNA in the synovial fluid of patients with rheumatoid arthritis. Rheumatology 39:950-954

Gaudray G, Gachon F, Basbous J, Biard-Piechaczyk M, Devaux C, Mesnard JM (2002) The complementary strand of the human T-cell leukemia virus type 1 RNA genome encodes a bZIP transcription factor that down-regulates viral transcription. J Virol 76:12813-22

Gazon H, Lemasson I, Polakowski N, Cesaire R, Matsuoka M, Barbeau B, Mesnard JM, Peloponese JM, Jr. (2012) Human T-Cell Leukemia Virus Type 1 (HTLV-1) bZIP Factor Requires Cellular Transcription Factor JunD To Upregulate HTLV-1 Antisense Transcription from the 3' Long Terminal Repeat. Journal of Virology 86:9070-9078

Georgiou I, Noutsopoulos D, Dimitriadou E, Markopoulos G, Apergi A, Lazaros L, Vaxevanoglou T, Pantos K, Syrrou M, Tzavaras T (2009) Retrotransposon RNA expression and evidence for retrotransposition events in human oocytes. Hum Mol Genet 18:1221-8

Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, Varela I, Phillimore B, Begum S, McDonald NQ, Butler A, Jones D, Raine K, Latimer C, Santos CR, Nohadani M, Eklund AC, Spencer-Dene B, Clark G, Pickering L, Stamp G, Gore M, Szallasi Z, Downward J, Futreal PA, Swanton C (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. The New England journal of medicine 366:883-892

Gifford R, Tristem M (2003) The evolution, distribution and diversity of endogenous retroviruses. Virus Genes 26:291-315

Gifford RJ, Katzourakis A, Tristem M, Pybus OG, Winters M, Shafer RW (2008) A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution. Proc Natl Acad Sci U S A 105:20362-7

Gilbert C, Maxfield DG, Goodman SM, Feschotte C (2009) Parallel germline infiltration of a lentivirus in two Malagasy lemurs. PLoS genetics 5:e1000425

Gimenez J, Montgiraud C, Oriol G, Pichon JP, Ruel K, Tsatsaris V, Gerbaud P, Frendo JL, Evain-Brion D, Mallet F (2009) Comparative Methylation of ERVWE1/Syncytin-1 and Other Human Endogenous Retrovirus LTRs in Placenta Tissues. DNA Res 16:195-211

Gimenez J, Montgiraud C, Pichon JP, Bonnaud B, Arsac M, Ruel K, Bouton O, Mallet F (2010) Custom human endogenous retroviruses dedicated microarray identifies self-induced HERV-W family elements reactivated in testicular cancer upon methylation control. Nucleic Acids Res 38:2229-2246

Gleason DF (1966) Classification of prostatic carcinomas. Cancer chemotherapy reports Part 1 50:125-128

Glenn TC (2011) Field guide to next-generation DNA sequencers. Molecular ecology resources 11:759-769

Glinsky GV, Berezovska O, Glinskii AB (2005) Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. J Clin Invest 115:1503-1521

Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL (2004) Gene expression profiling predicts clinical outcome of prostate cancer. The Journal of clinical investigation 113:913-923

Goedert JJ, Sauter ME, Jacobson LP, Vessella RL, Hilgartner MW, Leitman SF, Fraser MC, Mueller-Lantzsch NG (1999) High prevalence of antibodies against HERV-K10 in patients with testicular cancer but not with AIDS. Cancer Epidemiol Biomarkers Prev 8:293-6

Goering W, Ribarska T, Schulz WA (2011) Selective changes of retroelement expression in human prostate cancer. Carcinogenesis 32:1484-1492

Gogvadze E, Stukacheva E, Buzdin A, Sverdlov E (2009) Human-specific modulation of transcriptional activity provided by endogenous retroviral insertions. J Virol 83:6098-6105

Golan M, Hizi A, Resau JH, Yaal-Hahoshen N, Reichman H, Keydar I, Tsarfaty I (2008) Human endogenous retrovirus (HERV-K) reverse transcriptase as a breast cancer prognostic marker. Neoplasia 10:521-33

Goldstein AS, Huang J, Guo C, Garraway IP, Witte ON (2010) Identification of a cell of origin for human prostate cancer. Science 329:568-571

Gong R, Huang L, Shi J, Luo K, Qiu G, Feng H, Tien P, Xiao G (2007) Syncytin-A mediates the formation of syncytiotrophoblast involved in mouse placental development. Cell Physiol Biochem 20:517-26

Gonzalez-Hernandez MJ, Swanson MD, Contreras-Galindo R, Cookinham S, King SR, Noel RJ, Jr., Kaplan MH, Markovitz DM (2012) Expression of Human Endogenous Retrovirus Type K (HML-2) Is Activated by the Tat Protein of HIV-1. Journal of Virology 86:7790-7805

Gorlov IP, Sircar K, Zhao H, Maity SN, Navone NM, Gorlova OY, Troncoso P, Pettaway CA, Byun JY, Logothetis CJ (2010) Prioritizing genes associated with prostate cancer development. BMC cancer 10:599

Gotzinger N, Sauter M, Roemer K, Mueller-Lantzsch N (1996) Regulation of human endogenous retrovirus-K Gag expression in teratocarcinoma cell lines and human tumours. J Gen Virol 77 (Pt 12):2983-90

Griffiths DJ, Cooke SP, Herve C, Rigby SP, Mallon E, Hajeer A, Lock M, Emery V, Taylor P, Pantelidis P, Bunker CB, du BR, Weiss RA, Venables PJ (1999) Detection of human retrovirus 5 in patients with arthritis and systemic lupus erythematosus. Arthritis and rheumatism 42:448-454

Groom HC, Yap MW, Galao RP, Neil SJ, Bishop KN (2010) Susceptibility of xenotropic murine leukemia virus-related virus (XMRV) to retroviral restriction factors. Proceedings of the National Academy of Sciences of the United States of America 107:5166-5171

Guallar D, Perez-Palacios R, Climent M, Martinez-Abadia I, Larraga A, Fernandez-Juan M, Vallejo C, Muniesa P, Schoorlemmer J (2012) Expression of endogenous retroviruses is negatively regulated by the pluripotency marker Rex1/Zfp42. Nucl Acids Res 40: 8993–9007

Gundry CN, Vandersteen JG, Reed GH, Pryor RJ, Chen J, Wittwer CT (2003) Amplicon melting analysis with labeled primers: a closed-tube method for differentiating homozygotes and heterozygotes. Clinical chemistry 49:396-406

Hahn S, Ugurel S, Hanschmann KM, Strobel H, Tondera C, Schadendorf D, Lower J, Lower R (2008) Serological response to human endogenous retrovirus K in melanoma patients correlates with survival probability. AIDS Research and human retroviruses 24:717-723

Hahn T, Barth S, Weiss U, Mosgoeller W, Desoye G (1998) Sustained hyperglycemia in vitro down-regulates the GLUT1 glucose transport system of cultured human term placental trophoblast: a mechanism to protect fetal development? FASEB J 12:1221-1231

Hao W, Serreze DV, McCulloch JL, Neifing DK, Palmer JP (1993) Insulin (auto)antibodies from human IDDM crossreact with retroviral antigen p73. Autoimmunity 6:787-798

Harbig J, Sprinkle R, Enkemann SA (2005) A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. Nucl Acids Res 33:e31

Hart CA, Bennett M (1999) Hantavirus infections: epidemiology and pathogenesis. Microbes and infection / Institut Pasteur 1:1229-1237

Hart H, McCormick JN, Marmion BP (1979) Viruses and lymphocytes in rheumatoid arthritis. II. Examination of lymphocytes and sera from patients with rheumatoid arthritis for evidence of retrovirus infection. Annals of the rheumatic diseases 38:514-525

Hassler MR, Egger G (2012) Epigenomics of cancer - emerging new concepts. Biochimie 94: 2219–2230

Hasuike S, Miura K, Miyoshi O, Miyamoto T, Niikawa N, Jinno Y, Ishikawa M (1999) Isolation and localization of an IDDMK1,2-22-related human endogenous retroviral gene, and identification of a CA repeat marker at its locus. Journal of human genetics 44:343-347

Haupt S, Tisdale M, Vincendeau M, Clements MA, Gauthier DT, Lance R, Semmes OJ, Turqueti-Neves A, Noessner E, Leib-Mosch C, Greenwood AD (2011) Human endogenous retrovirus transcription profiles of the kidney and kidney-derived cell lines. The Journal of general virology 92:2356-2366

Heidmann O, Vernochet C, Dupressoir A, Heidmann T (2009) Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new "syncytin" in a third order of mammals. Retrovirology 6:107

Heinz S, Krause SW, Gabrielli F, Wagner HM, Andreesen R, Rehli M (2002) Genomic organization of the human gene HEP27: alternative promoter usage in HepG2 cells and monocyte-derived dendritic cells. Genomics 79:608-615

Herbst H, Sauter M, Kuhler-Obbarius C, Loning T, Mueller-Lantzsch N (1998) Human endogenous retrovirus (HERV)-K transcripts in germ cell and trophoblastic tumours. APMIS 106:216-220

Herbst H, Sauter M, Mueller-Lantzsch N (1996) Expression of human endogenous retrovirus K elements in germ cell and trophoblastic tumors. Am J Pathol 149:1727-35

Hermans KG, van der Korput HA, van MR, van de Wijngaart DJ, Ziel-van der MA, Dits NF, Boormans JL, van der Kwast TH, van DH, Bangma CH, Korsten H, Kraaij R, Jenster G, Trapman J (2008) Truncated ETV1, fused to novel tissue-specific genes, and full-length ETV1 in prostate cancer. Cancer Research 68:7541-7549

Hohenadl C, Germaier H, Walchner M, Hagenhofer M, Herrmann M, Sturzl M, Kind P, Hehlmann R, Erfle V, Leib-Mosch C (1999) Transcriptional activation of endogenous retroviral sequences in human epidermal keratinocytes by UVB irradiation. J Invest Dermatol 113:587-94

Hohn O, Krause H, Barbarotto P, Niederstadt L, Beimforde N, Denner J, Miller K, Kurth R, Bannert N (2009) Lack of evidence for xenotropic murine leukemia virus-related virus(XMRV) in German prostate cancer patients. Retrovirology 6:92

Holder BS, Tower CL, Forbes K, Mulla MJ, Aplin JD, Abrahams VM (2012) Immune cell activation by trophoblast-derived microvesicles is mediated by syncytin 1. Immunology 136:84-191

Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome research 19:1419-1428

Howard G, Eiges R, Gaudet F, Jaenisch R, Eden A (2008) Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. Oncogene 27:404-408

Hsiao FC, Lin M, Tai A, Chen G, Huber BT (2006) Cutting edge: Epstein-Barr virus transactivates the HERV-K18 superantigen by docking to the human complement receptor 2 (CD21) on primary B cells. Journal of immunology (Baltimore, Md : 1950) 177:2056-2060

Hsiao FC, Tai AK, Deglon A, Sutkowski N, Longnecker R, Huber BT (2009) EBV LMP-2A employs a novel mechanism to transactivate the HERV-K18 superantigen through its ITAM. Virology 385:261-266

Hu L, Hornung D, Kurek R, Ostman H, Blomberg J, Bergqvist A (2006) Expression of human endogenous gammaretroviral sequences in endometriosis and ovarian cancer. AIDS Res Hum Retroviruses 22:551-7

Hu N, Qiu X, Luo Y, Yuan J, Li Y, Lei W, Zhang G, Zhou Y, Su Y, Lu Q (2008) Abnormal histone modification patterns in lupus CD4+ T cells. The Journal of rheumatology 35:804-810

Huang W, Li S, Hu Y, Yu H, Luo F, Zhang Q, Zhu F (2011) Implication of the env gene of the human endogenous retrovirus W family in the expression of BDNF and DRD3 and development of recent-onset schizophrenia. Schizophrenia bulletin 37:988-1000

Huda A, Tyagi E, Marino-Ramirez L, Bowen NJ, Jjingo D, Jordan IK (2011) Prediction of transposable element derived enhancers using chromatin modification profiles. PLoS ONE 6:e27513

Hue S, Gray ER, Gall A, Katzourakis A, Tan CP, Houldcroft CJ, McLaren S, Pillay D, Futreal A, Garson JA, Pybus OG, Kellam P, Towers GJ (2010) Disease-associated XMRV sequences are consistent with laboratory contamination. Retrovirology 7:111

Hughes JF, Coffin JM (2001) Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat Genet 29:487-9

Huh JW, Kim DS, Kang DW, Ha HS, Ahn K, Noh YN, Min DS, Chang KT, Kim HS (2008) Transcriptional regulation of GSDML gene by antisense-oriented HERV-H LTR element. Archives of virology 153:1201-1205

Huland H (2001) Radical prostatectomy: options and issues. European urology 39 Suppl 1:3-9

Hull S, Fan H (2006) Mutational analysis of the cytoplasmic tail of jaagsiekte sheep retrovirus envelope protein. J Virol 80:8069-8080

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931-945

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England) 4:249-264

Ishida T, Hamano A, Koiwa T, Watanabe T (2006) 5' long terminal repeat (LTR)-selective methylation of latently infected HIV-1 provirus that is demethylated by reactivation signals. Retrovirology 3:69

Ishida T, Obata Y, Ohara N, Matsushita H, Sato S, Uenaka A, Saika T, Miyamura T, Chayama K, Nakamura Y, Wada H, Yamashita T, Morishima T, Old LJ, Nakayama E (2008) Identification of the HERV-K gag antigen in prostate cancer by SEREX using autologous patient serum and its immunogenicity. Cancer immunity 8:15

Iwabuchi H, Kakihara T, Kobayashi T, Imai C, Tanaka A, Uchiyama M, Fukuda T (2004) A gene homologous to human endogenous retrovirus overexpressed in childhood acute lymphoblastic leukemia. Leuk Lymphoma 45:2303-6

Jaeckel E, Heringlake S, Berger D, Brabant G, Hunsmann G, Manns MP (1999) No evidence for association between IDDMK(1,2)22, a novel isolated retrovirus, and IDDM. Diabetes 48:209-214

Jeanmougin M, de RA, Marisa L, Paccard C, Nuel G, Guedj M (2010) Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. PLoS ONE 5:e12336

Jern P, Sperber GO, Blomberg J (2004) Definition and variation of human endogenous retrovirus H. Virology 327:93-110

Jia Z, Wang Y, Sawyers A, Yao H, Rahmatpanah F, Xia XQ, Xu Q, Pio R, Turan T, Koziol JA, Goodison S, Carpenter P, Wang-Rodriguez J, Simoneau A, Meyskens F, Sutton M, Lernhardt W, Beach T, Monforte J, McClelland M, Mercola D (2011) Diagnosis of prostate cancer using differentially expressed genes in stroma. Cancer research 71:2476-2487

Jjingo D, Huda A, Gundapuneni M, Marino-Ramirez L, Jordan IK (2011) Effect of the transposable element environment of human genes on gene length and expression. Genome biology and evolution 3:259-271

Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics (Oxford, England) 8:118-127

Johnston JB, Silva C, Holden J, Warren KG, Clark AW, Power C (2001) Monocyte activation and differentiation augment human endogenous retrovirus expression: implications for inflammatory brain diseases. Ann Neurol 50:434-42

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research 110:462-467

Kalter SS, Heberling RL, Helmke RJ, Panigel M, Smith GC, Kraemer DC, Hellman A, Fowler AK, Strickland JE (1975) A comparative study on the presence of C-type viral particles in placentas from primates and other animals. Bibliotheca haematologica391-401

Kamat A, Alcorn JL, Kunczt C, Mendelson CR (1998) Characterization of the regulatory regions of the human aromatase (P450arom) gene involved in placenta-specific expression. Mol Endocrinol 12:1764-1777

Kammerer U, Germeyer A, Stengel S, Kapp M, Denner J (2011) Human endogenous retrovirus K (HERV-K) is expressed in villous and extravillous cytotrophoblast cells of the human placenta. Journal of reproductive immunology 91:1-8

Kamp C, Hirschmann P, Voss H, Huellen K, Vogt PH (2000) Two long homologous retroviral sequence blocks in proximal Yq11 cause AZFa microdeletions as a result of intrachromosomal recombination events. Hum Mol Genet 9:2563-72

Kapitonov VV, Jurka J (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS biology 3:e181

Karimi MM, Goyal P, Maksakova IA, Bilenky M, Leung D, Tang JX, Shinkai Y, Mager DL, Jones S, Hirst M, Lorincz MC (2011) DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. Cell stem cell 8:676-687

Karlsson H, Bachmann S, Schroder J, McArthur J, Torrey EF, Yolken RH (2001) Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia. Proc Natl Acad Sci U S A 98:4634-9

Kato N, Pfeifer-Ohlsson S, Kato M, Larsson E, Rydnert J, Ohlsson R, Cohen M (1987) Tissue-specific expression of human provirus ERV3 mRNA in human placenta: two of the three ERV3 mRNAs contain human cellular sequences. J Virol 61:2182-91

Katoh I, Mirova A, Kurata S, Murakami Y, Horikawa K, Nakakuki N, Sakai T, Hashimoto K, Maruyama A, Yonaga T, Fukunishi N, Moriishi K, Hirai H (2011) Activation of the long terminal repeat of human endogenous retrovirus K by melanoma-specific transcription factor MITF-M. Neoplasia (New York, N Y) 13:1081-1092

Katsumata K, Ikeda H, Sato M, Ishizu A, Kawarada Y, Kato H, Wakisaka A, Koike T, Yoshiki T (1999) Cytokine regulation of env gene expression of human endogenous retrovirus-R in human vascular endothelial cells. Clin Immunol 93:75-80

Katzourakis A, Pereira V, Tristem M (2007) Effects of recombination rate on human endogenous retrovirus fixation and persistence. J Virol 81:10712-7

Kawano N, Harada Y, Yoshida K, Miyado M, Miyado K (2011) Role of CD9 in Sperm-Egg Fusion and Its General Role in Fusion Phenomena. In: Larsson LI (ed) Cell Fusions: Regulation and Control, Springer Netherlands, pp 171-184

Keryer G, Alsat E, Tasken K, Evain-Brion D (1998) Cyclic AMP-dependent protein kinases and human trophoblast cell differentiation in vitro. J Cell Sci 111 (Pt 7):995-1004

Kim DS, Hahn Y (2011) Identification of human-specific transcript variants induced by DNA insertions in the human genome. Bioinformatics (Oxford, England) 27:14-21

Kim FJ, Battini JL, Manel N, Sitbon M (2004) Emergence of vertebrate retroviruses and envelope capture. Virology 318:183-191

Kim S, Kim N, Dong B, Boren D, Lee SA, Das GJ, Gaughan C, Klein EA, Lee C, Silverman RH, Chow SA (2008) Integration site preference of xenotropic murine leukemia virus-related virus, a new human retrovirus associated with prostate cancer. Journal of Virology 82:9964-9977

Kinjo Y, Matsuura N, Yokota Y, Ohtsu S, Nomoto K, Komiya I, Sugimoto J, Jinno Y, Takasu N (2001) Identification of nonsynonymous polymorphisms in the superantigen-coding region of IDDMK1,2 22 and a pilot study on the association between IDDMK1,2 22 and type 1 diabetes. Journal of human genetics 46:712-716

Kinsella M, Harismendy O, Nakano M, Frazer KA, Bafna V (2011) Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. Bioinformatics (Oxford, England) 27:1068-1075

Kjellman C, Sjogren HO, Salford LG, Widegren B (1999) HERV-F (XA34) is a full-length human endogenous retrovirus expressed in placental and fetal tissues. Gene 239:99-107

Klase Z, Winograd R, Davis J, Carpio L, Hildreth R, Heydarian M, Fu S, McCaffrey T, Meiri E, Ayash-Rashkovsky M, Gilad S, Bentwich Z, Kashanchi F (2009) HIV-1 TAR miRNA protects against apoptosis by altering cellular gene expression. Retrovirology 6:18

Kleiman A, Senyuta N, Tryakin A, Sauter M, Karseladze A, Tjulandin S, Gurtsevitch V, Mueller-Lantzsch N (2004) HERV-K(HML-2) GAG/ENV antibodies as indicator for therapy effect in patients with germ cell tumors. Int J Cancer 110:459-61

Klein SL, Calisher CH (2007) Emergence and persistence of hantaviruses. Current topics in microbiology and immunology 315:217-252

Knerr I, Beinder E, Rascher W (2002) Syncytin, a novel human endogenous retroviral gene in human placenta: evidence for its dysregulation in preeclampsia and HELLP syndrome. Am J Obstet Gynecol 186:210-3

Knerr I, Schnare M, Hermann K, Kausler S, Lehner M, Vogler T, Rascher W, Meissner U (2007) Fusiogenic endogenous-retroviral syncytin-1 exerts anti-apoptotic functions in staurosporine-challenged CHO cells. Apoptosis 12:37-43

Knerr I, Schubert SW, Wich C, Amann K, Aigner T, Vogler T, Jung R, Dotsch J, Rascher W, Hashemolhosseini S (2005) Stimulation of GCMa and syncytin via cAMP mediated PKA signaling in human trophoblastic cells under normoxic and hypoxic conditions. FEBS Lett 579:3991-8

Knerr I, Weigel C, Linnemann K, Dotsch J, Meissner U, Fusch C, Rascher W (2003) Transcriptional effects of hypoxia on fusiogenic syncytin and its receptor ASCT2 in human cytotrophoblast BeWo cells and in ex vivo perfused placental cotyledons. Am J Obstet Gynecol 189:583-8

Knossl M, Lower R, Lower J (1999) Expression of the human endogenous retrovirus HTDV/HERV-K is enhanced by cellular transcription factor YY1. J Virol 73:1254-61

Knudsen S (2005) Cluster Analysis. In A Biologist's Guide to Analysis of DNA Microarray Data, John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/0471227587

Kobayashi-Ishihara M, Yamagishi M, Hara T, Matsuda Y, Takahashi R, Miyake A, Nakano K, Yamochi T, Ishida T, Watanabe T (2012) HIV-1-encoded antisense RNA suppresses viral replication for a prolonged period. Retrovirology 9:38

Koide T, Salem-Izacc SM, Gomes SL, Vencio RZ (2006) SpotWhatR: a user-friendly microarray data analysis system. Genetics and molecular research : GMR 5:93-107

Koiwa T, Hamano-Usami A, Ishida T, Okayama A, Yamaguchi K, Kamihira S, Watanabe T (2002) 5'-long terminal repeat-selective CpG methylation of latent human T-cell leukemia virus type 1 provirus in vitro and in vivo. J Virol 76:9389-97

Koshi K, Ushizawa K, Kizaki K, Takahashi T, Hashizume K (2011) Expression of endogenous retrovirus-like transcripts in bovine trophoblastic cells. Placenta 32:493-499

Krach K, Badenhoop K, Tonjes RR (2003) The IDDM-associated solitary retroviral promoters DQ-LTR3 and DQ-LTR13 have a distinct impact on the expression of selected DQB1 genes in different cell lines in vitro. Immunogenetics 55:521-529

Krieg AM, Khan AS, Steinberg AD (1989) Expression of an endogenous retroviral transcript is associated with murine lupus. Arthritis and rheumatism 32:322-329

Krzysztalowska-Wawrzyniak M, Ostanek M, Clark J, Binczak-Kuleta A, Ostanek L, Kaczmarczyk M, Loniewska B, Wyrwicz LS, Brzosko M, Ciechanowicz A (2011) The distribution of human endogenous retrovirus K-113 in health and autoimmune diseases in Poland. Rheumatology (Oxford, England) 50:1310-1314

Kudo Y, Boyd CA, Sargent IL, Redman CW (2001) Tryptophan degradation by human placental indoleamine 2,3-dioxygenase regulates lymphocyte proliferation. J Physiol 535:207-215

Kumanogoh A, Kikutani H (2004) Biological functions and signaling of a transmembrane semaphorin, CD100/Sema4D. Cellular and molecular life sciences : CMLS 61:292-300

Kurn N, Chen P, Heath JD, Kopf-Sill A, Stephens KM, Wang S (2005) Novel isothermal, linear nucleic acid amplification systems for highly multiplexed applications. Clinical chemistry 51:1973-1981

La Mantia G, Majello B, Di Cristofano A, Strazzullo M, Minchiotti G, Lania L (1992) Identification of regulatory elements within the minimal promoter region of the human endogenous ERV9 proviruses: accurate transcription initiation is controlled by an Inr-like element. Nucleic Acids Res 20:4129-36

Laderoute MP, Giulivi A, Larocque L, Bellfoy D, Hou Y, Wu HX, Fowke K, Wu J, az-Mitoma F (2007) The replicative activity of human endogenous retrovirus K102 (HERV-K102) with HIV viremia. AIDS (London, England) 21:2417-2424

Lamprecht B, Bonifer C, Mathas S (2010) Repeat-element driven activation of proto-oncogenes in human malignancies. Cell cycle (Georgetown, Tex) 9:4276-4281

Landry JR, Rouhi A, Medstrand P, Mager DL (2002) The Opitz syndrome gene Mid1 is transcribed from a human endogenous retroviral promoter. Mol Biol Evol 19:1934-42

Langat DK, Johnson PM, Rote NS, Wango EO, Owiti GO, Isahakia MA, Mwenda JM (1999) Characterization of antigens expressed in normal baboon trophoblast and cross-reactive with HIV/SIV antibodies. J Reprod Immunol 42:41-58

Lao KQ, Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Tuch B, Bodeau J, Siddiqui A, Surani MA (2009) mRNA-sequencing whole transcriptome analysis of a single cell on the SOLiD system. Journal of biomolecular techniques : JBT 20:266-271

Lapointe J, Li C, Giacomini CP, Salari K, Huang S, Wang P, Ferrari M, Hernandez-Boussard T, Brooks JD, Pollack JR (2007) Genomic profiling reveals alternative genetic pathways of prostate tumorigenesis. Cancer Res 67:8504-8510

Lapointe J, Li C, Higgins JP, van de RM, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, Ekman P, DeMarzo AM, Tibshirani R, Botstein D, Brown PO, Brooks JD, Pollack JR (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proceedings of the National Academy of Sciences of the United States of America 101:811-816

Larsen JM, Christensen IJ, Nielsen HJ, Hansen U, Bjerregaard B, Talts JF, Larsson LI (2009) Syncytin immunoreactivity in colorectal cancer: potential prognostic impact. Cancer Lett 280:44-9

Larsson E, Andersson AC, Nilsson BO (1994) Expression of an endogenous retrovirus (ERV3 HERV-R) in human reproductive and embryonic tissues--evidence for a function for envelope gene products. Ups J Med Sci 99:113-20

Larsson E, Venables P, Andersson AC, Fan W, Rigby S, Botling J, Oberg F, Cohen M, Nilsson K (1997) Tissue and differentiation specific expression on the endogenous retrovirus ERV3 (HERV-R) in normal human tissues and during induced monocytic differentiation in the U-937 cell line. Leukemia 11 Suppl 3:142-144

Larsson LI, Holck S, Christensen IJ (2007) Prognostic role of syncytin expression in breast cancer. Hum Pathol 38:726-31

Laufer G, Mayer J, Mueller BF, Mueller-Lantzsch N, Ruprecht K (2009) Analysis of transcribed human endogenous retrovirus W env loci clarifies the origin of multiple sclerosis-associated retrovirus env sequences. Retrovirology 6:37

Lavie L, Kitova M, Maldener E, Meese E, Mayer J (2005) CpG methylation directly regulates transcriptional activity of the human endogenous retrovirus family HERV-K(HML-2). J Virol 79:876-83

Lavie L, Medstrand P, Schempp W, Meese E, Mayer J (2004) Human endogenous retrovirus family HERV-K(HML-5): status, evolution, and reconstruction of an ancient betaretrovirus in the human genome. J Virol 78:8788-8798

Lavillette D, Marin M, Ruggieri A, Mallet F, Cosset FL, Kabat D (2002) The envelope glycoprotein of human endogenous retrovirus type W uses a divergent family of amino acid transporters/cell surface receptors. J Virol 76:6442-52

Lawrence MG, Stephens CR, Need EF, Lai J, Buchanan G, Clements JA (2012) Long Terminal Repeats Act as Androgen-Responsive Enhancers for the PSA-Kallikrein Locus. Endocrinology 153:3199-3210

Lee WJ, Kwun HJ, Jang KL (2003) Analysis of transcriptional regulatory sequences in the human endogenous retrovirus W long terminal repeat. J Gen Virol 84:2229-35

Lee X, Keith JC, Jr., Stumm N, Moutsatsos I, McCoy JM, Crum CP, Genest D, Chin D, Ehrenfels C, Pijnenborg R, Van Assche FA, Mi S (2001) Downregulation of placental syncytin expression and abnormal protein localization in pre-eclampsia. Placenta 22:808-12

Lee YN, Bieniasz PD (2007) Reconstitution of an infectious human endogenous retrovirus. PLoS Pathog 3:e10

Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS genetics 3:1724-1735

Lefebvre G, Desfarges S, Uyttebroeck F, Munoz M, Beerenwinkel N, Rougemont J, Telenti A, Ciuffi A (2011) Analysis of HIV-1 expression level and sense of transcription by high-throughput sequencing of the infected cell. Journal of Virology 85:6205-6211

Lemmo AV, Rose DJ, Tisone TC (1998) Inkjet dispensing technology: applications in drug discovery. Current Opinion in Biotechnology 9:615-617

Li J, Akagi K, Hu Y, Trivett AL, Hlynialuk CJ, Swing DA, Volfovsky N, Morgan TC, Golubeva Y, Stephens RM, Smith DE, Symer DE (2012) Mouse endogenous retroviruses can trigger premature transcriptional termination at a distance. Genome research 22:870-884

Liang CY, Wang LJ, Chen CP, Chen LF, Chen YH, Chen H (2010) GCM1 regulation of the expression of syncytin 2 and its cognate receptor MFSD2A in human placenta. Biology of reproduction 83:387-395

Liang Q, Ding J, Xu R, Xu Z, Zheng S (2009a) Identification of a novel human endogenous retrovirus and promoter activity of its 5' U3. Biochem Biophys Res Commun 382:468-72

Liang Q, Ding J, Zheng S (2009b) Identification and detection of a novel human endogenous retrovirus-related gene, and structural characterization of its related elements. Genetics and molecular biology 32:704-708

Liang Q, Xu Z, Xu R, Wu L, Zheng S (2012) Expression Patterns of Non-Coding Spliced Transcripts from Human Endogenous Retrovirus HERV-H Elements in Colon Cancer. PLoS ONE 7:e29950

Liang QY, Xu ZF, Xu RZ, Zheng S, Ding JY (2007) [Deletion of the env region in HERV-H-X gene and its expression in colon cancer]. Ai zheng = Aizheng = Chinese journal of cancer 26:952-956

Lin HK, Hu YC, Yang L, Altuwaijri S, Chen YT, Kang HY, Chang C (2003) Suppression versus induction of androgen receptor functions by the phosphatidylinositol 3-kinase/Akt pathway in prostate cancer LNCaP cells with different passage numbers. The Journal of biological chemistry 278:50902-50907

Lindeskog M, Blomberg J (1997) Spliced human endogenous retroviral HERV-H env transcripts in T-cell leukemia cell lines and normal leukocytes: alternative splicing pattern of HERV-H transcripts. J Gen Virol 78:2575-2585

Ling J, Pi W, Bollag R, Zeng S, Keskintepe M, Saliman H, Krantz S, Whitney B, Tuan D (2002) The solitary long terminal repeats of ERV-9 endogenous retrovirus are conserved during primate evolution and possess enhancer activities in embryonic and hematopoietic cells. J Virol 76:2410-23

Lipka K, Tebbe B, Finckh U, Rolfs A (1996) Absence of human T-lymphotrophic virus type I in patients with systemic lupus erythematosus. Clinical and experimental dermatology 21:38-42

Liu M, Eiden MV (2011) Role of human endogenous retroviral long terminal repeats (LTRs) in maintaining the integrity of the human germ line. Viruses 3:901-905

Lombardi VC, Ruscetti FW, Das GJ, Pfost MA, Hagen KS, Peterson DL, Ruscetti SK, Bagni RK, Petrow-Sadowski C, Gold B, Dean M, Silverman RH, Mikovits JA (2009) Detection of an infectious retrovirus, XMRV, in blood cells of patients with chronic fatigue syndrome. Science (New York, N Y) 326:585-589

Long QM, Bengra C, Li CH, Kutlar F, Tuan D (1998) A long terminal repeat of the human endogenous retrovirus ERV-9 is located in the 5' boundary area of the human □-globin locus control region. Genomics 54:542-555

Löwer R (1999) The pathogenic potential of endogenous retroviruses: facts and fantasies. Trends Microbiol 7 (9):350-356

Lower R, Lower J, Kurth R (1996) The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. Proc Natl Acad Sci U S A 93:5177-84

Lower R, Lower J, Tondera-Koch C, Kurth R (1993) A general method for the identification of transcribed retrovirus sequences (R-U5 PCR) reveals the expression of the human endogenous retrovirus loci HERV-H and HERV-K in teratocarcinoma cells. Virology 192:501-11

Lower R, Tonjes RR, Korbmacher C, Kurth R, Lower J (1995) Identification of a Rev-related protein by analysis of spliced transcripts of the human endogenous retroviruses HTDV/HERV-K. J Virol 69:141-9

Ludwig JA, Weinstein JN (2005) Biomarkers in cancer staging, prognosis and treatment selection. Nature reviews Cancer 5:845-856

Lyden TW, Johnson PM, Mwenda JM, Rote NS (1994) Ultrastructural characterization of endogenous retroviral particles isolated from normal human placentas. Biol Reprod 51:152-7

Ma H, Ma Y, Ma W, Williams DK, Galvin TA, Khan AS (2011) Chemical induction of endogenous retrovirus particles from the vero cell line of African green monkeys. Journal of Virology 85:6579-6588

Ma XJ, Patel R, Wang X, Salunga R, Murage J, Desai R, Tuggle JT, Wang W, Chu S, Stecker K, Raja R, Robin H, Moore M, Baunoch D, Sgroi D, Erlander M (2006) Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. Archives of pathology & laboratory medicine 130:465-473

Macaulay EC, Weeks RJ, Andrews S, Morison IM (2011) Hypomethylation of functional retrotransposon-derived genes in the human placenta. Mammalian genome : official journal of the International Mammalian Genome Society 22:722-735

Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer O, Trono D, Pfaff SL (2012) Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature 487:57-63

Macfarlane C, Simmonds P (2004) Allelic variation of HERV-K(HML-2) endogenous retroviral elements in human populations. J Mol Evol 59:642-656

Mack M, Bender K, Schneider PM (2004) Detection of retroviral antisense transcripts and promoter activity of the HERV-K(C4) insertion in the MHC class III region. Immunogenetics 56:321-332

Mager DL (1989) Polyadenylation function and sequence variability of the long terminal repeats of the human endogenous retrovirus-like family RTVL-H. Virology 173:591-9

Mager DL, Hunter DG, Schertzer M, Freeman JD (1999) Endogenous retroviruses provide the primary polyadenylation signal for two new human genes (HHLA2 and HHLA3). Genomics 59:255-63

Mager DL, Medstrand P (2003) Retroviral repeat sequences. Nature Encyclopedia of the Human Genome, vol 5. Nature Publishing Group, pp 57-63

Magin C, Lower R, Lower J (1999) cORF and RcRE, the Rev/Rex and RRE/RxRE homologues of the human endogenous retrovirus family HTDV/HERV-K. J Virol 73:9496-507

Magistrelli C, Samoilova E, Agarwal RK, Banki K, Ferrante P, Vladutiu A, Phillips PE, Perl A (1999) Polymorphic genotypes of the *HRES-1* human endogenous retrovirus locus correlate with systemic lupus erythematosus and autoreactivity. Immunogenetics 49:829-834

Makawita S, Diamandis EP (2010) The bottleneck in the cancer biomarker pipeline and protein quantification through mass spectrometry-based approaches: current strategies for candidate verification. Clinical chemistry 56:212-222

Maksakova IA, Goyal P, Bullwinkel J, Brown JP, Bilenky M, Mager DL, Singh PB, Lorincz MC (2011) H3K9me3-binding proteins are dispensable for SETDB1/H3K9me3-dependent retroviral silencing. Epigenetics & chromatin 4:12

Maksakova IA, Mager DL, Reiss D (2008) Endogenous retroviruses : Keeping active endogenous retroviral-like elements in check: the epigenetic perspective. Cell Mol Life Sci 65:3329-47

Malassine A, Handschuh K, Tsatsaris V, Gerbaud P, Cheynet V, Oriol G, Mallet F, Evain-Brion D (2005) Expression of HERV-W Env glycoprotein (syncytin) in the extravillous trophoblast of first trimester human placenta. Placenta 26:556-62

Malik HS, Henikoff S, Eickbush TH (2000) Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 10:1307-18

Mallet F, Bouton O, Prudhomme S, Cheynet V, Oriol G, Bonnaud B, Lucotte G, Duret L, Mandrand B (2004) The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. Proc Natl Acad Sci U S A 101:1731-6

Mallet F, Prudhomme S (2004) [Retroviral inheritance in man]. J Soc Biol 198:399-412

Mamedov I, Lebedev Y, Hunsmann G, Khusnutdinova E, Sverdlov E (2004) A rare event of insertion polymorphism of a HERV-K LTR in the human genome. Genomics 84:596-599

Mameli G, Astone V, Khalili K, Serra C, Sawaya BE, Dolei A (2007) Regulation of the syncytin-1 promoter in human astrocytes by multiple sclerosis-related cytokines. Virology 362:120-30

Mangeney M, de Parseval N, Thomas G, Heidmann T (2001) The full-length envelope of an HERV-H human endogenous retrovirus has immunosuppressive properties. J Gen Virol 82:2515-2518

Mangeney M, Renard M, Schlecht-Louf G, Bouallaga I, Heidmann O, Letzelter C, Richaud A, Ducos B, Heidmann T (2007) Placental syncytins: Genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins. Proc Natl Acad Sci U S A 104:20534-9

Marguerat S, Wang WY, Todd JA, Conrad B (2004) Association of human endogenous retrovirus K-18 polymorphisms with type 1 diabetes. Diabetes 53:852-854

Marin M, Tailor CS, Nouri A, Kozak SL, Kabat D (1999) Polymorphisms of the cell surface receptor control mouse susceptibilities to xenotropic and polytropic leukemia viruses. Journal of Virology 73:9362-9368

Martens JH, O'Sullivan RJ, Braunschweig U, Opravil S, Radolf M, Steinlein P, Jenuwein T (2005) The profile of repeat-associated histone lysine methylation states in the mouse epigenome. Embo J 24:800-12

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nature reviews Genetics 12:671-682

Matouskova M, Blazkova J, Pajer P, Pavlicek A, Hejnar J (2006) CpG methylation suppresses transcriptional activity of human syncytin-1 in non-placental tissues. Exp Cell Res 312:1011-20

Matsui T, Leung D, Miyashita H, Maksakova IA, Miyachi H, Kimura H, Tachibana M, Lorincz MC, Shinkai Y (2010) Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. Nature 464:927-931

Mayer J, Blomberg J, Seal RL (2011) A revised nomenclature for transcribed human endogenous retroviral loci. Mobile DNA 2:7

Mayer J, Meese E (2005) Human endogenous retroviruses in the primate lineage and their influence on host genomes. Cytogenet Genome Res 110:448-456

McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG, Griffith M, Heravi MA, Senz J, Melnyk N, Pacheco M, Marra MA, Hirst M, Nielsen TO, Sahinalp SC, Huntsman D, Shah SP (2011a) deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS computational biology 7:e1001138

McPherson A, Wu C, Hajirasouliha I, Hormozdiari F, Hach F, Lapuk A, Volik S, Shah S, Collins C, Sahinalp SC (2011b) Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data. Bioinformatics (Oxford, England) 27:1481-1488

McPherson AW, Wu C, Wyatt A, Shah SP, Collins C, Sahinalp SC (2012) nFuse: Discovery of complex genomic rearrangements in cancer using high-throughput sequencing. Genome research 22: 2250-2261

Medstrand P, Landry JR, Mager DL (2001) Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. J Biol Chem 276:1896-903

Mellor AL, Munn DH (1999) Tryptophan catabolism and T-cell tolerance: immunosuppression by starvation? Immunology today 20:469-473

Menendez L, Benigno BB, McDonald JF (2004) L1 and HERV-W retrotransposons are hypomethylated in human ovarian carcinomas. Mol Cancer 3:12

Mey A, Acloque H, Lerat E, Gounel S, Tribollet V, Blanc S, Curton D, Birot AM, Nieto MA, Samarut J (2012) The endogenous retrovirus ENS-1 provides active binding sites for transcription factors in embryonic stem cells that specify extra embryonic tissue. Retrovirology 9:21

Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, Keith JC, Jr., McCoy JM (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403:785-9

Michael NL, Vahey MT, d'Arcy L, Ehrenberg PK, Mosca JD, Rappaport J, Redfield RR (1994) Negative-strand RNA transcripts are produced in human immunodeficiency virus type 1-infected cells and patients by a novel promoter downregulated by Tat. J Virol 68:979-87

Miyazawa T, Shojima T, Yoshikawa R, Ohata T (2011) Isolation of koala retroviruses from koalas in Japan. The Journal of veterinary medical science / the Japanese Society of Veterinary Science 73:65-70

Moles JP, Hadi JC, Guilhou JJ (2003) High prevalence of an IgG response against murine leukemia virus (MLV) in patients with psoriasis. Virus Res 94:97-101

Moles JP, Tesniere A, Guilhou JJ (2007) Reverse transcriptase activity in human normal and psoriatic skin samples. The British journal of dermatology 157:482-486

Moles JP, Tesniere A, Guilhou JJ (2005) A new endogenous retroviral sequence is expressed in skin of patients with psoriasis. Br J Dermatol 153:83-9

Molinaro RJ, Jha BK, Malathi K, Varambally S, Chinnaiyan AM, Silverman RH (2006) Selection and cloning of poly(rC)-binding protein 2 and Raf kinase inhibitor protein RNA activators of 2',5'-oligoadenylate synthetase from prostate cancer cells. Nucl Acids Res 34:6684-6695

Mortensen K, Christensen IJ, Nielsen HJ, Hansen U, Larsson LI (2004) High expression of endothelial cell nitric oxide synthase in peritumoral microvessels predicts increased disease-free survival in colorectal cancer. Cancer Lett 216:109-114

Mottet N, Bellmunt J, Bolla M, Joniau S, Mason M, Matveev V, Schmid HP, van der KT, Wiegel T, Zattoni F, Heidenreich A (2011) [EAU guidelines on prostate cancer. Part II: treatment of advanced, relapsing, and castration-resistant prostate cancer]. Actas urologicas espanolas 35:565-579

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-62

Moyes DL, Martin A, Sawcer S, Temperton N, Worthington J, Griffiths DJ, Venables PJ (2005) The distribution of the endogenous retroviruses HERV-K113 and HERV-K115 in health and disease. Genomics 86:337-341

Muir A, Ruan QG, Marron MP, She JX (1999) The IDDMK$_{1,2}$22 retrovirus is not detectable in either mRNA or genomic DNA from patients with type 1 diabetes. Diabetes 48:219-222

Mullins CS, Linnebacher M (2011) Endogenous retrovirus sequences as a novel class of tumor-specific antigens: an example of HERV-H env encoding strong CTL epitopes. Cancer immunology, immunotherapy 61: 1093-1100

Mullins CS, Linnebacher M (2012) Endogenous retrovirus sequences as a novel class of tumor-specific antigens: an example of HERV-H env encoding strong CTL epitopes. Cancer immunology, immunotherapy : CII 61:1093-1100

Mullis K (1990) L'invention insolite de l'amplification de gènes. Pour La Science 152:44-53

Munoz-Suano A, Hamilton AB, Betz AG (2011) Gimme shelter: the immune system during pregnancy. Immunol Rev 241:20-38

Muradrasoli S, Forsman A, Hu L, Blikstad V, Blomberg J (2006) Development of real-time PCRs for detection and quantitation of human MMTV-like (HML) sequences HML expression in human tissues. Journal of virological methods 136:83-92

Murphy L, Watson RW (2012) Patented prostate cancer biomarkers. Nature reviews Urology 9:464-472

Muster T, Waltenberger A, Grassauer A, Hirschl S, Caucig P, Romirer I, Fodinger D, Seppele H, Schanab O, Magin-Lachmann C, Lower R, Jansen B, Pehamberger H, Wolff K (2003) An endogenous retrovirus derived from human melanoma cells. Cancer Res 63:8735-41

Naito T, Ogasawara H, Kaneko H, Hishikawa T, Sekigawa I, Hashimoto H, Maruyama N (2003) Immune abnormalities induced by human endogenous retroviral peptides: with reference to the pathogenesis of systemic lupus erythematosus. Journal of clinical immunology 23:371-376

Nakagawa K, Brusic V, McColl G, Harrison LC (1997) Direct evidence for the expression of multiple endogenous retroviruses in the synovial compartment in rheumatoid arthritis. Arthritis Rheum 40:627-38

Nakagawa T, Kollmeyer TM, Morlan BW, Anderson SK, Bergstralh EJ, Davis BJ, Asmann YW, Klee GG, Ballman KV, Jenkins RB (2008) A tissue biomarker panel predicting systemic progression after PSA recurrence post-definitive prostate cancer therapy. PloS one 3:e2318

Nellaker C, Yao Y, Jones-Brando L, Mallet F, Yolken RH, Karlsson H (2006) Transactivation of elements in the human endogenous retrovirus W family by viral infection. Retrovirology 3:44

Nelson DT, Goodchild NL, Mager DL (1996) Gain of Sp1 sites and loss of repressor sequences associated with a young, transcriptionally active subset of HERV-H endogenous long terminal repeats. Virology 220:213-8

Netto GJ, Nakai Y, Nakayama M, Jadallah S, Toubaji A, Nonomura N, Albadine R, Hicks JL, Epstein JI, Yegnasubramanian S, Nelson WG, De Marzo AM (2008) Global DNA hypomethylation in intratubular germ cell neoplasia and seminoma, but not in nonseminomatous male germ cell tumors. Mod Pathol 21:1337-44

Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat Genet 19:233-240

Nilsson BO, Jin M, Andersson AC, Sundstrom P, Larsson E (1999) Expression of envelope proteins of endogeneous C-type retrovirus on the surface of mouse and human oocytes at fertilization. Virus Genes 18:115-20

Nissen PH, Christensen SE, Ladefoged SA, Brixen K, Heickendorff L, Mosekilde L (2012) Identification of rare and frequent variants of the CASR gene by high-resolution melting. Clinica chimica acta; international journal of clinical chemistry 413:605-611

Noerholm M, Balaj L, Limperg T, Salehi A, Zhu LD, Hochberg FH, Breakefield XO, Carter BS, Skog J (2012) RNA expression patterns in serum microvesicles from patients with glioblastoma multiforme and controls. BMC Cancer 12:22

Nyegaard M, Demontis D, Thestrup BB, Hedemand A, Sorensen KM, Hansen T, Werge T, Hougaard DM, Yolken RH, Mortensen PB, Mors O, Borglum AD (2012) No association of polymorphisms in human endogenous retrovirus K18 and CD48 with schizophrenia. Psychiatric genetics 22:146-148

Ogasawara H, Hishikawa T, Sekigawa I, Hashimoto H, Yamamoto N, Maruyama N (2000) Sequence analysis of human endogenous retrovirus clone 4-1 in systemic lupus erythematosus. Autoimmunity 33:15-21

Ogasawara H, Naito T, Kaneko H, Hishikawa T, Sekigawa I, Hashimoto H, Kaneko Y, Yamamoto N, Maruyama N, Yamamoto N (2001) Quantitative analyses of messenger RNA of human endogenous retrovirus in patients with systemic lupus erythematosus. The Journal of rheumatology 28:533-538

Ogasawara H, Okada M, Kaneko H, Hishikawa T, Sekigawa I, Hashimoto H (2003) Possible role of DNA hypomethylation in the induction of SLE: relationship to the transcription of human endogenous retroviruses. Clin Exp Rheumatol 21:733-8

Oja M, Peltonen J, Blomberg J, Kaski S (2007) Methods for estimating human endogenous retrovirus activities from EST databases. BMC Bioinformatics 8 Suppl 2:S11

Ono M, Kawakami M, Ushikubo H (1987) Stimulation of expression of the human endogenous retrovirus genome by female steroid hormones in human breast cancer cell line T47D. J Virol 61:2059-62

Oppelt P, Strick R, Strissel PL, Winzierl K, Beckmann MW, Renner SP (2009) Expression of the human endogenous retroviruse-W envelope gene syncytin in endometriosis lesions. Gynecol Endocrinol 25:741-747

Otowa T, Tochigi M, Rogers M, Umekage T, Kato N, Sasaki T (2006) Insertional polymorphism of endogenous retrovirus HERV-K115 in schizophrenia. Neuroscience letters 408:226-229

Oudes AJ, Campbell DS, Sorensen CM, Walashek LS, True LD, Liu AY (2006) Transcriptomes of human prostate cells. BMC genomics 7:92

Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. The New England journal of medicine 351:2817-2826

Palmarini M, Gray CA, Carpenter K, Fan H, Bazer FW, Spencer TE (2001) Expression of endogenous betaretroviruses in the ovine uterus: effects of neonatal age, estrous cycle, pregnancy, and progesterone. J Virol 75:11319-11327

Pani MA, Wood JP, Bieda K, Toenjes RR, Usadel KH, Badenhoop K (2002) The variable endogenous retroviral insertion in the human complement C4 gene: a transmission study in type I diabetes mellitus. Human immunology 63:481-484

Paprotka T, Venkatachari NJ, Chaipan C, Burdick R, viks-Frankenberry KA, Hu WS, Pathak VK (2010) Inhibition of xenotropic murine leukemia virus-related virus by APOBEC3 proteins and antiviral drugs. Journal of Virology 84:5719-5729

Pascal LE, Goo YA, Vencio RZ, Page LS, Chambers AA, Liebeskind ES, Takayama TK, True LD, Liu AY (2009a) Gene expression down-regulation in CD90+ prostate tumor-associated stromal cells involves potential organ-specific genes. BMC cancer 9:317

Pascal LE, Vencio RZ, Page LS, Liebeskind ES, Shadle CP, Troisch P, Marzolf B, True LD, Hood LE, Liu AY (2009b) Gene expression relationship between prostate cancer cells of Gleason 3, 4 and normal epithelial cells as revealed by cell type-specific transcriptomes. BMC Cancer 9:452

Patience C, Simpson GR, Colletta AA, Welch HM, Weiss RA, Boyd MT (1996) Human endogenous retrovirus expression and reverse transcriptase activity in the T47D mammary carcinoma cell line. J Virol 70:2654-7

Patzke S, Lindeskog M, Munthe E, Aasheim HC (2002) Characterization of a novel human endogenous retrovirus, HERV-H/F, expressed in human leukemia cell lines. Virology 303:164-73

Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proceedings of the National Academy of Sciences of the United States of America 91:5022-5026

Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB (2004) Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Dev Cell 7:597-606

Pedersen FS, Sørensen AB (2010) Pathogenesis of Oncoviral Infections. In: Kurth R, Bannert N (ed) Retroviruses: Molecular Biology, Genomics and Pathogenesis, Caister Academic Press, pp 237-267

Perl A, Fernandez D, Telarico T, Phillips PE (2010) Endogenous retroviral pathogenesis in lupus. Current Opin Rheumatol 22:483-492

Pérot P, Montgiraud C, Lavillette D, Mallet F (2011) A Comparative Portrait of Retroviral Fusogens and Syncytins. In: Larsson LI (ed) Cell Fusions: Regulation and Control, Springer Netherlands, pp 63-115

Perron H, Geny C, Gratacap B, Laurent A, Lalande B, Perret J, Pellat J, Seigneurin JM (1991) Isolations of an unknown retrovirus from CSF, blood and brain cells of patients with multiple sclerosis. Current concepts in multiple sclerosis111-116

Perron H, Geny C, Laurent A, Mouriquand C, Pellat J, Perret J, Seigneurin JM (1989) Leptomeningeal cell line from multiple sclerosis with reverse transcriptase activity and viral particles. Res Virol 140:551-61

Perron H, Germi R, Bernard C, Garcia-Montojo M, Deluen C, Farinelli L, Faucard R, Veas F, Stefas I, Fabriek BO, Van-Horssen J, Van-der-Valk P, Gerdil C, Mancuso R, Saresella M, Clerici M, Marcel S, Creange A, Cavaretta R, Caputo D, Arru G, Morand P, Lang AB, Sotgiu S, Ruprecht K, Rieckmann P, Villoslada P, Chofflon M, Boucraut J, Pelletier J, Hartung HP (2012) Human endogenous retrovirus type W envelope expression in blood and brain cells provides new insights into multiple sclerosis disease. Multiple sclerosis 0:1-16

Perron H, Jouvin-Marche E, Michel M, Ounanian-Paraz A, Camelo S, Dumon A, Jolivet-Reynaud C, Marcel F, Souillet Y, Borel E, Gebuhrer L, Santoro L, Marcel S, Seigneurin JM, Marche PN, Lafon M (2001) Multiple sclerosis retrovirus particles and recombinant envelope trigger an abnormal immune response in vitro, by inducing polyclonal Vbeta16 T-lymphocyte activation. Virology 287:321-32

Perron H, Mekaoui L, Bernard C, Veas F, Stefas I, Leboyer M (2008) Endogenous retrovirus type W GAG and envelope protein antigenemia in serum of schizophrenic patients. Biol Psychiatry 64:1019-23

Perron H, Suh M, Lalande B, Gratacap B, Laurent A, Stoebner P, Seigneurin JM (1993) Herpes simplex virus ICP0 and ICP4 immediate early proteins strongly enhance expression of a retrovirus harboured by a leptomeningeal cell line from a patient with multiple sclerosis. J Gen Virol 74 (Pt 1):65-72

Pestana ES, Tenev T, Gross S, Stoyanov B, Ogata M, Böhmer FD (1999) The transmembrane protein tyrosine phosphatase RPTPsigma modulates signaling of the epidermal growth factor receptor in A431 cells. Oncogene 18:4069-4079

Petersen T, Moller-Larsen A, Thiel S, Brudek T, Hansen TK, Christensen T (2009) Effects of interferon-beta therapy on innate and adaptive immune responses to the human endogenous retroviruses HERV-H and HERV-W, cytokine production, and the lectin complement activation pathway in multiple sclerosis. Journal of neuroimmunology 215:108-116

Pflueger D, Terry S, Sboner A, Habegger L, Esgueva R, Lin PC, Svensson MA, Kitabayashi N, Moss BJ, MacDonald TY, Cao X, Barrette T, Tewari AK, Chee MS, Chinnaiyan AM, Rickman DS, Demichelis F, Gerstein MB, Rubin MA (2011) Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome research 21:56-67

Pi W, Yang Z, Wang J, Ruan L, Yu X, Ling J, Krantz S, Isales C, Conway SJ, Lin S, Tuan D (2004) The LTR enhancer of ERV-9 human endogenous retrovirus is active in oocytes and progenitor cells in transgenic zebrafish and humans. Proc Natl Acad Sci U S A 101:805-10

Pichon JP, Bonnaud B, Cleuziat P, Mallet F (2006) Multiplex degenerate PCR coupled with an oligo sorbent array for human endogenous retrovirus expression profiling. Nucleic Acids Res 34:e46

Ponferrada VG, Mauck BS, Wooley DP (2003) The envelope glycoprotein of human endogenous retrovirus HERV-W induces cellular resistance to spleen necrosis virus. Arch Virol 148:659-675

Poste G (2011) Bring on the biomarkers. Nature 469:156-157

Pothlichet J, Heidmann T, Mangeney M (2006) A recombinant endogenous retrovirus amplified in a mouse neuroblastoma is involved in tumor growth in vivo. International journal of cancer Journal international du cancer 119:815-822

Pozhitkov A, Noble PA, Domazet-Loso T, Nolte AW, Sonnenberg R, Staehler P, Beier M, Tautz D (2006) Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted. Nucl Acids Res 34:e66

Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, Cao X, Jing X, Wang X, Siddiqui J, Wei JT, Robinson D, Iyer HK, Palanisamy N, Maher CA, Chinnaiyan AM (2011) Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nature Biotechnology 29:742-749

Prudhomme S, Oriol G, Mallet F (2004) A retroviral promoter and a cellular enhancer define a bipartite element which controls env ERVWE1 placental expression. J Virol 78:12157-68

Prusty BK, zur HH, Schmidt R, Kimmel R, de Villiers EM (2008) Transcription of HERV-E and HERV-E-related sequences in malignant and non-malignant human haematopoietic cells. Virology 382:37-45

Ptolemy AS, Rifai N (2010) What is a biomarker? Research investments and lack of clinical integration necessitate a review of biomarker terminology and validation schema. Scandinavian journal of clinical and laboratory investigation Supplementum 242:6-14

Pullmann R, Jr., Bonilla E, Phillips PE, Middleton FA, Perl A (2008) Haplotypes of the HRES-1 endogenous retrovirus are associated with development and disease manifestations of systemic lupus erythematosus. Arthritis and rheumatism 58:532-540

Ramos-Lopez E, Ghebru S, Van Autreve J, Aminkeng F, Herwig J, Seifried E, Seidl C, Van der Auwera B, Badenhoop K (2006) Neither an intronic CA repeat within the CD48 gene nor the HERV-K18 polymorphisms are associated with type 1 diabetes. Tissue Antigens 68:147-52

Ramskold D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, Schroth GP, Sandberg R (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nature Biotechnology 30:777-782

Rasko JE, Battini JL, Gottschalk RJ, Mazo I, Miller AD (1999) The RD114/simian type D retrovirus receptor is a neutral amino acid transporter. Proc Natl Acad Sci U S A 96:2129-2134

Rasmussen MH, Ballarin-Gonzalez B, Liu J, Lassen LB, Fuchtbauer A, Fuchtbauer EM, Nielsen AL, Pedersen FS (2010) Antisense transcription in gammaretroviruses as a mechanism of insertional activation of host genes. Journal of Virology 84:3780-3788

Rawn SM, Cross JC (2008) The evolution, regulation, and function of placenta-specific genes. Annu Rev Cell Dev Biol 24:159-81

Reiche J, Pauli G, Ellerbrok H (2010) Differential expression of human endogenous retrovirus K transcripts in primary human melanocytes and melanoma cell lines after UV irradiation. Melanoma research 20:435-440

Ren S, Peng Z, Mao JH, Yu Y, Yin C, Gao X, Cui Z, Zhang J, Yi K, Xu W, Chen C, Wang F, Guo X, Lu J, Yang J, Wei M, Tian Z, Guan Y, Tang L, Xu C, Wang L, Gao X, Tian W, Wang J, Yang H, Wang J, Sun Y (2012) RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. Cell research 22:806-821

Ren YZ, Dai QD, Xu RZ (2005) Expression of a novel human retroviral NP9 gene and potential roles of its protein in systemic lupus erythematosus patients. Zhonghua yi xue yi chuan xue za zhi = Zhonghua yixue yichuanxue zazhi = Chinese journal of medical genetics 22:248-250

Renaudineau Y, Youinou P (2011) Epigenetics and autoimmunity, with special emphasis on methylation. The Keio journal of medicine 60:10-16

Reymond N, Charles H, Duret L, Calevro F, Beslon G, Fayard JM (2004) ROSO: optimizing oligonucleotide probes for microarrays. Bioinformatics (Oxford, England) 20:271-273

Reynier F, Verjat T, Turrel F, Imbert PE, Marotte H, Mougin B, Miossec P (2009) Increase in human endogenous retrovirus HERV-K (HML-2) viral load in active rheumatoid arthritis. Scandinavian journal of immunology 70:295-299

Rhea JM, Molinaro RJ (2011) Cancer biomarkers: surviving the journey from bench to bedside. MLO: medical laboratory observer 43:10-2, 16, 18

Richardson B, Scheinbart L, Strahler J, Gross L, Hanash S, Johnson M (1990) Evidence for impaired T cell DNA methylation in systemic lupus erythematosus and rheumatoid arthritis. Arthritis and rheumatism 33:1665-1673

Rifai N, Gillette MA, Carr SA (2006) Protein biomarker discovery and validation: the long and uncertain path to clinical utility. Nature Biotechnology 24:971-983

Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science (New York, N Y) 328:636-639

Rodriguez J, Vives L, Jorda M, Morales C, Munoz M, Vendrell E, Peinado MA (2008) Genome-wide tracking of unmethylated DNA Alu repeats in normal and cancer cells. Nucl Acids Res 36:770-784

Roebke C, Wahl S, Laufer G, Stadelmann C, Sauter M, Mueller-Lantzsch N, Mayer J, Ruprecht K (2010) An N-terminally truncated envelope protein encoded by a human endogenous retrovirus W locus on chromosome Xq22.3. Retrovirology 7:69

Rolland A, Jouvin-Marche E, Viret C, Faure M, Perron H, Marche PN (2006) The envelope protein of a human endogenous retrovirus-W family activates innate immunity through CD14/TLR4 and promotes Th1-like responses. J Immunol 176:7636-44

Rosenfeld JA, Mason CE, Smith TM (2012) Limitations of the human reference genome for personalized genomics. PLoS ONE 7:e40294

Rowe HM, Jakobsson J, Mesnard D, Rougemont J, Reynard S, Aktas T, Maillard PV, Layard-Liesching H, Verp S, Marquis J, Spitz F, Constam DB, Trono D (2010) KAP1 controls endogenous retroviruses in embryonic stem cells. Nature 463:237-240

Roychowdhury S, Iyer MK, Robinson DR, Lonigro RJ, Wu YM, Cao X, Kalyana-Sundaram S, Sam L, Balbin OA, Quist MJ, Barrette T, Everett J, Siddiqui J, Kunju LP, Navone N, Araujo JC, Troncoso P, Logothetis CJ, Innis JW, Smith DC, Lao CD, Kim SY, Roberts JS, Gruber SB, Pienta KJ, Talpaz M, Chinnaiyan AM (2011) Personalized oncology through integrative high-throughput sequencing: a pilot study. Science translational medicine 3:111ra121

Ruebner M, Strissel PL, Langbein M, Fahlbusch F, Wachter DL, Faschingbauer F, Beckmann MW, Strick R (2010) Impaired cell fusion and differentiation in placentae from patients with intrauterine growth restriction correlate with reduced levels of HERV envelope genes. Journal of molecular medicine (Berlin, Germany) 88:1143-1156

Ruprecht K, Gronen F, Sauter M, Best B, Rieckmann P, Mueller-Lantzsch N (2008) Lack of immune responses against multiple sclerosis-associated retrovirus/human endogenous retrovirus W in patients with multiple sclerosis. J Neurovirol 14:143-51

Ruprecht K, Obojes K, Wengel V, Gronen F, Kim KS, Perron H, Schneider-Schaulies J, Rieckmann P (2006) Regulation of human endogenous retrovirus W protein expression by herpes simplex virus type 1: implications for multiple sclerosis. J Neurovirol 12:65-71

Sakuma T, Hue S, Squillace KA, Tonne JM, Blackburn PR, Ohmine S, Thatava T, Towers GJ, Ikeda Y (2011) No evidence of XMRV in prostate cancer cohorts in the Midwestern United States. Retrovirology 8:23

Samuelson LC, Wiebauer K, Gumucio DL, Meisler MH (1988) Expression of the human amylase genes: recent origin of a salivary amylase promoter from an actin pseudogene. Nucleic Acids Res 16:8261-76

Sanchez-Chapado M, Olmedilla G, Cabeza M, Donat E, Ruiz A (2003) Prevalence of prostate cancer and prostatic intraepithelial neoplasia in Caucasian Mediterranean males: an autopsy study. The Prostate 54:238-247

Sato E, Furuta RA, Miyazawa T (2010) An endogenous murine leukemia viral genome contaminant in a commercial RT-PCR kit is amplified using standard primers for XMRV. Retrovirology 7:110

Sauter M, Roemer K, Best B, Afting M, Schommer S, Seitz G, Hartmann M, Mueller-Lantzsch N (1996) Specificity of antibodies directed against Env protein of human endogenous retroviruses in patients with germ cell tumors. Cancer Res 56:4362-5

Sauter M, Schommer S, Kremmer E, Remberger K, Dolken G, Lemm I, Buck M, Best B, Neumann-Haefelin D, Mueller-Lantzsch N (1995) Human endogenous retrovirus K10: expression of Gag protein and detection of antibodies in patients with seminomas. J Virol 69:414-21

Schanab O, Humer J, Gleiss A, Mikula M, Sturlan S, Grunt S, Okamoto I, Muster T, Pehamberger H, Waltenberger A (2011) Expression of human endogenous retrovirus K is stimulated by ultraviolet radiation in melanoma. Pigment cell & melanoma research 24:656-665

Schiavetti F, Thonnard J, Colau D, Boon T, Coulie PG (2002) A human endogenous retroviral sequence encoding an antigen recognized on melanoma by cytolytic T lymphocytes. Cancer Res 62:5510-6

Schlaberg R, Choe DJ, Brown KR, Thaker HM, Singh IR (2009) XMRV is present in malignant prostatic epithelium and is associated with prostate cancer, especially high-grade tumors. Proceedings of the National Academy of Sciences of the United States of America 106:16351-16356

Schmitz-Winnenthal FH, Galindo-Escobedo LV, Rimoldi D, Geng W, Romero P, Koch M, Weitz J, Krempien R, Niethammer AG, Beckhove P, Buchler MW, Z'Graggen K (2007) Potential target antigens for immunotherapy in human pancreatic cancer. Cancer Lett 252:290-8

Schon U, Diem O, Leitner L, Gunzburg WH, Mager DL, Salmons B, Leib-Mosch C (2009) Human endogenous retroviral long terminal repeat sequences as cell type-specific promoters in retroviral vectors. Journal of Virology 83:12643-12650

Schon U, Seifarth W, Baust C, Hohenadl C, Erfle V, Leib-Mosch C (2001) Cell type-specific expression and promoter activity of human endogenous retroviral long terminal repeats. Virology 279:280-91

Schön U, Seifarth W, Baust C, Hohenadl C, Erfle V, Leib-Mosch C (2001) Cell type-specific expression and promoter activity of human endogenous retroviral long terminal repeats. Virology 279:280-291

Schroder FH, Hugosson J, Roobol MJ, Tammela TL, Ciatto S, Nelen V, Kwiatkowski M, Lujan M, Lilja H, Zappa M, Denis LJ, Recker F, Berenguer A, Maattanen L, Bangma CH, Aus G, Villers A, Rebillard X, van der KT, Blijenberg BG, Moss SM, de Koning HJ, Auvinen A (2009) Screening and prostate-cancer mortality in a randomized European study. The New England journal of medicine 360:1320-1328

Schubert SW, Lamoureux N, Kilian K, Klein-Hitpass L, Hashemolhosseini S (2008) Identification of integrin-alpha4, Rb1, and syncytin a as murine placental target genes of the transcription factor GCMa/Gcm1. J Biol Chem 283:5460-5

Schulte AM, Lai S, Kurtz A, Czubayko F, Riegel AT, Wellstein A (1996) Human trophoblast and choriocarcinoma expression of the growth factor pleiotrophin attributable to germ-line insertion of an endogenous retrovirus. Proc Natl Acad Sci U S A 93:14759-64

Seifarth W, Baust C, Murr A, Skladny H, Krieg-Schneider F, Blusch J, Werner T, Hehlmann R, Leib-Mosch C (1998) Proviral structure, chromosomal location, and expression of HERV-K-T47D, a novel human endogenous retrovirus derived from T47D particles. J Virol 72:8384-91

Seifarth W, Frank O, Zeilfelder U, Spiess B, Greenwood AD, Hehlmann R, Leib-Mosch C (2005) Comprehensive analysis of human endogenous retrovirus transcriptional activity in human tissues with a retrovirus-specific microarray. J Virol 79:341-52

Seifarth W, Krause U, Hohenadl C, Baust C, Hehlmann R, Leib-Mösch C (2000) Rapid identification of all known retroviral reverse transcriptase sequences with a novel versatile detection assay. AIDS Res Hum Retroviruses 16 (8):721-729

Seifarth W, Skladny H, Krieg-Schneider F, Reichert A, Hehlmann R, Leib-Mosch C (1995) Retrovirus-like particles released from the human breast cancer cell line T47-D display type B- and C-related endogenous retroviral sequences. J Virol 69:6408-16

Seifarth W, Spiess B, Zeilfelder U, Speth C, Hehlmann R, Leib-Mosch C (2003) Assessment of retroviral activity using a universal retrovirus chip. J Virol Methods 112:79-91

SenGupta D, Tandon R, Vieira RG, Ndhlovu LC, Lown-Hecht R, Ormsby CE, Loh L, Jones RB, Garrison KE, Martin JN, York VA, Spotts G, Reyes-Teran G, Ostrowski MA, Hecht FM, Deeks SG, Nixon DF (2011) Strong human endogenous retrovirus-specific T cell responses are associated with control of HIV-1 in chronic infection. Journal of Virology 85:6977-6985

Serafino A, Balestrieri E, Pierimarchi P, Matteucci C, Moroni G, Oricchio E, Rasi G, Mastino A, Spadafora C, Garaci E, Vallebona PS (2009) The activation of human endogenous retrovirus K (HERV-K) is implicated in melanoma cell malignant transformation. Exp Cell R 315:849-862

Shchepinov MS, Case-Green SC, Southern EM (1997) Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays. Nucl Acids Res 25:1155-1161

Shiroma T, Sugimoto J, Oda T, Jinno Y, Kanaya F (2001) Search for active endogenous retroviruses: identification and characterization of a HERV-E gene that is expressed in the pancreas and thyroid. J Hum Genet 46:619-25

Sicat J, Sutkowski N, Huber BT (2005) Expression of human endogenous retrovirus HERV-K18 superantigen is elevated in juvenile rheumatoid arthritis. J Rheumatol 32:1821-31

Silva IT, Vencio RZ, Oliveira TY, Molfetta GA, Silva WA, Jr. (2010) ProbFAST: Probabilistic functional analysis system tool. BMC Bioinformatics 11:161

Sin HS, Huh JW, Kim DS, Kang DW, Min DS, Kim TH, Ha HS, Kim HH, Lee SY, Kim HS (2006a) Transcriptional control of the HERV-H LTR element of the GSDML gene in human tissues and cancer cells. Arch Virol 151:1985-94

Sin HS, Huh JW, Kim DS, Kim TH, Ha HS, Kim WY, Park HK, Kim CM, Kim HS (2006b) Endogenous retrovirus-related sequences provide an alternative transcript of MCJ genes in human tissues and cancer cells. Genes Genet Syst 81:333-9

Sin HS, Koh E, Taya M, Iijima M, Sugimoto K, Maeda Y, Yoshida A, Iwamoto T, Namiki M (2011) A novel Y chromosome microdeletion with the loss of an endogenous retrovirus related, testis specific transcript in AZFb region. The Journal of urology 186:1545-1552

Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer cell 1:203-209

Singh R, Maganti RJ, Jabba SV, Wang M, Deng G, Heath JD, Kurn N, Wangemann P (2005) Microarray-based comparison of three amplification methods for nanogram amounts of total RNA. American journal of physiology Cell physiology 288:C1179-C1189

Sjottem E, Anderssen S, Johansen T (1996) The promoter activity of long terminal repeats of the HERV-H family of human retrovirus-like elements is critically dependent on Sp1 family proteins interacting with a GC/GT box located immediately 3' to the TATA box. J Virol 70:188-98

Skog J, Wurdinger T, van RS, Meijer DH, Gainche L, Sena-Esteves M, Curry WT, Jr., Carter BS, Krichevsky AM, Breakefield XO (2008) Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers. Nat Cell Biol 10:1470-1476

Smallwood A, Papageorghiou A, Nicolaides K, Alley MK, Jim A, Nargund G, Ojha K, Campbell S, Banerjee S (2003) Temporal regulation of the expression of syncytin (HERV-W), maternally imprinted PEG10, and SGCE in human placenta. Biol Reprod 69:286-93

Smit AFA, Hubley R, Green P (1996) RepeatMasker Open-3.0.

Søe K, Andersen TL, Hobolt-Pedersen AS, Bjerregaard B, Larsson LI, Delaisse JM (2011) Involvement of human endogenous retroviral syncytin-1 in human osteoclast fusion. Bone 48:837-846

Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, Desmedt C, Larsimont D, Cardoso F, Peterse H, Nuyten D, Buyse M, van d, V, Bergh J, Piccart M, Delorenzi M (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. Journal of the National Cancer Institute 98:262-272

Stauffer Y, Marguerat S, Meylan F, Ucla C, Sutkowski N, Huber B, Pelet T, Conrad B (2001) Interferon-alpha-induced endogenous superantigen. a model linking environment and autoimmunity. Immunity 15:591-601

Stauffer Y, Theiler G, Sperisen P, Lebedev Y, Jongeneel CV (2004) Digital expression profiles of human endogenous retroviral families in normal and cancerous tissues. Cancer immunity : a journal of the Academy of Cancer Immunology 4:2

Stengel A, Bach C, Vorberg I, Frank O, Gilch S, Lutzny G, Seifarth W, Erfle V, Maas E, Schatzl H, Leib-Mosch C, Greenwood AD (2006a) Prion infection influences murine endogenous retrovirus expression in neuronal cells. Biochemical and biophysical research communications 343:825-831

Stengel A, Roos C, Hunsmann G, Seifarth W, Leib-Mosch C, Greenwood AD (2006b) Expression profiles of endogenous retroviruses in Old World monkeys. Journal of Virology 80:4415-4421

Stengel S, Fiebig U, Kurth R, Denner J (2010) Regulation of human endogenous retrovirus-K expression in melanomas by CpG methylation. Genes, chromosomes & cancer 49:401-411

Stieler K, Schulz C, Lavanya M, Aepfelbacher M, Stocking C, Fischer N (2010) Host range and cellular tropism of the human exogenous gammaretrovirus XMRV. Virology 399:23-30

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 100:9440-9445

Strick R, Ackermann S, Langbein M, Swiatek J, Schubert SW, Hashemolhosseini S, Koscheck T, Fasching PA, Schild RL, Beckmann MW, Strissel PL (2007) Proliferation and cell-cell fusion of endometrial carcinoma are induced by the human endogenous retroviral Syncytin-1 and regulated by TGF-beta. J Mol Med 85:23-38

Subramanian RP, Wildschutte JH, Russo C, Coffin JM (2011) Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology 8:90

Sun C, Skaletsky H, Rozen S, Gromoll J, Nieschlag E, Oates R, Page DC (2000) Deletion of azoospermia factor a (AZFa) region of human Y chromosome caused by recombination between HERV15 proviruses. Hum Mol Genet 9:2291-6

Sun T, Zhao Y, Mangelsdorf DJ, Simpson ER (1998) Characterization of a region upstream of exon I.1 of the human CYP19 (aromatase) gene that mediates regulation by retinoids in human choriocarcinoma cells. Endocrinology 139:1684-1691

Sun Y, Ouyang DY, Pang W, Tu YQ, Li YY, Shen XM, Tam SC, Yang HY, Zheng YT (2010) Expression of syncytin in leukemia and lymphoma cells. Leukemia research 34:1195-1202

Sutkowski N, Chen G, Calderon G, Huber BT (2004) Epstein-Barr virus latent membrane protein LMP-2A is sufficient for transactivation of the human endogenous retrovirus HERV-K18 superantigen. J Virol 78:7852-60

Sutkowski N, Conrad B, Thorley-Lawson DA, Huber BT (2001) Epstein-Barr virus transactivates the human endogenous retrovirus HERV-K18 that encodes a superantigen. Immunity 15:579-589

Svensson AC, Raudsepp T, Larsson C, Di CA, Chowdhary B, La MG, Rask L, Andersson G (2001) Chromosomal distribution, localization and expression of the human endogenous retrovirus ERV9. Cytogenetics and cell genetics 92:89-96

Szpakowski S, Sun X, Lage JM, Dyer A, Rubinstein J, Kowalski D, Sasaki C, Costa J, Lizardi PM (2009) Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements. Gene 448:151-167

Tai AK, Luka J, Ablashi D, Huber BT (2009) HHV-6A infection induces expression of HERV-K18-encoded superantigen. Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology 46:47-48

Takeuchi K, Katsumata K, Ikeda H, Minami M, Wakisaka A, Yoshiki T (1995) Expression of endogenous retroviruses, ERV3 and lambda 4-1, in synovial tissues from patients with rheumatoid arthritis. Clinical and experimental immunology 99:338-344

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods 6:377-382

Tarlinton R, Meers J, Hanger J, Young P (2005) Real-time reverse transcriptase PCR for the endogenous koala retrovirus reveals an association between plasma viral load and neoplastic disease in koalas. The Journal of general virology 86:783-787

Tarlinton R, Meers J, Young P (2008) Biology and evolution of the endogenous koala retrovirus. Cell Mol Life Sci 65:3413-3421

Tarlinton RE, Meers J, Young PR (2006) Retroviral invasion of the koala genome. Nature 442:79-81

Temin HM (1980) Origin of retroviruses from cellular movable genetic elements. Cell 21:599-600

Tempel S (2012) Using and understanding RepeatMasker. Methods in molecular biology (Clifton, N J) 859:29-51

Tempel S, Jurka M, Jurka J (2008) VisualRepbase: an interface for the study of occurrences of transposable element families. BMC Bioinformatics 9:345

Ting CN, Rosenberg MP, Snow CM, Samuelson LC, Meisler MH (1992) Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. Genes Dev 6:1457-65

Toda K, Nomoto S, Shizuta Y (1996) Identification and characterization of transcriptional regulatory elements of the human aromatase cytochrome P450 gene (CYP19). J Steroid Biochem Mol Biol 56:151-9

Tomita N, Horii A, Doi S, Yokouchi H, Ogawa M, Mori T, Matsubara K (1990) Transcription of human endogenous retroviral long terminal repeat (LTR) sequence in a lung cancer cell line. Biochemical and biophysical research communications 166:1-10

Tomlins SA, Laxman B, Dhanasekaran SM, Helgeson BE, Cao X, Morris DS, Menon A, Jing X, Cao Q, Han B, Yu J, Wang L, Montie JE, Rubin MA, Pienta KJ, Roulston D, Shah RB, Varambally S, Mehra R, Chinnaiyan AM (2007a) Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature 448:595-599

Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, Shah RB, Chinnaiyan AM (2007b) Integrative molecular concept modeling of prostate cancer progression. Nature genetics 39:41-51

Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science (New York, N Y) 310:644-648

Tong SY, Xie S, Richardson LJ, Ballard SA, Dakh F, Grabsch EA, Grayson ML, Howden BP, Johnson PD, Giffard PM (2011) High-resolution melting genotyping of Enterococcus faecium based on multilocus sequence typing derived single nucleotide polymorphisms. PLoS ONE 6:e29189

Toufaily C, Landry S, Leib-Mosch C, Rassart E, Barbeau B (2011) Activation of LTRs from different human endogenous retrovirus (HERV) families by the HTLV-1 tax protein and T-cell activators. Viruses 3:2146-2159

Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature reviews Genetics 13:36-46

Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M (2011) Next generation sequence assembly with AMOS. Current protocols in bioinformatics Chapter: Unit 11.8

Trejbalova K, Blazkova J, Matouskova M, Kucerova D, Pecnova L, Vernerova Z, Heracek J, Hirsch I, Hejnar J (2011) Epigenetic regulation of transcription and splicing of syncytins, fusogenic glycoproteins of retroviral origin. Nucl Acids Res 39:8728-8739

Trembath RC, Clough RL, Rosbotham JL, Jones AB, Camp RD, Frodsham A, Browne J, Barber R, Terwilliger J, Lathrop GM, Barker JN (1997) Identification of a major susceptibility locus on chromosome 6p and evidence for further disease loci revealed by a two stage genome-wide search in psoriasis. Human molecular genetics 6:813-820

Tristem M (2000) Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. J Virol 74:3715-30

True L, Coleman I, Hawley S, Huang CY, Gifford D, Coleman R, Beer TM, Gelmann E, Datta M, Mostaghel E, Knudsen B, Lange P, Vessella R, Lin D, Hood L, Nelson PS (2006) A molecular correlate to the Gleason grading system for prostate adenocarcinoma. Proceedings of the National Academy of Sciences of the United States of America 103:10991-10996

Turcanova VL, Bundgaard B, Hollsberg P (2009) Human herpesvirus-6B induces expression of the human endogenous retrovirus K18-encoded superantigen. Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology 46:15-19

Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J (2001) Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr Biol 11:1531-1535

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America 98:5116-5121

U.S. Preventive Services Task Force (2012) Screening for Prostate Cancer: U.S. Preventive Services Task Force Recommendation Statement. Annals of internal medicine 157:120-134

Urisman A, Molinaro RJ, Fischer N, Plummer SJ, Casey G, Klein EA, Malathi K, Magi-Galluzzi C, Tubbs RR, Ganem D, Silverman RH, DeRisi JL (2006) Identification of a novel Gammaretrovirus in prostate tumors of patients homozygous for R462Q RNASEL variant. PLoS pathogens 2:e25

van Dam L, Kuipers EJ, van Leerdam ME (2010) Performance improvements of stool-based screening tests. Best practice & research Clinical gastroenterology 24:479-492

van de Lagemaat LN, Landry JR, Mager DL, Medstrand P (2003a) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 19:530-536

van de Lagemaat LN, Landry JR, Mager DL, Medstrand P (2003b) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 19:530-6

van de Lagemaat LN, Medstrand P, Mager DL (2006) Multiple effects govern endogenous retrovirus survival patterns in human gene introns. Genome Biol 7:R86

van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van d, V, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R (2002) A gene-expression signature as a predictor of survival in breast cancer. The New England journal of medicine 347:1999-2009

Van der Kuyl AC, Cornelissen M, Berkhout B (2010) Of Mice and Men: On the Origin of XMRV. Frontiers in microbiology 1:147

van Regenmortel MH, Mayo MA, Fauquet CM, Maniloff J (2000) Virus nomenclature: consensus versus chaos. Arch Virol 145:2227-2232

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biology 3:research0034-research0034.11

Varambally S, Yu J, Laxman B, Rhodes DR, Mehra R, Tomlins SA, Shah RB, Chandran U, Monzon FA, Becich MJ, Wei JT, Pienta KJ, Ghosh D, Rubin MA, Chinnaiyan AM (2005) Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. Cancer cell 8:393-406

Vargas A, Moreau J, Landry S, LeBellego F, Toufaily C, Rassart E, Lafond J, Barbeau B (2009) Syncytin-2 plays an important role in the fusion of human trophoblast cells. J Mol Biol 392:301-318

Vargas A, Toufaily C, LeBellego F, Rassart E, Lafond J, Barbeau B (2011) Reduced expression of both syncytin 1 and syncytin 2 correlates with severity of preeclampsia. Reproductive sciences (Thousand Oaks, Calif) 18:1085-1091

Venables PJ, Brookes SM, Griffiths D, Weiss RA, Boyd MT (1995) Abundance of an endogenous retroviral envelope protein in placental trophoblasts suggests a biological function. Virology 211:589-92

Vencio RZ, Koide T (2005) HTself: self-self based statistical test for low replication microarray studies. DNA research : an international journal for rapid publication of reports on genes and genomes 12:211-214

Vencio RZ, Koide T, Gomes SL, Pereira CA (2006) BayGO: Bayesian analysis of ontology term enrichment in microarray data. BMC Bioinformatics 7:86

Venet D, Dumont JE, Detours V (2011) Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS computational biology 7:e1002240

Vernochet C, Heidmann O, Dupressoir A, Cornelis G, Dessen P, Catzeflis F, Heidmann T (2011) A syncytin-like endogenous retrovirus envelope gene of the guinea pig specifically expressed in the placenta junctional zone and conserved in Caviomorpha. Placenta 32:885-892

Villesen P, Aagaard L, Wiuf C, Pedersen FS (2004) Identification of endogenous retroviral reading frames in the human genome. Retrovirology 1:32

Vinogradova T, Leppik L, Kalinina E, Zhulidov P, Grzeschik KH, Sverdlov E (2002) Selective Differential Display of RNAs containing interspersed repeats: analysis of changes in the transcription of HERV-K LTRs in germ cell tumors. Mol Genet Genomics 266:796-805

Vinogradova TV, Leppik LP, Nikolaev LG, Akopov SB, Kleiman AM, Senyuta NB, Sverdlov ED (2001) Solitary human endogenous retroviruses-K LTRs retain transcriptional activity in vivo, the mode of which is different in different cell types. Virology 290:83-90

Vlaeminck-Guillem V, Ruffion A, Andre J, Devonec M, Paparel P (2010) Urinary prostate cancer 3 test: toward the age of reason? Urology 75:447-453

Vogetseder W, Dumfahrt A, Mayersbach P, Schönitzer D, Dierich MP (1993) Antibodies in human sera recognizing a recombinant outer membrane protein encoded by the envelope gene of the human endogenous retrovirus K. AIDS Res Hum Retroviruses 9 (7):687-694

Voisset C, Bouton O, Bedin F, Duret L, Mandrand B, Mallet F, Paranhos-Baccala G (2000) Chromosomal distribution and coding capacity of the human endogenous retrovirus HERV-W family. AIDS Res Hum Retroviruses 16:731-40

Vyse TJ, Todd JA (1996) Genetic analysis of autoimmune disease. Cell 85:311-318

Walsh CP, Chaillet JR, Bestor TH (1998) Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nature Genet 20:116-117

Wang X, Kruithof-de JM, Economides KD, Walker D, Yu H, Halili MV, Hu YP, Price SM, bate-Shen C, Shen MM (2009) A luminal epithelial stem cell that is a cell of origin for prostate cancer. Nature 461:495-500

Wang Y, Pelisson I, Melana SM, Holland JF, Pogo BG (2001) Detection of MMTV-like LTR and LTR-env gene sequences in human breast cancer. Int J Oncol 18:1041-1044

Wang-Johanning F, Frost AR, Jian B, Azerou R, Lu DW, Chen DT, Johanning GL (2003a) Detecting the expression of human endogenous retrovirus E envelope transcripts in human prostate adenocarcinoma. Cancer 98:187-97

Wang-Johanning F, Frost AR, Jian B, Epp L, Lu DW, Johanning GL (2003b) Quantitation of HERV-K env gene expression and splicing in human breast cancer. Oncogene 22:1528-35

Wang-Johanning F, Frost AR, Johanning GL, Khazaeli MB, LoBuglio AF, Shaw DR, Strong TV (2001) Expression of human endogenous retrovirus k envelope transcripts in human breast cancer. Clin Cancer Res 7:1553-60

Wang-Johanning F, Liu J, Rycaj K, Huang M, Tsai K, Rosen DG, Chen DT, Lu DW, Barnhart KF, Johanning GL (2007) Expression of multiple human endogenous retrovirus surface envelope proteins in ovarian cancer. Int J Cancer 120:81-90

Wang-Johanning F, Radvanyi L, Rycaj K, Plummer JB, Yan P, Sastry KJ, Piyathilake CJ, Hunt KK, Johanning GL (2008) Human endogenous retrovirus K triggers an antigen-specific immune response in breast cancer patients. Cancer Res 68:5869-77

Wang-Johanning F, Rycaj K, Plummer JB, Li M, Yin B, Frerich K, Garza JG, Shen J, Lin K, Yan P, Glynn SA, Dorsey TH, Hunt KK, Ambs S, Johanning GL (2012) Immunotherapeutic potential of anti-human endogenous retrovirus-K envelope protein antibodies in targeting breast tumors. Journal of the National Cancer Institute 104:189-210

Watanabe T, Takeda A, Tsukiyama T, Mise K, Okuno T, Sasaki H, Minami N, Imai H (2006) Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. Genes Dev 20:1732-43

Watson JD, Wang S, Von Stetina SE, Spencer WC, Levy S, Dexheimer PJ, Kurn N, Heath JD, Miller DM, III (2008) Complementary RNA amplification methods enhance microarray identification of transcripts expressed in the C. elegans nervous system. BMC Genomics 9:84

Webber MM, Quader ST, Kleinman HK, Bello-DeOcampo D, Storto PD, Bice G, Mendonca-Calaca W, Williams DE (2001) Human cell lines as an in vitro/in vivo model for prostate carcinogenesis and progression. The Prostate 47:1-13

Weis S, Llenos IC, Sabunciyan S, Dulay JR, Isler L, Yolken R, Perron H (2007) Reduced expression of human endogenous retrovirus (HERV)-W GAG protein in the cingulate gyrus and hippocampus in schizophrenia, bipolar disorder, and depression. Journal of neural transmission (Vienna, Austria : 1996) 114:645-655

Weiss RA (2010) A cautionary tale of virus and disease. BMC biology 8:124

Wentzensen N, Coy JF, Knaebel HP, Linnebacher M, Wilz B, Gebert J, von Knebel Doeberitz M (2007) Expression of an endogenous retroviral sequence from the HERV-H group in gastrointestinal cancers. Int J Cancer 121:1417-23

Wentzensen N, Wilz B, Findeisen P, Wagner R, Dippold W, von Knebel Doeberitz M, Gebert J (2004) Identification of differentially expressed genes in colorectal adenoma compared to normal tissue by suppression subtractive hybridization. Int J Oncol 24:987-94

Wick LM, Rouillard JM, Whittam TS, Gulari E, Tiedje JM, Hashsham SA (2006) On-chip non-equilibrium dissociation curves and dissociation rate constants as methods to assess specificity of oligonucleotide probes. Nucl Acids Res 34:e26

Wiegel T, Hinkelbein W (1998) [Locally advanced prostate carcinoma (T2b-T4 N0) without and with clinical evidence of local progression (Tx N+) with lymphatic metastasis. Is radiotherapy for pelvic lymphatic metastasis indicated or not?]. Strahlentherapie und Onkologie : Organ der Deutschen Rontgengesellschaft [et al] 174:231-236

Willer A, Saussele S, Gimbel W, Seifarth W, Kister P, Leib-Mösch C, Hehlmann R (1997) Two groups of endogenous MMTV related retroviral *env* transcripts expressed in human tissues. Virus Genes 15:123-133

Wittwer CT, Reed GH, Gundry CN, Vandersteen JG, Pryor RJ (2003) High-resolution genotyping by amplicon melting analysis using LCGreen. Clinical chemistry 49:853-860

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9:3353-3362

Yang C, Compans RW (1996) Analysis of the cell fusion activities of chimeric simian immunodeficiency virus-murine leukemia virus envelope proteins: inhibitory effects of the R peptide. J Virol 70:248-254

Yao Y, Schroder J, Nellaker C, Bottmer C, Bachmann S, Yolken RH, Karlsson H (2008) Elevated levels of human endogenous retrovirus-W transcripts in blood cells from patients with first episode schizophrenia. Genes Brain Behav 7:103-12

Yin H, Medstrand P, Andersson ML, Borg A, Olsson H, Blomberg J (1997) Transcription of human endogenous retroviral sequences related to mouse mammary tumor virus in human breast and placenta: similar pattern in most malignant and nonmalignant breast tissues. AIDS Research and human retroviruses 13 (6):507-516

You WC, Jin F, Devesa S, Gridley G, Schatzkin A, Yang G, Rosenberg P, Xiang YB, Hu YR, Li Q (2002) Rapid increase in colorectal cancer rates in urban Shanghai, 1972-97, in relation to dietary changes. Journal of cancer epidemiology and prevention 7:143-146

Yu C, Shen K, Lin M, Chen P, Lin C, Chang GD, Chen H (2002) GCMa regulates the syncytin-mediated trophoblastic fusion. J Biol Chem 277:50062-8

Yu YP, Landsittel D, Jing L, Nelson J, Ren B, Liu L, McDonald C, Thomas R, Dhir R, Finkelstein S, Michalopoulos G, Becich M, Luo JH (2004) Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 22:2790-2799

Zeilfelder U, Frank O, Sparacio S, Schon U, Bosch V, Seifarth W, Leib-Mosch C (2007) The potential of retroviral vectors to cotransfer human endogenous retroviruses (HERVs) from human packaging cell lines. Gene 390:175-179

Zhao J, Rycaj K, Geng S, Li M, Plummer JB, Yin B, Liu H, Xu X, Zhang Y, Yan Y, Glynn SA, Dorsey TH, Ambs S, Johanning GL, Gu L, Wang-Johanning F (2011) Expression of Human Endogenous Retrovirus Type K Envelope Protein is a Novel Candidate Prognostic Marker for Human Breast Cancer. Genes & cancer 2:914-922

1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061-1073

IV Annexes

IV.1. Revue : A Comparative Portrait of Retroviral Fusogens and Syncytins

Chapter 4
A Comparative Portrait of Retroviral Fusogens and Syncytins

Philippe Pérot, Cécile Montgiraud, Dimitri Lavillette, and François Mallet

Abstract The strongest candidates for developmentally regulated cellular fusogens in mammals are Syncytins which contribute to cell–cell fusion leading to placental syncytiotrophoblast in higher primates, rodents, lagomorphs and sheeps. They consist of domesticated endogenous retroviral envelope glycoproteins (Env) whose fusion properties depend on the initial recognition of a specific receptor. In order to clearly understand Syncytins characteristics, we will first illustrate molecular details characterizing the maturation of class I fusion proteins by introducing envelope-driven fusion in an infectious context, i.e. virus cell fusion, exemplifying each step that lead to functional virions with the most relevant model such as HIV-1 lentivirus or MLV and type D interference group retroviruses. In a second part, we will comparatively present the current knowledge concerning Syncytins and the associated three levels of complexity. First, the placenta is probably more variable in structure than any of the mammalian organs. Second, Syncytins recognize specific and highly function-divergent/unrelated receptors. Third, some Syncytins were shown to exhibit other functions than fusion, such as proliferation, immunomodulation, receptor interference and anti-apoptotic properties. We will conclude by a brief overview of the consequences of Syncytin expression outside of its privileged tissue.

Keywords Fusion · placenta · retrovirus · endogenous retrovirus · envelope · Syncytin · enJSRV · receptor · hASCT1 · hASCT2 · MFSD2 · HYAL2

Abbreviations

ASCT	Alanine, serine and cysteine selective transporters
ALV	Avian leukosis virus
ASLV	Avian sarcoma leukosis virus
BaEV	Baboon endogenous retrovirus

F. Mallet (✉)
Laboratoire Commun de Recherche Hospices Civils de Lyon – bioMérieux, Cancer Biomarkers Research Group, Centre Hospitalier Lyon Sud, 69495 Pierre Bénite cedex, France
e-mail: francois.mallet@eu.biomerieux.com

Cécile Montgiraud and Dimitri Lavillette contributed equally to in this chapter

L.-I. Larsson (ed.), *Cell Fusions*, DOI 10.1007/978-90-481-9772-9_4,

BLV	Bovine leukemia virus
CA	Capsid
CAT-1	Cationic amino acid transporter-1
CT	Cytotrophoblast
cyt	Cytoplasmic tail
DRM	Rafts?
EBV	Epstein-Barr virus
ECT	Extravillous cytotrophoblasts
en	Endogenous
EnCa	Endometrial carcinoma
Env	Envelope
ER	Endoplasmic reticulum
ERV	Endogenous retrovirus
ESCRT	Endosomal sorting complex required for transport
Exo	Exogenous
FeLV	Feline leukemia virus
FcEV	Felis catus endogenous retrovirus
FP	Fusion peptide
GaLV	Gibbon ape leukemia virus
GCM	Glial cell missing
Gp	Glycoprotein
GPI	Glycosylphosphatidylinositol
h	Human
HELLP	Hemolysis, elevated liver enzymes and low platelets
HERV	Human endogenous retrovirus
HIV	Human immunodeficiency virus
HR	Heptad repeats
HTDV	Human teratocarcinoma-derived virus
HTLV	Human T-cell leukemia virus
HYAL2	Hyaluronidase 2
IDO	Indoleamine 2,3-dioxygenase
IFN	Interferon
Ig	Immunoglobulin
IL	Interleukin
IP3	Inositol-3-phosphate
JSRV	Jaagsiekte sheep retrovirus
KoRV	Koala retrovirus
LLP	Lentivirus lytic peptides
LTR	Long terminal repeat
m	Mouse
M	Mesenchyme
MA	Matrix
MAO	Morpholino antisense oligonucleotide
MFSD2	Major facilitator superfamily domain containing 2
MLV	Murine leukemia virus
MMTV	Mouse mammary tumor virus
Mo-MLV	Moloney murine leukemia virus

M-PMV	Mason-Pfizer monkey virus
MS	Multiple sclerosis
MSRV	Multiple sclerosis associated retrovirus
MuLV	Murine leukemia virus
MVB	Multivesicular bodies
MYA	Million years ago
NC	Nucleocapsid
NO	Nitric oxide
NWM	New world monkeys
OASIS	Old astrocytes specifically induced substance
ORF	Open reading frame
OWM	Old world monkeys
PCR	Polymerase chain reaction
PcRV	Papio cynocephalus retrovirus
PDI	Protein disulfide isomerase
PE	Preeclampsia
Pit	Sodium-dependent phosphate symporter
RBD	Receptor-binding domain
RT	Reverse transcriptase
SERV	Simian endogenous retrovirus
SIV	Simian immunodeficiency virus
SRV-1	Simian retrovirus-1
SNARE	Soluble N-ethylmaleimide-sensitive fusion protein attachment protein receptor
SP	Signal peptide
SRP	Signal recognition particle
ST	Syncytiotrophoblast
SU	Surface unit
T	Trophoblast
TfR1	Transferrin receptor 1
TGF	Transforming growth factor
TGN	Trans-Golgi network
Th	T helper cells
TM	Transmembrane unit
tm	Transmembrane domain
TNF	Tumor necrosis factor
TSE	Trophoblast specific enhancer
UA	Uterine arteries
URE	Upstream regulatory element

Contents

4.1 Introduction

Our life begins with fusion of our mother's oocyte with one of our father's spermatozoids. A consequent successful pregnancy depends on the 10 m^2 placental syncytiotrophoblast resulting from fusion of abundant cytotrophoblastic cells. As the embryo develops, skeletal muscle differentiation depends on the fusion of mononucleated myoblasts to form multinucleated muscle fibers. In the adult body, macrophages can fuse to form either multinucleated osteoclasts that control the maintenance of the bones or multinucleated giant cells that are important for the immune response. Last but not least, fusion as a driver of embryonic stem cell differentiation suggests a new role of cell fusion in mammalian development.

Overall, cell fusion is a process in which two or more cells become one by merging their plasma membranes. Fusion, with the exception of gametes and stem cells, produces only terminally differentiated, non-proliferating tissue, and is thus mainly involved in tissue maintenance or regeneration. The fused cells (syncytia) that contain several nuclei within a single cytoplasm may be homokaryons (homotypic fusion) or heterokaryons (heterotypic fusion) as derived from the fusion of similar or different origin cell types, respectively. In any case, fusion of two separate lipid bilayers in non aqueous environment first requires that they come in close contact. Second, an intermediate stage is characterized by the merger of only the closest contacting monolayer, a process called hemifusion. Third, the fully completed fusion results in whole bilayer merging following by the opening of the pore.

It remains questionable whether cell–cell fusion involves the same type of mechanisms than in other membrane fusion events, such as intracellular vesicle fusion mainly based on SNAREs proteins (soluble N-ethylmaleimide-sensitive fusion protein attachment protein receptor) and virus-cell fusion achieved by transmembrane viral fusion proteins (Chen and Olson 2005). However, we might expect that long-conserved syncytial structures, such as skeletal muscle that have become integral to the body plans of multiple phyla, may be formed by mechanisms that have been mostly conserved during evolution (Mohler 2009). It is thus striking to notice that little is known about the molecular actors that are involved in regulating and completing cell–cell fusion, and of what is known there is little conservation between species, suggesting that these mechanisms might have evolved independently (Chen et al. 2007).

The multiple transmembrane-domain tetraspanin protein CD9 on the egg surface and the single transmembrane-domain protein IZUMO with an extracellular

immunoglobulin (Ig)-like domain on the spermatozoid surface contribute to the fusion of the mice gametes, although it is still a matter of conjecture whether the two molecules interact directly in trans to achieve the membrane fusion reaction (Inoue et al. 2005). Transmembrane-domain proteins with extracellular Ig-like domains have been implicated in homotypic macrophage fusion in rats (Han et al. 2000) and in cell–cell tethering prior myoblast fusion in drosophila melanogaster (Bour et al. 2000). Interestingly, the strongest candidates for developmentally regulated cellular fusogens in mammals are Syncytins, a family of single-pass transmembrane proteins, which contribute to cell–cell fusion leading to placental syncytiotrophoblast in higher primates, rodents, lagomorphs and sheeps. They consist of domesticated endogenous retroviral envelope glycoproteins whose fusion properties depend on the initial recognition of a specific receptor.

With the exception of retrovirus-derived Syncytins, none of the cell surface proteins identified in the various cell–cell fusion processes resemble SNAREs or class I fusion protein, (i.e. fusion does not appear to be mediated by an α-helical bundle). Though, fusogenic proteins contribute to decrease the kinetic barrier to allow the fusion of the two bilayer membranes. Viral fusion proteins do so by using the force energy released during a protein conformational change to draw together the membranes. The understanding of the Syncytins dependant cell–cell fusions will likely parallel the mechanism of at least retroviral infection. Indeed, Syncytins are host domesticated genes derived from ancient retroviruses infections of the host germ line. Syncytins appear to group relatively distinct actors that may exhibit common principles leading to membrane fusion and hence are good examples of the various scheme of evolution to establish similar but different structures (microscopic and macroscopic) with similar roles. Such a dichotomy between distinct players with common principles was indeed proposed for all fusion processes by Martens and McMahon (2008).

Three classes of viral fusogens have been described. The class I and II fusion proteins are characterized by trimers of hairpins containing a central α-helical coiled-coil or β-sheets structure, respectively, while the class III fusion proteins have a mixed secondary structure (Weissenhorn et al. 2007). We will first introduce envelope-driven fusion in an infectious context, i.e. virus-cell fusion, by illustrating each step leading to functional virions with the most relevant model such as HIV-1 lentivirus or MLV and type D interference group retroviruses. The purpose is to illustrate molecular details characterizing the maturation of class I fusion proteins, defined by the following four characteristics: the cleavage of an envelope protein precursor leading to surface and transmembrane subunits, a fusion peptide located just next to the cleavage site (except avian ASLV), a trimeric complex association, and the ability to form a hairpin structure, also called a coiled-coil structure, in its active fusion conformation. The progression of these structural rearrangements slows down the kinetic barrier between hemifusion and fusion-pore formation.

Intriguingly, without unequivocal evidence of infectious agent, retroviral particles were observed in physiological situation (Lyden et al. 1994) but also in pathological ones (Perron et al. 1989; Boller et al. 1993) in man. They could derived from endogenous retroviral sequences, as the human genome (Lander et al. 2001) but also the mouse genome (Waterston et al. 2002) contain a huge amount of

endogenous retroviruses, reaching 8.5 and 9% of these genomes sizes, respectively. Although human and mouse species contain essentially different retroviral families, some of their coding sequences are still intact, e.g. 18 envelopes ORFs were identified in the human genome (Blaise et al. 2003; Blaise et al. 2005) including the two human Syncytins. This huge repertoire awaiting the identification of functions will be illustrated, as it may represent a third condition beside infectious viruses and domesticated envelope Syncytins.

In a second part, we will comparatively present the current knowledge on Syncytins. Outstandingly, the situation comprises at least three levels of complexity. First, the placenta is probably the more variable in structure than any of the mammalian organ. Placentas are variously classified, as regard to (i) their form, being discoid (primates, lagomorphs and rodents) or cotyledonary (ruminant), (ii) the type of layer between fetal trophoblast and maternal endometrial surface, hemochorial (primates, lagomorphs and rodents) or epithelio- and syndesmochorial (ruminants), and (iii) the structure of the maternal-fetal interdigitation, villous type or villi (primates and sheep) or labyrinth (lagomorphs and rodents) (Bernirschke K, Comparative placentation at http://placentation.ucsd.edu). The latter, a continuous syncytiotrophoblast layer that covers the entire surface of the human placental villi which floats in maternal blood, is responsible of ion and nutrient exchanges and synthesizes steroid and peptide hormones such as progesterone and human chorionic gonadotropin (hCG) required for human gestation. Second, Syncytins recognize specific and highly function-divergent/unrelated receptors. In human, Syncytin-1 recognizes hASCT1 (Blond et al. 2000) and hASCT2 (Lavillette et al. 2002) receptors while Syncytin-2 binds to MFSD2 receptor (Esnault et al. 2008). In rodents, Syncytin-A and Syncytin-B possess unidentified but distinct receptors (Dupressoir et al. 2005), and in lagomorphs Syncytin-Ory1 functionally recognizes hASCT2 (Heidmann et al. 2009). In sheep, enJSRV envelope(s) interacts with HYAL2 (Dunlap et al. 2006). This illustrates that proteins involved in cell–cell fusion, such as Syncytin partner receptors, are likely to play pleiotropic roles in other cellular processes, e.g. transport of small molecules, but also modulation of membrane structures, with specificity being achieved through the coupling of these proteins to different upstream and downstream effectors. Third, Syncytins were shown to exhibit other functions than fusion, such as proliferation (Strick et al. 2007; Larsen et al. 2009), immunomodulation (Mangeney et al. 2007), receptor interference (Blond et al. 2000; Ponferrada et al. 2003) and anti-apoptotic properties (Strick et al. 2007; Knerr et al. 2007), these functions being not shared by all these proteins. We will conclude by a brief overview of the consequences of Syncytin-1 expression outside of its privileged tissue.

4.2 Contribution of the Envelope to the Retroviral Life Cycle

Retroviral classification was initially based on virion morphology observed with electronic microscopy during maturation and assembly of particles (Coffin 1992). Accordingly, retroviruses are designated A-, B-, C- and D-type. The

International Comity of Taxonomy has now established seven genera of Retroviridae based on sequence homologies: Alpharetoviruses correspond to avian type C (Avian leukosis virus, ALV) retroviruses, Betaretroviruses to type B (Mouse Mammary Tumor Virus, MMTV) and D (Simian retrovirus-1, SRV-1) retroviruses, Gammaretroviruses to mammalian type C retroviruses (Murine Leukemia Virus, MLV), Deltaretroviruses to the ancient group of HTLV-BLV (Human T-cell Leukemia Virus/Bovine Leukemia Virus), Epsilonretroviruses (Waileye Dermal Sarcoma Viruses family), Lentiviruses group which contains HIV and SIV (Human and Simian Immunodeficiency Viruses) and Spumavirus including Human Foamy Viruses (van Regenmortel et al. 2000).

Retroviruses are RNA enveloped viruses. They infect cells *via* a cellular receptor recognition followed by the fusion of virus and cell membranes. Upon entry, the next step of the retroviral life cycle consists of a retrotranscription stage mediated by the viral reverse transcriptase protein that converts the viral genomic RNA in double strand DNA. Subsequently, the viral genetic material is targeted to the nucleus and stably integrated in the host cell genome by the viral integrase. The integrated viral DNA is named provirus and is flanked by two Long Terminal Repeats (LTR) that act as transcriptional regulatory elements. The 5′LTR contains the promoter and enhancer signals while the 3′LTR contains the polyadenylation signal terminating the transcription. All the replication competent retroviruses contain at least three genes coding for the structural proteins (*gag*), the enzymatic proteins (*pro-pol*) and the envelope glycoprotein (*env*). During its life cycle the virus uses the gene replication machinery of the host cell. Herein, we will focus on the characteristics of the envelope (Env) protein that is composed of one surface unit (SU) and one transmembrane unit (TM) which is itself subdivided into three domains, an ectodomain, a strict transmembrane domain (tm) and a cytoplasmic tail (cyt) (Fig. 4.1a). Env glycoprotein will undergo several modifications to generate a mature and functional glycoprotein addressed to the plasma membrane in order to contribute to the virus infection-competency. Functionally, the SU domain is involved in receptor recognition and the TM subunit contains both the fusion peptide and the heptad repeat domains involved in fusion.

4.2.1 Synthesis of Env Glycoprotein and Viral Assembly

During virus production, the host cell is basically preserved since the expression of fusogenic competent glycoproteins is highly controlled. Sequentially, the Env protein synthesis is initiated by the free-ribosomes, next modifications take place in the endoplasmic reticulum and then an oligomerized precursor is transported by vesicles to the golgi apparatus. Abundant glycoprotein at the surface of the cell could induce cellular death by syncytia formation, toxicity *via* receptor interaction, or immune recognition. That's why the localization and the amount of the oligomerized retroviral envelope glycoprotein at the host cellular surface are highly modulated by fine trafficking and sequestration mechanisms. The receptor interference mechanism

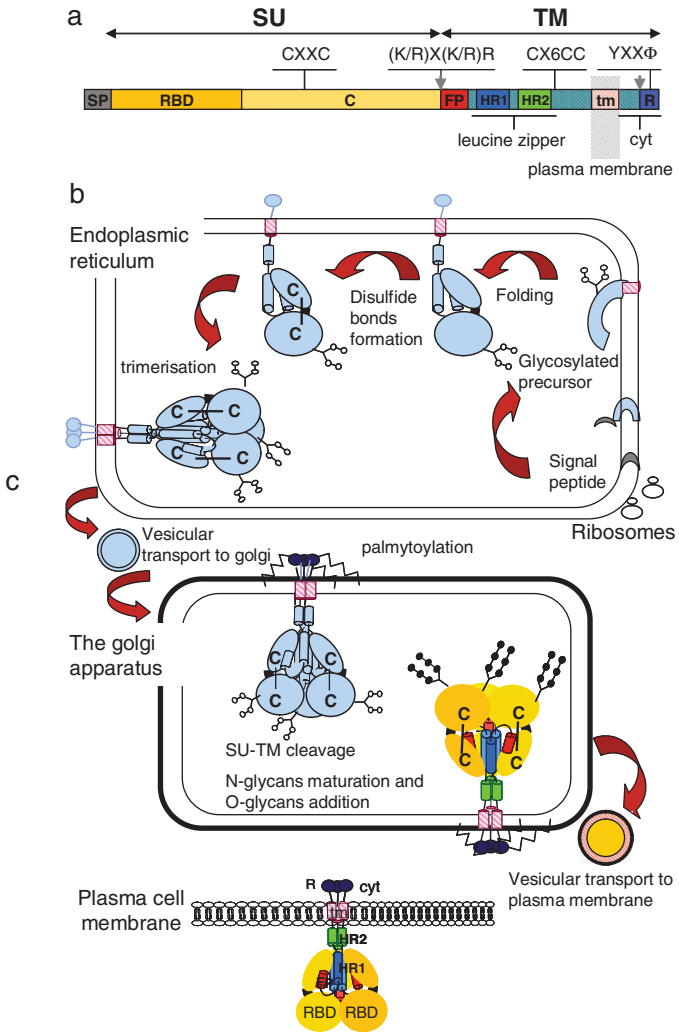


Fig. 4.1 Structure and synthesis of retroviral envelope. (**a**) Schematic portrait of an envelope prototype. SP, signal peptide (*grey*). SU, surface unit (*yellow*): contains RBD, receptor binding domain (*gold yellow*) and C, C-terminal domain (*light yellow*) with CXXC motif (generally CΦΦC with Φ = L,I,V,F,M or W). (K/R)X(K/R)R, furin cleavage site. TM, transmembrane unit (*hatched boxes*): contains FP, fusion peptide (*red*); leucine zipper motif with HR1 (*blue*) and HR2 (*green*) heptad repeats followed by the CX_6CC motif; tm, trans-membrane anchorage domain (*red*, hatched); cyt, cytoplasmic tail with C-terminal R peptide (*blue*) containing YXXΦ motif. Note that the ectodomain part of the TM contains a so-called immunosuppressive domain $QNRX_2LDXLX_5GXC$ joining the CX_6CC motif (not illustrated). (**b**) Schematic maturation process of the envelope glycoprotein. The successive immature forms of the envelope glycoprotein are illustrated (*petrol blue*). Initial glycosylation sites (branch trees with *open circles*), disulfide bonds (C–C), palmitoylation sites (*broken lines*) and final glycosylation sites (branch trees with *dark circles*) are indicated. Color codes and abbreviations used in the final trimeric structure expressed at the plasma cell membrane are as given in **a**; post translational modifications and disulfide bonds are omitted

can also limit the amount of receptors available for fusion between infected cells. Finally, Env fusion competency may be a late event that occurs during virus budding as described for MLV.

4.2.1.1 Synthesis and Maturation of Env Glycoprotein

As common cellular proteins, retroviral Env translation is initiated by free ribosomes in the cytosol. The signal peptide (SP) located in the N-terminus of the Env glycoprotein is the first synthesized segment. This initial step and the following ones are illustrated in Fig. 4.1b. SP length varies depending on the retrovirus family but its composition is conserved with an hydrophobic signal, recognized by the SRP (Signal Recognition Particle), that allows the anchorage of the nascent chain to the endoplasmic reticulum (ER). The nascent protein is translocated through the membrane of ER. At the end of the synthesis, the extracellular part of Env is folded in the lumen of the ER, as for cellular membrane proteins. The release of the protein in the lumen is impaired by a stop transfer region composed of a hydrophobic sequence followed by aromatic and charged amino acid which will delimitate the membrane anchored domain or transmembrane domain (tm) (Hunter and Swanstrom 1990). This tm domain is an α–helix constituted by at least 23 amino acids (for HIV-1) and a maximum of 36 amino acids (for MMTV) but it contains unexpected residus for alpha helix structure in the context of a membrane (helix breaker amino acids like glycine and proline or positively charged tryptophane or cysteine). The Env N-extremity (ectodomain in the future virion) is then located in the lumen of ER while the C-terminal part (cyt) of the protein remains in the cytosol. ER is then the site of co- and post-translational modifications such as N-glycosylations, protein folding, disulfide bonds formation and oligomerization (Ratner 1992).

After the SP cleavage, the precursor is modified by N-glycosylations. Depending on the retrovirus, the number and the location of glycosylation sites are variable. For HIV-1, the protein presents an unusual highly glycosylation with 24–32 sites and sugars account for half of the molecular weight of the Env protein (Mizuochi et al. 1990). For the other retroviruses, the number of glycosylation is around 8. Almost all glycosylations are in the SU, and except for gammaretroviruses for which there is no glycosylation in the TM, others have 1 (Betaretroviruses like BaEV) to 7 (HIV-1) glycosylations in TM reflecting the weak exposition of this subunit hidden by SU. In all cases, glycosylations are essential for the folding, the trafficking, the cleavage and the recognition of the receptor by the Env protein (Polonoff et al. 1982). For example, N-linked glycans are critical determinants for the efficient recognition of CD4 T cells by HIV-1 gp120, since mutant protein lacking one N-glycan did not effectively stimulate CD4+ T cells (Li et al. 2008a). For MLV, due to the fewer number of glycosylations than for HIV-1 Env, their roles have been more characterized and if some glycosylations are not crucial for incorporation, they are involved in the stability of the postcleavage envelope complex (Li et al. 1997).

Following glycosylation events, intramolecular disulfide bonds are formed in SU and TM subunits to generate loops in the secondary structure of the envelope protein. The cysteines of the SU involved in these bonds are well conserved in a

subgroup of retroviruses suggesting a similar domain organization. These cysteines are crucial for the folding and the transport of the envelope to different cellular compartments. The substitution of some cysteines in SU and TM of HIV-1 (Bolmstedt et al. 1991; Dedera et al. 1992) or other retroviruses like MuLV (Thomas and Roth 1995; Freed and Risser 1987) leads to a non cleaved envelope glycoprotein precursor retains into the cells. Finally, the loops generated by the disulfide bonds are essential for cellular receptor recognition (MacKrell et al. 1996). In addition, the TM of γ– and some β–retroviruses (but not for lentiviruses) contains specific cysteines that are important for SU-TM intermolecular association as it will be developed later. The disulfide bond formation is part of a more global control of modifications of the retroviral Env by the cell (Braakman and van Anken 2000). This control involved protein disulfide isomerase (PDI) (Fenouillet et al. 2007), and also chaperone molecules like GRB78 Bip (Earl et al. 1991), calnexin (Li et al. 1996) and calreticulin (Otteken and Moss 1996) as described for HIV-1. The failure to pass the quality control leads to a non cleaved envelope glycoprotein precursor that is kept either in endoplasmic reticulum or in the golgi apparatus. Retroviral Env glycoprotein leaves the ER in a trimeric form to reach the Golgi where N-glycans are matured, O-glycans added (Pinter and Honnen 1988; Bernstein et al. 1994) and the cysteines at the hedge of the tm domain are palmitoylated. (Yang et al. 1995) (Fig. 4.1b). The appropriate trimeric conformation (with glycosylation and disulfide bonds) being obtained, the precursor is then cleaved at a highly conserved site (Fig. 4.1a) among retroviruses by the furin-convertase protein into its two subunits e.g. gp120 (SU) and gp41 (TM) for HIV-1. The Env complex of HIV-1, as described for most retroviral Env glycoproteins, is trimeric with six individual subunits (three gp120 and three gp41 subunits). It is the TM subunit that triggers the oligomerization, as the TM (associated or not to the SU) is always detected in the oligomerized forms (Einfeld and Hunter 1988; Earl et al. 1990). The main determinant of this trimerization is a region in heptad repeat (Gallaher et al. 1989; Poumbourios et al. 1997) with high homology with leucine zipper domain (Fass et al. 1996; Weissenhorn et al. 1997; Owens and Compans 1990). Interestingly, the SU is also a trimer when it is shedded (Tucker 1991; Owens and Compans 1990) whereas soluble SU expressed alone is usually a monomer (Poumbourios et al. 1997). Hence, TM initiates the trimerization and, after that, SU can stay as a non-covalently linked fragile trimer. The trimerization gives the required environment for the fusion by masking the fusion peptide that will be later freed following receptor binding and also confers the fusogenic potential to the glycoprotein. The two SU and TM subunits are either linked in a covalent or non-covalent way. For HIV-1, the existence of the soluble gp120 protein indicates a non-covalent link between SU and TM (Kowalski et al. 1987). The regions implicated in this interaction are principally the C1 and C5 region of the SU and the leucine zipper domain and the CX_5C region of the TM (Lopalco et al. 1993; Schulz et al. 1992). For most others retroviruses a covalent link was described at one point. In all the case, except MMTV and JSRV, a disulfide bond between the SU and the TM is formed between the highly conserved CX_6CC motif of the TM and the CXXC of the SU (Sitbon et al. 1991; Schulz et al. 1992; Pinter et al. 1997). This CXXC motif is extremely rare in cellular proteins

and is similar to a motif found in the catalytic site of enzyme involved in thiol isomerisation like PDI or thioredoxin (Sanders 2000; Pinter et al. 1997). This motif in the SU has been shown to be part of an autocatalytic isomerisation function of SU to destroy the initial bond between SU and TM that was established during Env synthesis and to create an intra-SU bond inside the CXXC motif (Wallin et al. 2004; Li et al. 2008b). This disulfide bond isomerisation is crutial for the fusogenicity of gammaretrovirus (MLV) (Fenouillet et al. 2008). It can be interesting to mention that HIV-1 bond can be reconstituted after recreating the motif in SU and TM (Binley et al. 2000).

4.2.1.2 Cellular Localization of Env Glycoprotein and Viral Assembly

Complex multilevel interactions have been described between Env, Gag and sorting proteins involved in traffic of molecules or vesicles inside the cells. These proteins are involved in both Env trafficking and virus budding. To limit the quantity of Env at the cell surface, the Env undergoes endocytosis and is trafficked in endosomal pathways. These cell localizations are driven by traffic peptidic motifs, like for cell proteins, that have been characterized to direct cellular transmembrane proteins into different endosomal compartment (Bonifacino and Traub 2003). Lentiviruses, including HIV-1, are unusual in having transmembrane glycoproteins with much longer cyt intracytoplasmic tail (150 amino acids) than most other retroviruses (20–50 amino acids) (Hunter and Swanstrom 1990), suggesting that this domain has one or more functions specific to lentivirus replication or persistence. Two groups of motifs have been identified in HIV-1 cyt. The first group consists of three structurally conserved amphipathic alpha-helical domains, designated as lentivirus lytic peptides 1, 2, and 3 (LLP-1, LLP-2, and LLP-3) (Xu et al. 2006). LLP domains have been implicated in various functions, including Env cell surface expression, Env fusogenicity, and Env incorporation into virus particles (Piller et al. 2000). Several studies have suggested that Env is incorporated into virions *via* interactions between LLP and the matrix region of Gag. The second group of motifs regulates the intracellular trafficking of Env. At steady state, Env is predominantly located in the trans-Golgi network (TGN) (Takeda et al. 2003). This intracellular distribution results from dynamic cycling of Env between the cell surface, the endosomal compartment, and the TGN. Newly synthesized Env proteins undergo endocytosis after their arrival at the cell surface. Env internalization is mediated by the interaction of Y712SPL (YXXΦ on prototype, Fig. 4.1a), a membrane-proximal tyrosine-based signal in the gp41 cyt tail, with the adaptor protein (AP) complexes of the cellular sorting machinery, involving the clathrin adaptor AP-2 in particular (Berlioz-Torrent et al. 1999). The cytoplasmic tail of many other retroviruses also contains a YXXΦ motif, including gammaretrovirus like MLV (Song and Hunter 2003), RD114 (Sandrin and Cosset 2006), HTLV-1 (Berlioz-Torrent et al. 1999), M-PMV (Song and Hunter 2003; Song et al. 2005) and BLV (Inabe et al. 1999; Novakovic et al. 2004). The gp41 cyt also interacts with the TGN- and endosome-based clathrin-associated adaptor AP-1 *via* a dileucine motif (Berlioz-Torrent et al. 1999; Wyss et al. 2001) which induces its cellular retention. Some dileucine motifs

are also present in the cytoplasmique tail of MLV, RD114, M-PMV or BLV (Sandrin and Cosset 2006; Grange et al. 2000). Finally, Y802W803, a diaromatic motif involved in the retrograde transport of Env to the TGN, was also identified in the gp41 cyt (Takeda et al. 2003) and interacts with TIP47, a protein required for the retrograde transport from cell surface to the TGN in link with matrix (MA) interaction (Lopez-Verges et al. 2006). A retrograde acid cluster motif has also been identified in RD114 Env (Bouard et al. 2007) that induces a retrograde transport of the Env from the late endosome *via* interaction with PACS1 complexes. It should be note that the palmitoylated cysteines located in the tm are also involved in cell distribution of Env by contributing to the association with lipid rafts (Bhattacharya et al. 2004). These rafts seem to serve as platforms for virus assembly and budding (Suomalainen 2002). For example, membranes of HIV-1 virions have a higher cholesterol rate than the original infected cell, and HIV-1 virions produced in cells with synthesis defects in sphingolipid and cholesterol are less infectious (Brugger et al. 2006).

If the cellular distribution of Env is linked to interaction with trafficking cell molecules, there is evidence indicating that an interaction between the TM cytoplasmic tail and the MA domain of the viral Gag polyprotein mediates Env packaging into particles. Gag structural HIV-1 polyprotein precursor consists of MA, CA (capsid), NC (nucleocapsid), and p6 proteins. The budding ability of retroviruses requires only Gag proteins. Indeed, Gag expression in the absence of other retroviral proteins is sufficient to the liberation of Env-free virions, but these pseudo virions are released independently of the cellular poles. However, when Env is co-expressed, the budding is restricted to the basolateral membrane and mutations of the cyt tyrosine in the membrane-proximal tyrosine-based signal Y712SPL disturb the polarized release of HIV-1 in polarized epithelial cellular models (Lodge et al. 1997; Owens et al. 1991). It is unclear how viral RNA, Gag and Env proteins reach the same site of the cellular plasmic membrane for perfect assembling and budding of the virion. It should be mention that HIV-1 and SIV budding are not only polarized in epithelial cells. In lymphocytes, the release of viruses is restricted to domain of contact between two cells or even between the cell and the culture plate (Bugelski et al. 1995; Pearce-Pratt et al. 1994). This is a budding in a virologic synapse (Morita and Sundquist 2004) and it is also driven by Env cyt motif. This polarisation of budding might have a physiological importance for the cell–cell transmission of the virus. In the case of MLV virus, a model has been developed that allows a better comprehension of Gag and viral RNA trafficking. It was shown that recruitment of glycoproteins by the gammaretroviral core proteins takes place in the intracellular compartments and not at the cell surface. Moreover, gammaretroviral core proteins could relocalize Env glycoproteins in late endosomes and could allow incorporation on viral particles (Sandrin et al. 2004; Bouard et al. 2007). Finally, it was proposed that the retrovirus budding depends on the cell types but might depend of the status of the infection and condition of the cells. In T-lymphocytes, it was initially described that the assembly and budding takes places at the plasma membrane (Barre-Sinoussi et al. 1983; Gelderblom et al. 1987), but recent reports indicated also an assembly in intracellular vesicle containing virus (Grigorov et al. 2006; Joshi et al. 2009). Similarly, in macrophages, HIV-1 assemble and bud in MVB

(multivesicular bodies) (Nydegger et al. 2003; Sherer et al. 2003; von Schwedler et al. 2003) and the viruses are then released outside the cell by fusion of the MVB with plasma membrane (Trojan Horse hypothesis). This is in agreement with the suggestion that retroviruses exploit a cell-encoded pathway of intercellular vesicle traffic, exosome exchange, for both the biogenesis of retroviral particles and a low-efficiency mode of infection (Gould et al. 2003; Fang et al. 2007). However, it was recently proposed that the vesicle containing viruses might have different genesis with some vesicles coming from the plasma membrane invaginations (Welsch et al. 2007). The reason for this discrepancies are not clear but involved Gag interaction with membrane, ESCRT (endosomal sorting complex required for transport) localizations, interferon induced proteins and lipidic composition of microdomain (Ono et al. 2004).

4.2.1.3 Fusion Competency

Gammaretrovirus virions assemble and bud from the infected cells as immature particles that must undergo an additional proteolytic maturation to become infectious (Brody et al. 1992; Christodoulopoulos and Cannon 2001; Green et al. 1981; Rein et al. 1994). This maturation concerns the viral protease dependent cleavage of the so-called R peptide at the C-terminus of the cytoplasmic tail (Green et al. 1981; Rein et al. 1994) (see location on Fig. 4.1a). The R peptide inhibits the fusion, and different hypotheses have been proposed. Firstly, the R peptide contains the YXXϕ internalization motif and the removal of this motif following the cleavage of the R peptide might induce higher amount of envelope at the surface membrane and consequently more fusion (Song and Hunter 2003). Secondly, another explanation is that following the R peptide cleavage, the remaining cyt tail forms a membrane-embedded amphiphilic alpha-helix domain destabilizing the membrane (Zhao et al. 1998; Rozenberg-Adler et al. 2008). Thirdly, it was proposed that, as the R peptide contains a palmitoylation site, its removal induces the close trimerization of the cyt tail and drastic conformational changes in the ectodomain of Env (Aguilar et al. 2003) which might influence Env fusogenicity by destabilizing the SU-TM complexes. These conformational changes are necessary for the isomerisation of the SU-TM disulfide MLV Env (Loving et al. 2008). The R peptide cleavage is the last step leading to a fusion competent infectious MLV retrovirus but this final modification does not exist in lentiviruses which harbour a long cytoplasmic tail. However, studies indicate that artificial (HTLV, HIV or SIV) or natural (SIV) shortening of the cytoplasmic tails change the conformation of ectodomain (Edwards et al. 2002; Spies et al. 1994) and increase the fusogenicity of the Env in cell–cell fusion (Kim et al. 2003; Edwards et al. 2002; Spies et al. 1994).

4.2.2 Virus-Host Cell Membrane Fusion: A Multistep Mechanism

Glycoproteins from enveloped viruses evolved to combine two main features. They have the capacity to bind with a specific cellular receptor and they harbour a fusion domain (peptide fusion and transmembrane domain) that can be activated to mediate the merging (fusion) of viral and cellular membranes.

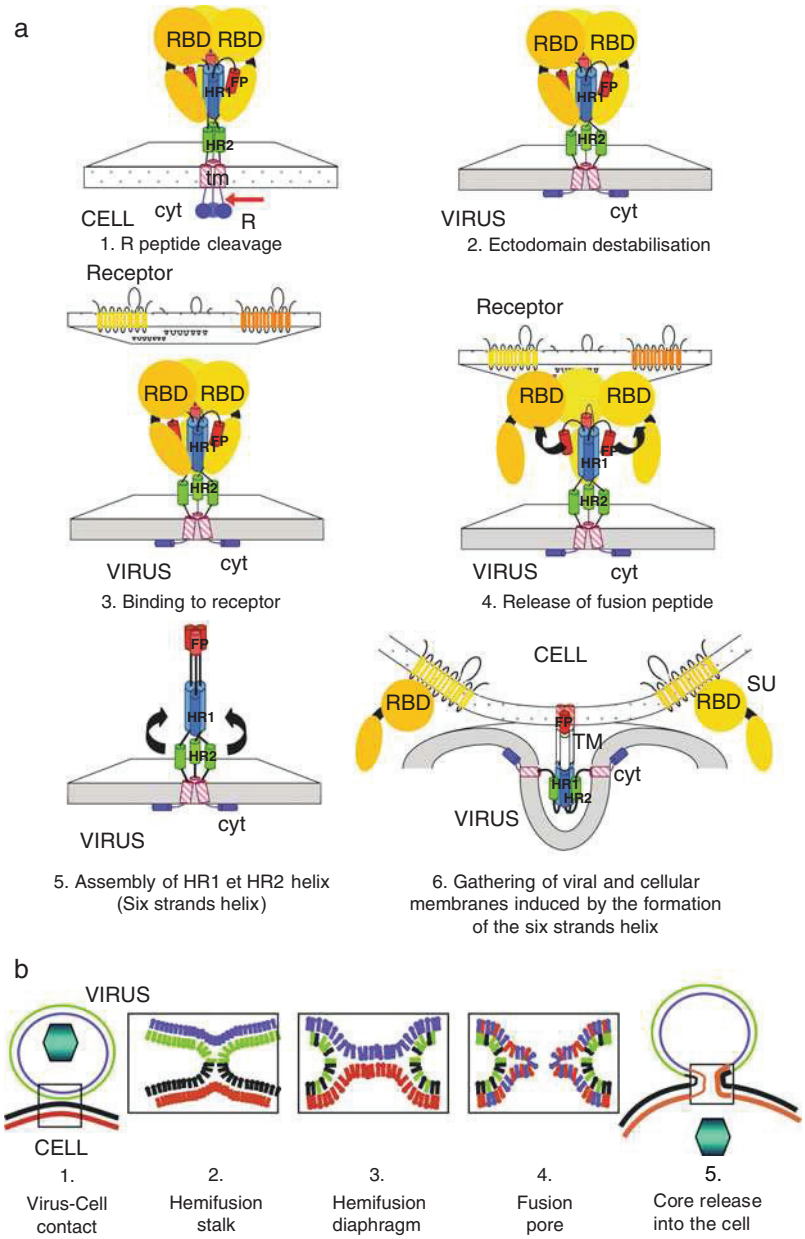


Fig. 4.2 Virus-host cell membrane fusion. **a** Schematic representation of six prototype stages beginning with the fusion competency acquisition of the envelope glycoprotein (1) based on R peptide release by viral protease and ending with the gathering of viral and cellular membranes (6) induced by the anchorage of the fusion peptide into the cell membrane. *Red arrow* symbolizes the R peptide cleavage. **b** Schematic drawing of the successive steps leading to lipidic pore formation. (1) proximal leaflets of viral (*green*) and cell (*black*) membranes coming into immediate contact,

Three different classes of viral fusion proteins have been identified to date based on key structural features at pre- and post-fusion stages. Many studies mentioned that the structural transition from a pre- to a post-fusion conformation leads to a stable hairpin conformation. This concerns the class I fusion proteins, characterized by trimers of hairpins containing a central α-helical coiled-coil structure, and the class II fusion proteins characterized by trimers of hairpins composed of beta sheets structures. A third class of fusion proteins has been described recently, that also forms trimers of hairpins by combining the two structural elements alpha-helix and beta-sheet structures (Weissenhorn et al. 2007).

Three main steps are described for achieving the pre- to post conformational changes. The first one, after Env activation upon receptor binding or acidification of the endosomal compartment, exposes the fusion peptide that is projected toward the top of the glycoprotein, allowing the initial interaction with the target membrane (Fig. 4.2a, drawings 1–4). The second one is the folding back of the C-terminal region onto a trimeric N-terminal region (Fig. 4.2a, drawings 5–6) that leads to the formation of a post-fusion protein structure with the outer regions zipped up against the inner trimeric core in an antiparallel coiled coil structure. The final and third step also requires further refolding of the membrane proximal and transmembrane regions in order to obtain a full-length post-fusion structure where both membrane anchors are present in the same membrane (Fig. 4.2b).

4.2.2.1 Receptor Binding and Peptide Fusion Liberation

HIV-1 fusion is mediated by specific interaction of the viral envelope glycoprotein with the cell surface CD4 molecule that serves as the primary receptor, and additionally a chemokine receptor CCR5 or CXCR4 as HIV-1 co-receptors. Both receptors and the co-receptor binding sites are on gp120 although the membrane fusion is triggered by conformational changes in the transmembrane protein gp41. The viral entry can be blocked by three categories of agents (Qian et al. 2009) (i) attachment inhibitors/antagonists targeting CD4, CCR5 and CXCR4 (ii) inhibition of the post-binding conformational changes, (iii) fusion inhibition. During fusion process, heptad repeats HR1 and HR2 form a six helix bundle structure. Synthetic peptides based on the HR1 and HR2 sequences of gp41 have anti-HIV-1 properties; this is up to date the most successful HIV-1 entry inhibitors class.

Receptors of type C and D retroviruses are cell membrane anchored proteins that transport small molecules. The receptor of ectopic MLV type C retroviruses is CAT-1, a cationic amino acid (like lysine or arginine) transporter (Kavanaugh et al. 1994). The receptor of amphotrophic MLV is Pit-2 and the receptor of MLV-10A1 is Pit-1. Pits-1 is also the receptor for GALV and FeLV. Pit-1 and Pit-2 are two inorganic phosphate transporters. Nevertheless, in some cases, these viruses are able

Fig. 4.2 (continued) (2) hemifusion stalk with proximal leaflets fused and unfused viral (*blue*) and cell (*red*) distal leaflets, (3) unfused stalk expansion leading to the hemifusion diaphragm, (4) fusion pore forming the hemifusion diaphragm bilayer, (5) core release into the cell

to recognize the common parts of Pit-1 and Pit-2 receptors and to infect human cells by using one or the other indifferently (Tailor et al. 2000). Two receptors have been identified for the type D retroviruses interference group: ASCT1 and ASCT2, two neutral amino acid transporters and receptors for BaEV, SRV and RD114 viruses. Retroviruses of the RD114 family recognize an hypervariable region in the second bundle of ASCT2 receptor. Most of the time the viruses use the part of the transporter that is involved in the transport function of this molecule: this region is under strong selection pressure to keep the function and therefore the polymorphism is limited and the infectivity, i.e. the recognition of the receptor by the viruses, is conserved. Furthermore these type D retroviruses can use either ASCT1 or ASCT2 by recognizing the conserved domain.

Two types of fusion mechanisms can occur, namely pH independent and pH-dependent. In the first case, the recognition between virus and receptor directly triggers conformational changes in the glycoprotein that leads to the direct fusion between the two membranes (viral and plasma) and to the liberation of the viral genetic material. The activation of Env at neutral pH allows the fusion in vitro and in vivo of Env-expressing cells co-cultured with receptor-expressing cells. The fusion leads to the merging of cytoplasms and to the generation of multinucleated cells named syncytia. In the second case, for pH-dependent fusion, the interaction between the Env and the receptor is followed by an endocytosis of the virus-receptor complex before the acidification of the endosome triggers conformational changes in the glycoprotein. For the pH dependent virus, such a fusion can be reproduced in vitro in cell culture or in liposome-virus fusion assay after decreasing the pH in the test tube, but cannot occur in vivo. Most retroviruses use a pH-independent fusion mechanism, with a few exceptions for MMTV, JSRV and ASLV. The proposed mechanism for ASLV virions is an intermediate since entry occurs in two steps, beginning with a receptor-priming that in turn induces Env conformational changes allowing the Env to become sensitive to the low pH. This hybrid mechanism does not lead to cell–cell fusion in vivo. JSRV also uses a receptor-priming for fusion activation of Env at low pH but the mechanism is slightly different that for ASLV (Cote et al. 2009) and requires dynamin-associated endocytosis (Bertrand et al. 2008). MMTV is so far considered as a classical pH-dependent virus that uses mouse transferrin receptor 1 (TfR1) and trafficking to a low pH compartment (Wang et al. 2008).

Finally, it should be note that viruses that use a pH independent mechanism of activation of Env may still enter the cell by endocytosis without any requirement for acidification activation of Env into the endosomes. So far, there are many different endocytosis pathways that have been described (Marsh and Helenius 2006; Mercer and Helenius 2009) as being used by both pH-dependant and pH-independent viruses. However, re-investigations of the entry pathways are clearly needed for many pH-independent viruses that were thought not to rely on endocytosis. For example, Nipah paramyxoviruses that can fuse cells at neutral pH seem to use macropinocytosis for entry (Pernet et al. 2009). Macropinocytosis and phagocytosis have also been proposed for HIV-1 entry even if it is unsure that this entry can lead to productive infection (Marechal et al. 2001; Trujillo et al. 2007).

Let's describe the several critical domains which are directly involved in the fusion process. For the most part, in retrovirus-cell fusion, a fusion unit typically contains a unique transmembrane domain and a fusion peptide – a sequence of 10–30 residues that forms an amphiphilic domain usually at the N-terminus of the TM (Chernomordik and Kozlov 2003). The hydrophobic fusion peptide domain is sequestered in all previously described Env biosynthesis steps. The final acquisition of the fusion state competency is triggered by the receptor binding alone and/or a low pH surrounding the endosomes and globular head domains dissociation. This movement allows a loop-to-helix transition of a polypeptide segment of TM that was previously buried underneath the SU heads, projecting the fusion peptide ~100 Å towards the target membrane, where it inserts irreversibly. In the case of class I fusion proteins like retroviruses, this occurs by a "spring-loaded" mechanism. This initial change is proposed to result in a "pre-hairpin intermediate", an extended structure that is anchored both in the target membrane by the fusion peptide and in the virus membrane by the tm segment (Fig. 4.2a drawings 5 and 6). The HR2 C-terminal end of the long TM α-helix jackknifes back, reversing the direction of the viral-membrane-proximal segment of TM, which then interacts in an anti-parallel fashion with the groove formed by the N-terminal HR1 trimeric coiled coil. The final post-fusion conformation of TM is therefore a highly stable rod with the TM and fusion-peptide segments together at the same end of the molecule, a structure termed a "trimer of hairpins". The hairpin structure brings the two membranes proximal and provides free energy to overcome the barrier of membrane merging (Melikyan 2008). Membrane fusion occurs, which leads to pore formation and release of the viral genome into the cytoplasm.

In addition, compare to cellular glycoproteins, the retroviral TM ectodomain also contains a hydrophobic domain abnormally enriched in tryptophane in the juxtamembrane domain (Salzwedel et al. 1999; Suarez et al. 2000). This domain contributes to the conformational change and membrane destabilization during the fusion process of HIV-1 (Munoz-Barroso et al. 1999), and antibodies (Lorizate et al. 2006; Purtscher et al. 1994) or peptides (Moreno et al. 2006) directed against this domain inhibit the entry. This juxtamembrane domain is also critical for fusion of many envelope viruses beside retroviruses, including paramyxoviruses and coranaviruses.

4.2.2.2 Pore Formation and Fusion of the Target Membranes

The hypothesis of the pore model in viral membrane fusion mechanism (Fig. 4.2b) is supported by experimental results. The first evidence for a hemifusion intermediate was achieved by studying influenza virus entry that occurs after the hemagglutinin glycoprotein binding to the host cell. The substitution of the hemagglutinin transmembrane domain by a glycosylphosphatidylinositol (GPI) revealed the importance of the transmembrane region for the fusion pore opening and expansion. Hemifusion structures are connections between outer leaflets of apposed membranes, whereas the inner leaflets remain distinct. This is a transient structure that either dissociates or gives rise to the fusion pore (Chernomordik and Kozlov 2008). Interestingly, the

helix breaker residues within the tm domain are critical for the hemifusion and pore opening step of the fusion process mediated by different retroviral Env, like HIV-1 (Owens et al. 1994) and Mo-MLV (Taylor and Sanders 1999). In addition, a hemifusion intermediate has been detected in the case of HIV-1 Env-mediated fusion (Munoz-Barroso et al. 1998) by using peptide inhibitors that target a pre-fusion or prehairpin structure such as HIV-1 gp41 T-20. Then, the pore is formed and it allows a connection between two compartments initially separated by the apposed membranes.

The membrane ability to hemifuse and develop fusion pore has been found to depend on the lipid microdomain composition, e.g. cholesterol (Chernomordik and Kozlov 2003). A potential lipid dependence of virus entry processes has been first deducted from experiments on influenza virus suggesting the implication of lipid rafts (Takeda et al. 2003). For retroviruses, the tm palmitoylations which contribute to the Env localization in raft domains (Li et al. 2002) influence indirectly the fusion process (Gebhardt et al. 1984; Ochsenbauer-Jambor et al. 2001). As an alternative to lipidic pore hypothesis, a direct fusion has also been proposed. The fusion pore is a full proteic channel-like structure dependent only on the transmembrane domains of the glycoproteins. In this model, the pore is opened by the joining of two hemipores located on each membrane (Chernomordik and Kozlov 2008; Chernomordik and Kozlov 2005). After fusion pore opening and enlargement (Melikyan et al. 2005), the genetic material enters the cytoplasm of the cell and enters the nucleus.

4.2.3 Rous Meets Mendel

In humans, virus-like particles without trivial evidence of inter-individual transmissibility were identified in disparate contexts such as placenta, autoimmune diseases, e.g. body fluids of multiple sclerosis patients, and cancers, e.g. seminomas, lymphomas or plasma of breast cancer patients. In the seventies, numerous electron microscopic studies have described the presence of virus related particles in placental chorionic villous tissues of humans (and primates). Further studies then revealed some retroviral characteristics of these particles such as ultrastructural features and RT activity (Lyden et al. 1994). In addition, retroviral envelopes were detected in placenta sections by immunohistochemical techniques in human (Venables et al. 1995) and in baboon (Langat et al. 1999).

Retrovirus-like particles associated with reverse transcriptase (RT) activity have been described by several groups in cell cultures from patients with multiple sclerosis (MS) (Perron et al. 1989; Haahr et al. 1994). Infectious properties of these particles are at least not trivial to ascertain if not doubtful. However, using PCR techniques, a reconstructed retroviral genome was defined as Multiple Sclerosis associated RetroVirus (MSRV) (Perron et al. 1997; Komurian-Pradel et al. 1999). MSRV is closely related to the HERV-W (Human Endogenous RetroVirus) family and particularly the ERVWE1/Syncytin-1 locus (Blond et al. 1999). Though, no full length replication competent virus has been experimentally isolated (Voisset et al. 1999). Nevertheless, it has been demonstrated that MSRV particles cause

T lymphocyte-dependent death with brain haemorrhage in humanized SCID mice (Firouzi et al. 2003). MSRV envelope protein has been proposed to exert various immune properties, e.g. as triggering a superantigen effect (Perron et al. 2001) and activating innate immunity (Rolland et al. 2006). Note that HERV-H related elements were also associated with particles observed in MS (Christensen et al. 1998), dually found with viruses from the herpes group, most likely Epstein Barr Virus (EBV).

Cancers of the male reproductive system appears to be a favourable context for virus-like particles detection. Thus, HTDV (Human Teratocarcinoma-Derived Virus) is only expressed in the male germ line tumor context (Boller et al. 1993), and similar particles were observed in testicular germ cell tumors (or seminomas) (Herbst et al. 1999). In both situations, HERV-K transcripts could be associated with the particles. By electron microscopy and immunogold staining, HERV-K like particles were also visualized in the plasma of individuals with lymphoma, but these particles seem to be defective, as surface spikes and free mature virus particles were never observed (Contreras-Galindo et al. 2008).

In all the situations exemplified above, although nucleic acid material could be associated with the particles, it remains unclear whether such particles could result from expression of a single retroviral loci, a trans complementation process or even more complex phenomenon involving genetic material recombination. As a corollary, infectivity of these particles has not been demonstrated.

A clearer view was expected from the publication of several mammalian genomes, including human (Lander et al. 2001) and mouse (Waterston et al. 2002) genomes. Genomes of mammalian species harbor a large amount of retrovirus-like sequences. These endogenous retroviruses (ERVs) are remnants of ancient retroviral infections that initially occurred in the host germline. Throughout evolutionary time, these initially stably-integrated sequences have derived into gene families by retrotransposition events, and have accumulated genetic defects as a consequence of the host domestication. This general drift basically resulted in gene silencing. Generally, the retro-elements are free-Env and are not able to dissemination between cells. Intriguingly, the human genome but also the mouse genome contains a huge amount of endogenous retroviruses, reaching 8.5 and 9% of these genomes sizes, respectively.

Deciphering the human genome showed that the HERV-K family contains tens of almost complete but mutated proviruses that allow the expression of viral proteins which appears able to form retroviral particles. However, no complete proviruses able to produce replication competent and infectious viral particles have been detected. The HERV-K113 locus though to be the more recent element of the family and that contains intact ORFs for all the viral proteins does not produce any particles (Lee and Bieniasz 2007). Trans-complementation between different HERV-K(HML2) proviruses could theoretically produce infectious particles, although not demonstrated to date. Interestingly, the infectious potential of HERV-K particles was artificially restored by generating a consensus HERV-K (HML-2) provirus named Phoenix supposed to be the HERV-K family progenitor (Dewannieux et al. 2006); this consensus contained at least 20 amino acid changes

on the overall sequence as compared to individual proximal HERV-K loci. By electronic microscopy, this resurrected HERV-K forms viral particles in transfected cells. The budding of its particles is similar to γ-, δ-retroviruses or lentiviruses with no particles preassembling into the cytoplasm.

As cited earlier, MSRV is closely related to the HERV-W family including the Syncytin-1 encoding ERVWE1 locus which is the only W-locus bearing a full-length envelope. The sequencing of ERVWE1 envelope confirmed that the MSRV envelope was not encoded by the ERVWE1 locus (Mallet et al. 2004). It was thus proposed that MSRV particles (if not derived from an as yet uncharacterised exogenous retrovirus) may result from transcomplementation of dispersed HERV-W copies simultaneously activated (Dolei 2005), what appears poorly probable as regards to the HERV-W elements identified in the human genome. However, it could not be formally excluded that MSRV/HERV-W genome (associated with particles) may result from very complex recombination events involving several loci on distinct chromosomes (Laufer et al. 2009).

Although complete genomes analyses did not clearly explained the mechanisms leading to the formation of endogenous retroviral particle, they uncovered the extreme plasticity of these retroviral elements. Koala retrovirus (KoRV) provides a unique opportunity to study the process of ongoing endogenisation as it still appears to be spreading through the koala population. Interestingly, infectious viral particles are produced by the endogenous form of KoRV and high levels of viraemia have been linked to neoplasia and immunosuppression (Tarlinton et al. 2008). It remains unclear how the host can react when exogenous and endogenous forms of a virus are coexisting within the genome and his environment. Studies on Koala might answer this question. Interplay between the primitive virus world and the evaluated eukaryotic one could be observed at the env level. Thus, infectious retroviruses appear to have burst from our far ancestors genome by transcomplementation of cellular retrotransposons with viral envelopes genes (Malik et al. 2000). Another type of capture exists between retroviruses of distant species, consisting in the swapping of envelopes observed for species in the same environment or linked by the food chain. For example, the RD114 virus comes from two genetic recombinations resulting in two env-captures. First, the SERV (simian endogenous retrovirus) env was captured by the PcRV (Papio cynocephalus retrovirus) leading to the BaEV (baboon endogenous retrovirus). Second, the acquisition of BaEV env by FcEV (felis catus endogenous retrovirus) led to the emergence of RD114 virus (Kim et al. 2004). Last, endogenous retroviruses as remnants of ancient retroviral infections that initially occurred in the host germline represent an intriguing heritage. More precisely, as a consequence of at least 30 distinct chapters of retroviral infection during the past 90 million years, the current human genome contains 18 coding envelope genes (de Parseval et al. 2003; Blaise et al. 2005) (Table 4.1). The most represented family is the HERV-K(HML2) family which contains six coding env genes lacking fusogenic activity. Three Env proteins belonging to HERV-W, HERV-FRD and HERV-P families, namely Syncytin-1, Syncytin-2 and EnvP(b) respectively, have fusogenic properties.

Table 4.1 Human viral heritage of envelopes open reading frames containing canonical retroviral motives

Name	RNA expression	Protein expression	Fusogenity	References
*env*H1	NE[c]	NE[c]	ND[d]	Lindeskog et al. (1999)
*env*H2	NE[c]	NE[c]	ND[d]	de Parseval et al. (2001)
*env*H3	NE[c]	NE[c]	ND[d]	de Parseval et al. (2001)
*env*K1	NE[c]	NE[c]	ND[d]	de Parseval et al. (2003)
*env*K2	NE[c]	NE[c]	ND[d]	Barbulescu et al. (1999)
*env*K3	NE[c]	NE[c]	ND[d]	Donner et al. (1999)
*env*K4	NE[c]	NE[c]	ND[d]	Barbulescu et al. (1999)
*env*K5	NE[c]	NE[c]	ND[d]	Turner et al. (2001)
*env*K6	NE[c]	NE[c]	ND[d]	Turner et al. (2001)
*env*T	NE[c]	NE[c]	ND[d]	de Parseval et al. (2003)
*env*W[a]	Placenta	Placenta	Yes[e]	Blond et al. (1999)
*env*FRD[b]	Placenta	Placenta	Yes[f]	de Parseval et al. (2003)
*env*R	All tissue	ND[d]	ND[d]	Cohen et al. (1985)
*env*R(b)	NE[c]	NE[c]	ND[d]	de Parseval et al. (2003)
*env*F(c)2	NE[c]	NE[c]	ND[d]	de Parseval et al. (2003)
*env*F(c)1	NE[c]	NE[c]	ND[d]	de Parseval et al. (2003)
*env*V	Placenta	ND[d]	No[e]	Blaise et al. (2005)
*env*P(b)	All tissue	ND[d]	Yes[e]	Blaise et al. (2005)

[a]Syncytin-1.
[b]Syncytin-2.
[c]No expression.
[d]No determined.
[e]In vitro.
[f]In vivo.

4.3 Syncytins and Cell–Cell Fusion

In spite of ERVs have been thought to be a non-functional part of the genome for a while, the past 10 years identified open reading frames of envelope genes in human, mouse, rabbit and sheep genomes (Fig. 4.3a), and associated with transcription activity and fusogenic glycoproteins synthesis (Fig. 4.3b) likely involved in biological functions. This is the case for the two human envelopes genes ERVWE1/Syncytin-1 and ERVFRDE1/Syncytin-2, located on chromosome 7q21.2 and 6p24.1, respectively, as well as for the two Syncytins-related A and B in mice, both pairs associated with fusion steps occurring in placental development process. Recently the novel Syncytin-Ory1 was identified in rabbit given a new example of syncytin gene within a third order of mammals (Heidmann et al. 2009). The ovine species also provide a quite interesting model of endogenisation process since the exogenous and pathogenic Jaagsiekte Sheep Retrovirus (JSRV) coexists with at least 27 highly related endogenous counterparts (enJSRVs), accounting for envelope genes in the ovine genomic DNA with evidences for open reading frames (Arnaud et al. 2007). enJSRVs play a crucial role in the sheep placental morphogenesis and

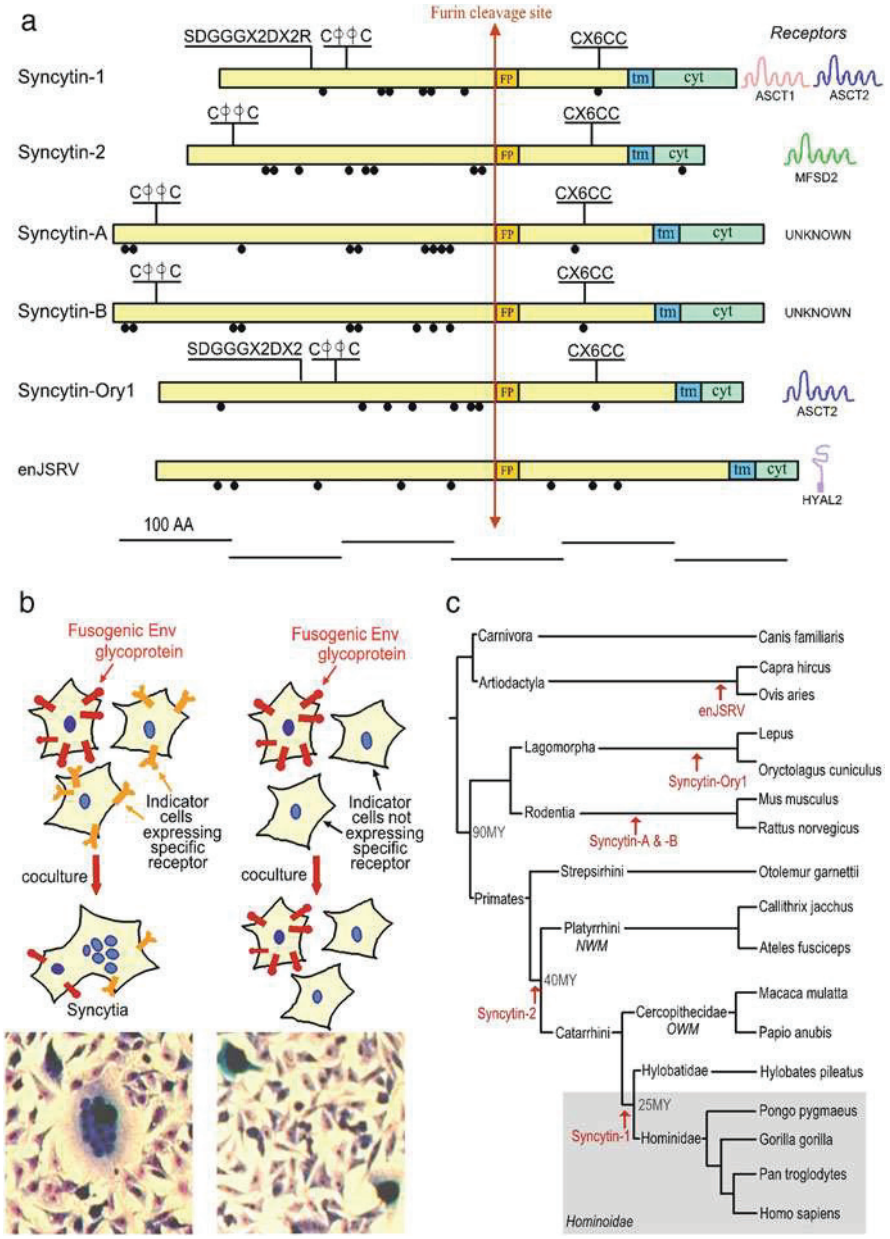


Fig. 4.3 Structure, phylogeny and fusion capacities of Syncytins. **a** Envelopes structures of Syncytins and schematic representation of their cognate receptors. FP: fusion peptide; tm: transmembrane domain; cyt: cytoplasmic tail. *Black dots* indicate the predicted N-glycosylation sites. SDGGGX2DX2R, consensus motif conserved in type D retroviral interference group, is indicated in human Syncytin-1 and rabbit Syncytin-Ory1. **b** Demonstration of Syncytin-1 cell–cell fusion property. TELacZ cells (*dark blue* nucleus) expressing Syncytin-1 envelope glycoprotein

their envelope expression in the reproductive tract is mandatory for a successful pregnancy (Dunlap et al. 2006).

The focus point in this chapter will be now to discuss if these endogenous viral proteins of the genome still remain fusogenic in the same way an exogenous retrovirus envelope glycoprotein is (including transcription strategy, maturation steps, receptor recognition and fusion process), and to address the question of the domestication by the host and adaptive response though integrative and evolutionary points of view.

4.3.1 Integration, Domestication Steps and Biological Functions of Endogenous Viral Glycoproteins

4.3.1.1 Integration Dating and Orthologues

One starting point in the discussion about the endogenous envelopes found in genomes could be the estimation of the age of the proviruses. This can basically be done by two approaches, bringing additional informations. One way is to assess a phylogenic lineage by tracing the presence of a similar DNA sequence at the same genomic loci in the genome of different species, and to conclude by a unique hypothetical integration event into the germline of a common ancestor. Another way is to consider the divergence between the 5′ and 3′ LTR and assuming a molecular clock is acting randomly through the genome, to generate variations over time between two originally and identical provirus sequences (the 5′LTR and 3′LTR being identical at the time of integration in the host genome).

The first striking point are the unshared properties of both families and integration times within the humans, mice, rabbits and sheeps endogenous envelopes (Fig. 4.3c). ERV-W elements have been identified in *hominoidae* (human, chimpanzees, gorillas, orangutans and gibbons) and *Cercopithecidae* (old world monkeys) (Kim et al. 1999; Voisset et al. 1999) indicating that what we call today the human ERV-W family, HERV-W, derived from an ancestral virus which entered the genome after the divergence between *Catarrhini* and *Platyrrhini*, i.e. less that 40 million years ago (MYA). The ERVWE1/Syncytin-1 locus results from a complete proviral retrotransposition event into the germ line of an ancestor before *Hominoidae* and *Cercopithecidae* divergence more than 19–25 MYA (Caceres and Thomas 2006; Bonnaud et al. 2005). A full length envelope ORF corresponding to functional envelope glycoprotein was preserved in *Hominoidae* but genetic drift led to truncated envelope genes in old world monkeys. In contrast, the FRD family containing the HERV-FRD envelope Syncytin-2 is found in all simians, from New

Fig. 4.3 (continued) co-cultured with indicator XC cells (*light blue* nucleus) expressing hASCT2 receptor generates multinucleated large syncytia (*left part*). TELacZ-Syncytin-1 cells co-cultured with XC cells lacking hASCT2 do not fuse (*right part*). **c** Phylogenetic tree depicting the conservation among species of the six envelope-open reading frames harbouring retroviral canonical motifs (branches of the tree are only illustrative). NWM: new world monkeys; OWN: new world monkeys

World monkeys to human, suggesting a divergence split at least 40 MYA (Blaise et al. 2003). Moreover, the human, mouse and rabbit genes are not orthologs since Syncytin-A and -B entered into the rodent lineage before speciation of *Muridae* about 20 MYA (Dupressoir et al. 2005), i.e. after the speciation of rodents and primates while the Syncytin-Ory1 integration took place before the divergence between *Lepus* and *Oryctolagus/Sylvilagus* around 12 MYA (Heidmann et al. 2009). The situation is also different in domesticated sheep (Ovis aries) and other species within the *Caprinae* subfamily, where the endogenous retroviruses JSRV (enJSRVs) start to invade the genome at least 5–7 MYA and are likely still colonizing it today as given evidences by the restricted presence of recent enJSRVs loci into the genome of only some breeds or even some animals of the same breed of domesticated sheep (Arnaud et al. 2007).

In order to better understand which mechanisms may have led to a positive selection of organisms harboring embarked viral genes, many arguments in favor of a domestication scenario have been deployed, especially about the ERVWE1/Syncytin-1 locus. Indeed, the proposed evolutionary pathway occurring in *Hominoidae* is opposed to the genetic drift in *Cercopithecus*. A ~4.3 kb region, comprising the HERV-W 5′LTR-*gag-pol* fraction, was deleted in *Cercopithecus* and was followed by a genetic drift of the Env/Syncytin-1 ORF (Bonnaud et al. 2005; Caceres and Thomas 2006). Remarkably, the Syncytin-1 ORF has been conserved in all *Hominoidae*, while *gag* and *pol* regions have accumulated numerous stop and frameshift mutations (Mallet et al. 2004), supporting the idea of a specific preservation. Meanwhile, the analysis of 155 individuals including Caucasians, Asians, Africans, Metis and Ashkenazi people has revealed a positional conservation of the Syncytin-1 locus and the preservation of the envelope ORF (Bonnaud et al. 2004; Mallet et al. 2004), while a close examination of 24 ERVWE1 provirus sequences has showed an unusually low polymorphism in the 5′LTR (1 base per 18 kb as compared to 1 base per 2 kb for coding sequences) (Mallet et al. 2004). An additional specificity of the ERVWE1 provirus is the MaLR-LTR trophoblast specific enhancer (TSE) located upstream the ERVWE1 provirus that is highly conserved with no polymorphism observed in the 48 human sequences analyzed. Although the envelope gene may be considered under selective pressure depending on the part of the gene we focus on, the striking feature for the ERVWE1 locus is a 12 bp deletion observed in the Syncytin-1 intracytoplasmic tail gene region and that constitutes a specific signature of this locus. This deletion is unique among all ERV-W copies in available human and chimpanzee genomes (Bonnaud et al. 2005), and is crucial for the envelope fusogenicity (Bonnaud et al. 2004; Cheynet et al. 2005).

Interestingly, the comparison of the FRD/Syncytin-2 envelopes among simians has also revealed a limited number of mutations, and pseudotypes experiments demonstrated that only one mutation occurring in the transmembrane subunit of the protein can be responsible for the loss of infectivity (Blaise et al. 2004). Notably, the alignment of endogenous and exogenous JSRV envelopes reveals similar deletions in the cytoplasmic tail of enJSRVs env as compared to the exogenous one (Palmarini et al. 2001). Altogether, these elements may infer the hypothesis of a positive selection and domestication of retroviral envelopes.

4.3.1.2 Endogenous Retrovirus Envelopes Are Expressed in the Placenta and in the Testis Suggesting a Direct Involvement in Developmental Process

The HERV-W family was molecularly characterized following the isolation of cDNA clones in the placenta that revealed viral sequences genes expression, especially with similarities to the avian retrovirus primer binding site (Blond et al. 1999). In 2000, protein truncation tests within this endogenous family revealed only one open reading frame (ORF) coding for a putative envelope gene associated with a functional U3 promoter (Voisset et al. 2000). One year later, Blond and Mi concomitantly associated an HERV-W envelope protein with fusion events in TE671 and BeWo cells, and the name Syncytin was proposed by Mi (Blond et al. 2000; Mi et al. 2000). Heidmann and colleagues then conducted a genome wide screening that identified a second envelope protein, belonging to the HERV-FRD family, and expressed exclusively in the human placenta. They named Syncytin-2 this putative new fusogenic Env-FRD protein (Blaise et al. 2003). A similar *in silico* approach was done in the murine genome, identifying the two coding envelopes genes present as unique copies and with a placenta specific expression: Syncytin-A and Syncytin-B (Dupressoir et al. 2005), and recently in the rabbit genome, identifying the Syncytin-Ory1 gene (Heidmann et al. 2009). If the situation is much more different in the ovine genome, where approximately 27 copies of endogenous betaretrovirus (enJSRVs) were detected, RT-PCR and in situ hybridization clearly indicate a conceptus (embryo/fetus and extra embryonic membranes) localization of enJSRVs env transcripts during gestation (Dunlap et al. 2006).

Although human Syncytins were abundantly described in the placental tissues, initial works also mentioned a weaker but significant transcription in the testis without any protein evidence (Mi et al. 2000). Envelope-specific RT-PCR established expression in the human testis of both Syncytin-1 and Syncytin-2 (de Parseval et al. 2003), and a multiplex degenerated PCR screening for a consensus *pol* region has revealed a general expression of the HERV-W family in testis (Pichon et al. 2006) and epididymis (Crowell and Kiessling 2007). This is consistent with old studies that identified the epididymal epithelium as a principle reservoir for retrovirus expression in the mouse (Kiessling et al. 1989).

This knowledge points out that endogenous envelopes expression are usually associated with developmental tissues, and so far raise the question of whether or not Syncytins play a direct role in the mammalian placentation.

4.3.1.3 Biological Function of ERVs Envelopes

The keen interest in ERVs envelopes expressed in placentas is fed by in vivo or ex vivo demonstrations that directly link Syncytins with fusion events during placental development.

Although the role of Syncytin-1 in human placentation awaits a definitive demonstration (e.g. infertility associated mutation), recent knock-out gene experiments in mice clearly achieved this goal in rodent model and demonstrated for the first

time the critical role of Syncytin-A in placenta morphogenesis. Using a homologous recombination strategy, Syncytin-A null mouse embryos exhibited growth retardation with an altered placenta labyrinth architecture and died in utero (Heidmann et al. 2009). This is consistent with previous in vitro works that used specific antibodies and antisense oligonucleotides to show a decrease in syncytia cell formation after Syncytin-A inhibition (Gong et al. 2007). In addition, the endogenous retroviruses of sheep, enJSRVs, play a fundamental role in sheep conceptus growth and trophectoderm differentiation *via* their envelope glycoproteins. Indeed, in vivo experiments using an enJSRV envelopes specific morpholino injection trigger the lost of pregnancy by day 20 after injection (Dunlap et al. 2006).

These kind of in vivo experiments obviously cannot be performed in human. Yet, primary cultures of human villous cytotrophoblasts cells give a unique opportunity

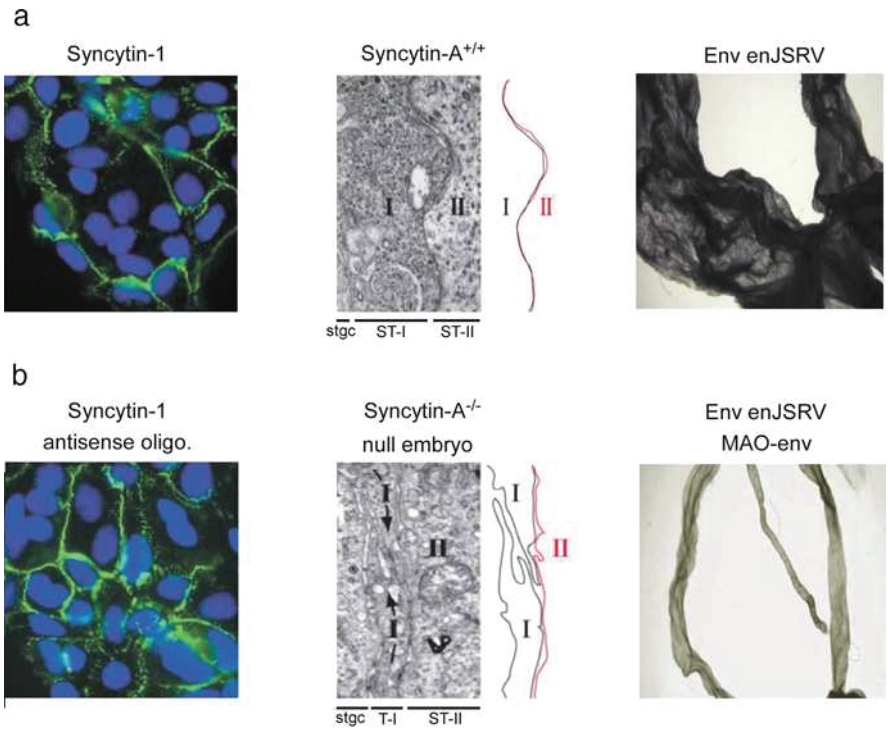


Fig. 4.4 Involvement of Syncytins in placenta development. **a** Assays reporting the biological effect of Syncytins. **b** Ex vivo or in vivo specific inhibition of Syncytins expressions. From *left* to *right*: Syncytin-1-induced human primary trophoblasts fusion and differentiation results in syncytia formation ex vivo (**a**). Inhibition by specific antisense oligonucleotide largely reduces syncytia formation (**b**). Electron micrograph of Syncytin-A$^{+/+}$ mouse placenta shows tight apposition of the syncytiotrophoblast I and II layers (ST-I; ST-II); stgc: sinusoidal trophoblast giant cells (**a**). Syncytin-A$^{-/-}$ null mouse embryo interhemal domains shows unfused trophoblast I cells (T-I) (**b**). Micrograph of the normal development of a sheep conceptus (**a**). Retarded growth of a sheep conceptus recovered after an envelope enJSRV morpholino antisense oligonucleotide (MAO-env) injection (**b**)

to study placenta cells as closely related as possible to tissue environment. Thus, by using specific antisense oligonucleotides and siRNA strategies, expression of Syncytin-1 mRNA and protein as well as the syncytium formation by cell fusion events were dramatically reduced (Frendo et al. 2003b; Vargas et al. 2009). In addition to that, Vargas and colleagues recently compared these results using the same targeting strategy against Syncytin-2, and interestingly showed that Syncytin-2 inhibition in primary cells culture also leads to a decrease in fusion index that is more important than for Syncytin-1 (Vargas et al. 2009). The conclusion is that Syncytin-2 could also be a major determinant of trophoblast cell fusion, and in a coherent vision this underlines there should be more than one ERV envelopes proteins acting upon trophoblast cell fusion in human. Parallel procedures demonstrating the involvement of human, mouse and sheep Syncytins in placenta development are illustrated in Fig. 4.4.

Note that early works identified ERV-3 (HERV-R) envelope as the first candidate for placental functions. The ERV-3 envelope protein is detected specifically in the multinucleated syncytiotrophoblast in vivo (Venables et al. 1995) and ERV-3 Env expression affects proliferation and differentiation of BeWo cells in vitro (Lin et al. 1999; Lin et al. 2000). However, the observation that approximately 1% of the Caucasian population has a mutation in ERV-3 *env* inducing a stop codon, and consequently resulting in a truncated envelope lacking both the fusion peptide and the immunosuppressive domain (de Parseval and Heidmann 1998) has drastically lowered the scientific efforts regarding involvement of ERV-3 in placental development. Indeed, a second hypothetical function of Syncytins is related to their putative immunosuppressive activity (see below) due to the presence of a putative immunosuppressive region conserved among murine, feline, and human retroviruses (Cianciolo et al. 1985).

So far we saw that Syncytins are involved in developmental fusion process. In the next part of this chapter we aim to focus on the mechanistic comparison between exogenous and endogenous envelope glycoproteins at the synthesis and maturation steps.

4.3.2 Fusion Mechanism and Receptor Recognition

4.3.2.1 Maturation

The different steps leading from a brand-new translational product in the cytosol to a functional membrane-anchored envelope glycoprotein has been discussed previously. Basically, endogenous envelopes still remain glycoproteins, engaged in the classical reticulum-golgi apparatus where post translational events occur, before to be address to the plasma membrane and to become functional. Thus, precursor synthesis and glycosylation, disulfide bonds, trimerization, peptide cleavage and the importance of the cytoplasmic tail will be illustrated here introducing specific Syncytins knowledge, in order to support our previous descriptions as well as to focus on endogenous envelope specific characteristics.

Precursor, Furin Cleavage and Glycosylation

Studies in BeWo cells models have described in a fine way the maturation of Syncytin-1. Syncytin-1 is first synthesized as a 73-kDa precursor (gPr73) before to be cleaved at a conserved RNKR furin cleavage site into two subunits, SU (gp50) and TM (gp24) (Cheynet et al. 2006). Although polypeptides size of Syncytin-1 and -2 is the same (538 amino acids), sizes of SU and TM are different after processing (Chen et al. 2008). PNGase F digestion and tunicamicyn treatment predicted and confirmed seven N-glycosylation sites for the Syncytin-1. These results indicate that Syncytin-1 is a moderately glycosylated protein, with one glycosylation site in the TM subunit which is essential for correct envelope protein folding, and with high-mannose N-glycans on the six glycosylation sites of the carboxy-terminal domain of the SU (Cheynet et al. 2006). Furin inhibition experiments conducted on Syncytin-2 have also established the furin to play a major role in the proteolytic cleavage of the HERV-FRD envelope proprotein (Chen et al. 2008), where the cleavage consensus sequence is also found. Interestingly, using knock down experiment, furin has been proposed to have a possible role in promoting trophoblast cell migration and invasion in human placenta (Zhou et al. 2009).

Bioinformatics analyses and sequence alignments suggest that Syncytin-A and -B exhibit most features of membrane fusion proteins, including the conserved cleavage site RNKR, which separate the SU and TM subunits (Dupressoir et al. 2005; Peng et al. 2007). Finally, the same feature is observed for the Syncytin-Ory1 sequence that exhibits a RQKR site (Heidmann et al. 2009) and for the enJSRV sequences that harbour the cleavage furin motif site.

Disulfide Bonds and Trimerization

Considering the Syncytin-1 TM gp24 sequence, it appeared that a leucin zipper-like $LX_6LX_6NX_6LX_6L$ and a CX_6CC motifs are present, suggesting that SU and TM may covalently link together and form homotrimers (Cheynet et al. 2006). Indeed, Syncytin-1 sequence contains a typical disulfure isomerase motif in the SU domain (CΦΦC). As previously mentioned for MLV, the first two cysteines of the CX_6CC motif can form a stable disulfide bond, leaving the third cystein free to form a disulfide bond with the CΦΦC motif (Fass et al. 1997). Mutational experiments using neutral substitution in the CX_6CC motif did not affect the protein precursor expression level, but impaired syncytia formation, suggesting that disulfide bonds likely contribute to the correct folding of the envelope. In accordance to that, Chen and colleagues demonstrated that the disulfide bridge-forming CX_7C motif of the Syncytin-2 was essential for the fusogenic activity (Chen et al. 2008).

Cytoplasmic Tail and R Peptide

The cytoplasmic tail region of numerous retroviral envelopes plays a critical role in the fusion triggering. In the retrovirus life cycle, the presence of an R peptide basically prevents the fusion to occur, notably because of the presence of the YXXΦ

motif described above. During viral packaging, the R peptide is proteolytically cleaved and this event enables envelopes to cause membrane fusion, as described by mutagenesis experiments (Yang and Compans 1996). We illustrate how Syncytins used various strategies that diverge from envelopes of infectious retroviruses to adapt to their physiological functions.

Surprisingly, sequences comparison of the Syncytin-1 locus with all other HERV-W envelope elements revealed a 12-bp (corresponding to four LQMV amino acids) deletion in its cytoplasmic tail (Bonnaud et al. 2004). Moreover, insertion of these four amino acids into Syncytin-1 tail completely abolished the fusogenic potential (Bonnaud et al. 2004). This result argues that Syncytin-1 is constitutively fusion competent, as opposed to exogenous retroviruses envelopes, and is coherent with a domestication point of view since no viral protease open reading frames exist anymore in the human genome (Voisset et al. 2000) (Fig. 4.5). Furthermore, the role of the cytoplasmic domain of Syncytin-1 has been systematically investigated by producing a series of C-terminal truncated variants, leading to the conclusion that residues adjacent to the membrane domain are required for optimal fusion probably by forming a helical structure, while final C-terminal residues more likely act as a fusion inhibitor domain (Drewlo et al. 2006; Cheynet et al. 2005). Remarkably, a truncation mutant which shortens the cytoplasmic tail precisely at the site of the LQMV-deletion motif exhibits higher fusogenic properties than the wild-type protein (Cheynet et al. 2005). Even if no work on Syncytin-2 has been done in such a fine way to assess the fusogenic properties modulation its cytoplasmic tail, we

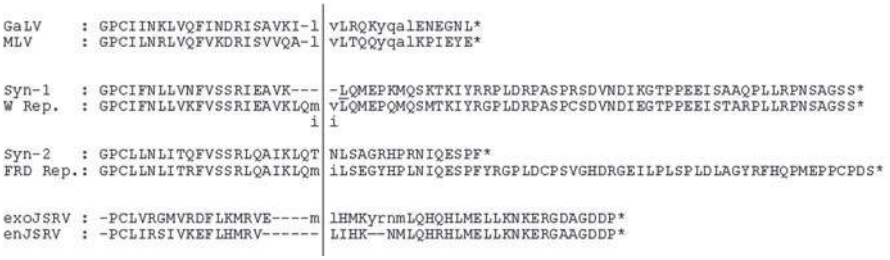


Fig. 4.5 Comparative evolution of Syncytins cytoplasmic tails: from viruses to genes. The first five amino acids correspond to the transmembrane domain. Experimentally determined (GaLV, MLV, exoJSRV) and putative (W Rep. and FRD Rep.) protease cleavage site (*black line*) and YXXΦ signaling motif are indicated in lowercase. Comparison of the Syncytin-1 protein (Syn-1) with the HERV-W family consensus sequence obtained from Repbase (W Rep.) shows a four amino acids deletion (LQMV) in the domesticated fusogenic protein, overlapping the ancestral viral protease cleavage site. The underlined leucine indicates a C-terminal truncation mutant exhibiting hyperfusogenic activity and significant pseudotyping capacity. Comparison of the Syncytin-2 protein (Syn-2) with the Repbase FRD consensus sequence (FRD Rep.) shows a stop codon that shortens the Syncytin-2 cytoplasmic tail and no evidence of viral protease cleavage site. Alignment of enJSRV and exoJSRV shows the placenta-expressed enJSRV has accumulated mutations surrounding the protease cleavage site and lacks downstream tyrosine (Y) residue. Genebank accession numbers: MLV: M14702; GaLV: AF055060, Syncytin-1: GQ919057, Syncytin-2: HEU27240, enJSRV: enJS56A1 and exoJSRV: AF105220

identified a stop codon in the cyt of Syncytin-2, as opposed to the RepBase prototype, resulting in a shortening of the tail (Fig. 4.5). Moreover, the protease cleavage site appears absent as regard to the FRD family consensus genome.

Unlike the endogenous retrovirus enJSRV, the exogenous JSRV is pathogenic for sheep and is responsible for a transmissible lung cancer in sheep *via* its envelope glycoprotein acting as a dominant oncoprotein (Palmarini et al. 1999). Studies on the cytoplasmic tail of JSRV envelope protein first focused on the VR3 region that was described as the least conserved region between exogenous and endogenous forms. The VR3 region includes the putative membrane-spanning domain as well as the cytoplasmic tail, and series of envelope chimeras revealed that mutations in a YXXM motif of the cytoplasmic tail of JSRV env were sufficient to inhibit its transforming abilities (Palmarini et al. 2001). Further mutational amino acid substitutions have proven the tyrosine residue to be essential for transformation of exogenous JSRV. It is noteworthy that the VR3 region of all exogenous stains of JSRV sequences exhibit this tyrosine residue whereas all the enJSRVs envelopes described so far lacked this motif critical for JSRV transformation (Fig. 4.5). However, despite differences in terms of motif and sequence, JSRV and enJSRVs envelopes use the same cellular receptor called HYAL2. Further systematic mutagenesis studies of the cytoplasmic tail of JSRV envelope TM protein have established four categories of mutants that allow the TM to be devised into subdomains with regard to the transformation efficiency. Among them, mutations in the YXXM motif have various effects including the generation of "supertransformers" while the last nine amino acids of the cytoplasmic tail appear not essentials for the envelope-induced transformation (Hull and Fan 2006).

4.3.2.2 Receptor Binding

Consistent with the virology paradigms, the comprehensive search for endogenous retroviral envelope functions led to the identification of the associated receptor (or co-receptors) that allows fusion events to be complete. In 2000, state of the art was to consider three main virus receptor types, PiT-1 and PiT-2, two independent inorganic-phosphate symporters for GaLV and MLV viruses, respectively, and the RDR/Type D receptor, a neutral-amino acid transporter for the cat endogenous retrovirus RD114 and type D simian retroviruses. When we first attempted to check for the right one that could trigger cell fusion, evidence in favor of the RDR/Type D hASCT2 mammalian receptor involvement was revealed through receptor-blocked experiments in cell lines transfected with Syncytin-1 gene (formerly HERV-W Env in Blond's report) (Blond et al. 2000). Two years later, cell–cell fusion and pseudotypes virion infection assays demonstrated that Syncytin-1 efficiently uses both hASCT2 and the related hASCT1 transporter as receptors, and could recognize the mouse mASCT2 and mASCT1 transporters lacking their N-glycosylation sites removed by mutagenesis (Lavillette et al. 2002). Very interestingly, Syncytin-Ory1 also uses the hASCT2 transporter as specific receptor (Heidmann et al. 2009). Anti-RDR serum used for histochemical and flow cytometric biodistribution analyzes, further unveiled a broad expression pattern of RDR/hASCT2 in many normal tissues

including colon, testis, ovary, bone morrow and skeletal muscle (Green et al. 2004). To move forward into the comprehensive relationship between Syncytin-1 and its hASCT2 receptor, Cheynet and coworkers designed a truncated series of HERV-W SU subunits and performed cell fusion assays. This led to the identification of a minimal N-terminal 124 amino acids of the mature SU glycoprotein required as a receptor-binding domain (RBD) (Cheynet et al. 2006). Furthermore, within this domain, one especially conserved sub-domain among retroviruses belonging to the interference group, the $SDGGGX_2DX_2R$ motif, was proved to be essential for Syncytin-1-hASCT2 interaction (Cheynet et al. 2006).

The human Syncytin-2 receptor has been identified more recently. By using an old powerful approach based on a human/Chinese hamster radiation hybrid panel mapping, one candidate gene was identified to encode a putative transmembrane protein (Esnault et al. 2008). The major facilitator superfamily domain containing 2 (MFSD2) was showed to confer cell–cell fusogenicity in the presence of Syncytin-2 and infectivity to Syncytin-2 pseudotypes. MFSD2 belongs to a large family of presumptive carbohydrate transporters, and is highly conserved among mammalian genomes.

Although RDR and MFSD2 are two different receptors, they belong to the ion channels and small molecules transporters category, exhibiting a classical hydrophobic profile composed of transmembrane helices. The situation is outstandingly different for the HYAL2 JSRV receptor. Interestingly, early experiments to localize the JSRV receptor gene region were also based on pseudotyping and radiation panel screening (Rai et al. 2000). After identifying a set of overlapping clones, genetic analyses confirmed that HYAL2 was the only protein that functions as a JSRV env receptor (Rai et al. 2001). If the name HYAL2 first suggests a mainly strong hyaluronidase function, studies indeed showed low but detectable hyaluronidase activity pH-dependant (Lepperdinger et al. 1998) Actually, amino acid sequence analyses established one transmembrane domain as well as an hydrophobic carboxyl terminus and upstream signal, indicating that HYAL2 is likely attached to the membrane by a glycosylphosphatidylinositol (GPI) anchor (Rai et al. 2001). As compared to other GPI-anchored proteins, HYAL2 can potentially be involved in signal transduction and mitogenic responses, strongly suggesting a role of HYAL2 in JSRV Env-mediated oncogenesis (Rai et al. 2001).

No receptors for mice Syncytin-A and -B have been identified to date. However, fusogenic experiments showed that each Syncytin-A and Syncytin-B expression vectors triggered fusion in two different cell lines, respectively. This result argues in favor of a divergent receptor usage for these two envelopes proteins (Dupressoir et al. 2005).

4.3.2.3 Incorporation in Particles

Although one may speculate that domesticated endogenous envelopes are not supposed to spread using viral particle carrier, in an initial attempt to test the fusion capacity of HERV-W Env/Syncytin-1, we sought to generate retroviral pseudotypes in which MLV core particles were coated with HERV-W Env glycoproteins.

The absence of infectivity of HERV-W Env pseudotypes was due to the inability of HERV-W Env glycoproteins to be incorporated on MLV particles (Blond et al. 2000). Conversely, pseudotyping of HIV-1 virions with the HERV-W envelope resulted in infectious viruses, although with a poor efficiency (An et al. 2001). Such a difference in incorporation efficiency could be attributed to the lenght of the cytoplasmic tail of Syncytin-1, which is longer than the MLV one, but closer to the HIV-1 one. On line with this, shortening of the Syncytin-1 cytoplasmic tail (Fig. 4.5) significantly enhances pseudotyping of HIV-1 viral cores (Lavillette et al. 2002).

4.3.3 Retroviral Envelopes Are Involved in the Placenta Development

Syncytins or envelope glycoproteins, originated from ancient retroviral infections, are now fully integrated in mouse, rabbit, sheep and human genomes. They are naturally expressed in the placenta, following a biological protein maturation process and interacting with specific receptors. Based on these observations, the purpose is now to understand how they contribute to the placenta morphogenesis. As *bona fide* genes, their transcription is strongly regulated by different cellular mechanisms, they are under local and temporal expression control, and finally they cooperate to an integrative way in a biological system as a whole, involving numerous co actors. The next part aims to develop these points.

4.3.3.1 Envelope and Receptor Localization Throughout Mammalian Gestation

In humans, the embryo implantation process and the placental development are driven by fetal cytotrophoblasts stem cells. Placental growth is characterized by proliferation and differentiation of the villous cytotrophoblast pool into a multinucleated syncytiotrophoblast layer upon fusion events. This polarized layer constitutes the main fetomaternal interface in direct contact with the maternal blood (Fig. 4.6a). The general attempt to finely localize Syncytin-1 protein within the

Fig. 4.6 (continued) detected either in CT or ST and MFSD2 receptor is detected in ST. **b** Immunohistochemical detection of Syncytin-1 protein (SC-Syn1) at the apical syncytiotrophoblast microvillus membrane of a 10 weeks gestation normal placenta (*upper*). Note that desmoplakin, a protein of the desmosomiale plaque involved in intercellular junctions, is absent from the syncytiotrophoblast fused tissue and lines the plasmatic membranes of the cytotrophoblasts (CT-d). The hASCT2 receptor is observed at the membrane of cytotrophoblastic cells (CT-hASCT2) underlying the syncytiotrophoblast (ST). **c** Transcriptional and epigenetic control of Syncytins and associate receptors during human gestation. Promoter regions are indicated as *boxes* and CpG schematized by *circles*. TSE, trophoblast specific enhancer, U3, LTR promoter, R, transcription initiation site. CpG methylation is determined by bisulfite sequencing PCR in cytotrophoblasts (CT) at different times of gestation. Each *line* represents an independent molecule. Methylated CpGs are schematized by *black circles* and unmethylated CpGs by *white circles*

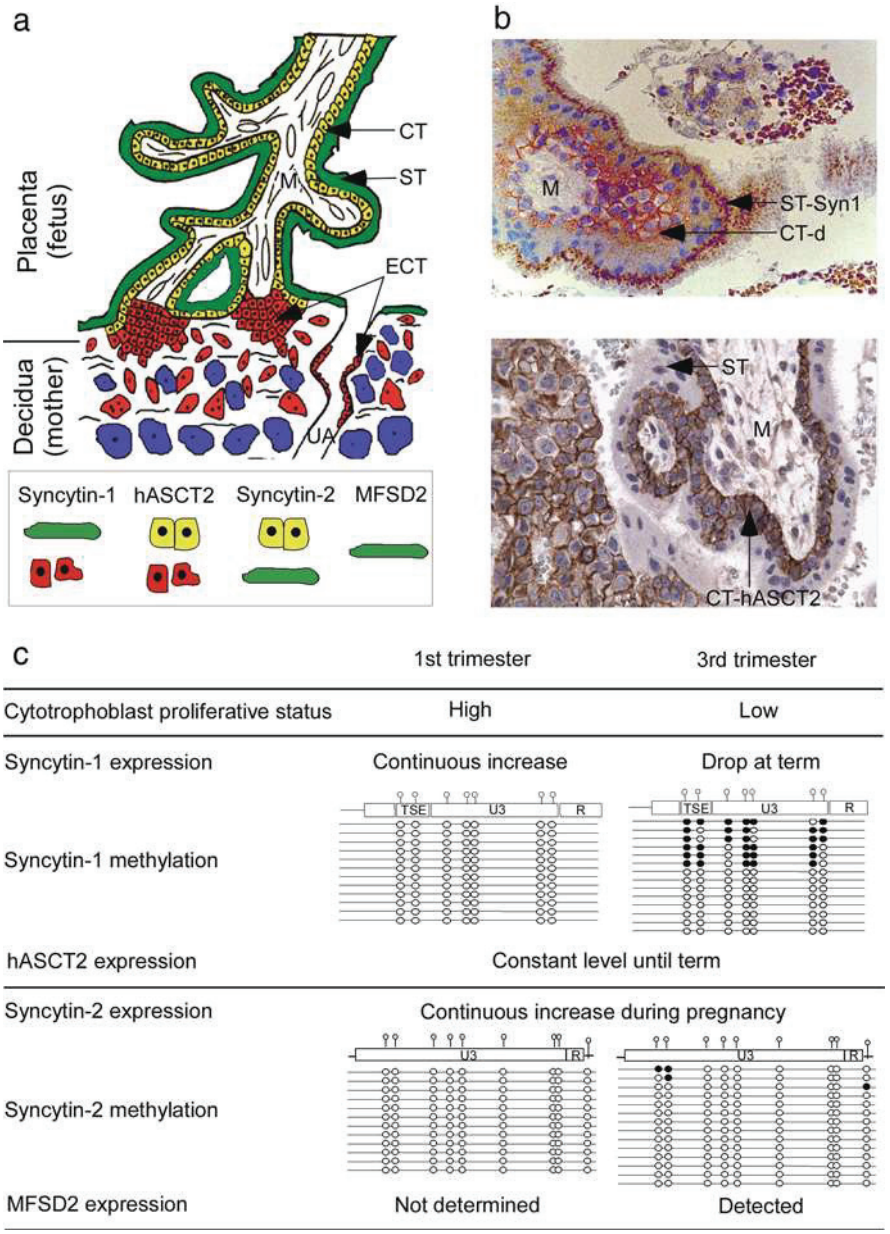


Fig. 4.6 Local and temporal expression of Syncytins in human placenta. **a** Schema of human chorionic villi. Cytotrophoblastic cells (CT, in *yellow*) differentiate by fusion to generate the syncytiotrophoblast (ST, in *green*). In the anchoring villi the cytotrophoblast cells proliferate and invade the decidua. The extravillous cytotrophoblast cells (ECT) invade the uterin stroma and differentiate into multinucleated giant cells and invade also the lumen of uterine arteries (UA). M: mesenchyme. Sites of expression of Syncytins and receptors are symbolized in the *lower box*: Syncytin-1 protein is detected in ST and ECT, hASCT2 receptor is detected in CT and ECT, Syncytin-2 protein is

placenta led the authors to use homemade antibodies that resulted in variable conclusions. With full knowledge of that facts, we decided to focus on the hypothesis that Syncytin-1 is preferentially detected at the apical membrane of the syncytiotrophoblast (Frendo et al. 2003b; Muir et al. 2006) (Fig. 4.6b). Indeed, two convergent studies mentioned that Syncytin-1 is located on specific membrane areas, enriched in cholesterol and called detergent-resistant membrane (DRMs) or rafts (Cheynet et al. 2005; Strick et al. 2007). Moreover, in primary trophoblast cells, Syncytin-1 detection was likely to be associated with cell-to-cell contact zones (Vargas et al. 2009). According to several authors, the level of Syncytin-1 protein increases during early pregnancy but remarkably decreases in very late pregnancy (Smallwood et al. 2003; Muir et al. 2006). The hASCT2 receptor expression is restricted to the cytotrophoblast compartment (Hayward et al. 2007) (Frappart, Cheynet, Mallet, unpublished data), being largely absent in the syncytiotrophoblast (Fig. 4.6b) and no spacial or temporal changes in the hASCT2 expression has been associated with the proliferative status of cytotrophoblast cells (Hayward et al. 2007). On another hand, divergent observations mentioned Syncytin-2 expression to be either restricted to villous cytotrophoblast cells (Malassine et al. 2007) or in the syncytiotrophoblast (Chen et al. 2008). Compromising positions accordingly associated Syncytin-2 with cell-to-cell contact regions, likely at the interface between the cytotrophoblast and the syncytiotrophoblast (Malassine et al. 2007; Vargas et al. 2009). Remarkably, the level expression pattern of Syncytin-2 follows an inverse correlation compared to Syncytin-1, since a significant increase in Syncytin-2 mRNA and protein is monitoring through pregnancy time and primary trophoblast culture evolution (Chen et al. 2008; Vargas et al. 2009). Finally, the MFSD2 receptor expression is unambiguously reported at the level of the syncytiotrophoblast (Esnault et al. 2008). In addition to the villous phenotype, the cytotrophoblastic cells of the anchoring villi can proliferate and invade the endometrium to be in contact with the spiral arterioles of the mother uterus. In these cells that do not fuse, both Syncytin-1 and his receptor hASCT2 have been detected (Malassine et al. 2005; Muir et al. 2006), suggesting that the trophoblastic cell fusion is indeed a complex multifactorial process (Malassine et al. 2005).

The mouse placenta is composed of spongiotrophoblasts, giant cells and a so-called labyrinth zone. In this placenta labyrinth, two layers of multinucleated syncytiotrophoblast cells, resulting from cell–cell fusion, function as the major transport surface for nutrient and gas exchange between the maternal and fetal circulation. Early in situ hybridizations showed that Syncytin-A and Syncytin-B are expressed at the level of syncytiotrophoblats, all over the labyrinth zone (Dupressoir et al. 2005). In coherence with that, latter studies more precisely indicated a clear expression of Syncytin-A in the syncytiotrophoblast while expression in trophoblast stem cells and in trophoblast giant cells could hardly be detected (Gong et al. 2007).

The rabbit placenta can be divided into the maternal decidua (the uterus modifications after implantation) and the placental lobe, in which a labyrinthine structure results from the fetal invading process. At this interface, a junctional zone presumably formed by cellular cytotrophoblasts takes place and defines a broad syncytial front. In situ hybridizations on paraffin sections of rabbit placenta have shown Syncytin-Ory1 expression to be restricted at the junctional zone and limited

to the trophoblast cells surrounding the invading fetal vessels (Heidmann et al. 2009). The authors also suggest that the labeling is compatible with an expression of Syncytin-Ory1 in the cytotrophoblast just before fusion takes place and so is consistent with a role for Syncytin-Ory1 in the formation of the syncytiotrophoblast (Heidmann et al. 2009).

In the sheep conceptus, trophoblast binucleated cells are in many respects analogous to the trophoblast giant cells of the human syncytiotrophoblast (Hoffman and Wooding 1993). They first appear at day 14 post-coitum and progressively form the outer layer of the fetal placental cotyledon giving rise to the syncytial plaques (Wooding 1984; Palmarini et al. 2001). The plaques then may cover the surface of the endometrial carunucles and aid in development of placentomes that are required for hematrophic nutrition exchanges from the maternal uterus. RT-PCR analyses have showed that endogenous JSRV envelopes and HYLA2 were expressed in the trophoblast giant binucleated cells and in the multinucleated syncytial plaques (Dunlap et al. 2006). If both endogenous JSRV envelopes and HYAL2 are detected in placentome throughout gestation, HYAL2 expression was not detected in endometrium (Dunlap et al. 2006). Endogenous JSRV envelopes were first detected in the day 12 conceptus, whereas HYAL2 first detection appeared at day 16, in a coherent way with the initial differentiation start of the binucleated cells at day 14.

4.3.3.2 Splicing Strategy, Transcription Factors and Epigenetic Control

For a brief reminder, Syncytins's chromosomal localizations are 7q21.2 (ERVWE1/Syncytin-1) (Blond et al. 1999) and 6p24.1 (ERVFRDE1/Syncytin-2) (Blaise et al. 2003) in human, and 5qG2 (Syncytin-A) (Dupressoir et al. 2005) and 14qD1 (Syncytin-B) in mouse (Dupressoir et al. 2005). To date the rabbit genome is not available in a definitive assembly to check the Syncytin-Ory1 localization and several integrations sites for enJSRV exist in the sheep genome. Putative splice donor and acceptor sites have been identified for all of them (Blond et al. 1999), although only specific transcripts have been detected depending on the biological context. ERVWE1 produces three major singly-spliced transcripts in placenta (Blond et al. 1999). The first one, 7.4-kb long, containing the *gag* and *pro/pol* pseudogenes and *env* gene, is found both in testis and placenta (Mi et al. 2000), while the 3.1 kb long, strictly including the open reading frame for the envelope protein Syncytin-1, is exclusively detected in the placenta. Early northern blot experiments also detected a 1.3-kb largely-spliced transcript in placenta (Blond et al. 1999), indeed containing the cytoplasmic tail of Syncytin-1. These observations are similar to lentivirus or oncovirus transcription patterns such as human immunodeficiency virus (HIV), the mouse mammary tumor virus (MMTV) or the human T-cell leukemia virus (HTLV), in which several genomic and subgenomic transcripts derive from a single locus by alternative splice variations.

Like any other classical retrovirus, endogenous retroviruses may display all the signals required for the transcription initiation and regulation within their long terminal repeat sequences (LTRs) (subdivided in three regions named U3, R and U5). Typically, the U3 region of the 5′LTR possesses a promoter activity. Fine studies

have well described how Syncytin-1 is under upstream regions control. The core promoter domain within the U3 region contains CAAT box and TATA box located upstream of the CAP site marking the beginning of the R region (Prudhomme et al. 2004). Mutant analyses have confirmed the functional role of these boxes. Moreover, the 5′ end of the U3 region harbors multiple binding sites contributing to overall promoter efficiency including GATA, Sp-1, AP-2, Oct-1, and PPAR-γ/RXR. Although Sp-1 and Ap-2 binding sites remain putative, they have been found to be essential for LTR activity (Prudhomme et al. 2004). It is noteworthy that Syncytin-1 regulation elements not only include the 5′LTR but also a so-called upstream regulatory element (URE), a cellular 436 bp sequence immediately upstream the Syncytin-1 proviral integration site, that define together with the 5′LTR a bipartite control element (Prudhomme et al. 2004). This URE is composed of two main domains: (i) a distal regulatory region, including the previously putative binding sites found in the promoter core as well as binding sites for the NF-κB and AP-1 important for the stimulation by TNFα, IFNγ, IL-1β, IL-6, and the inhibition by IFNβ (Mameli et al. 2007) (ii) a MaLR retrotransposon with binding sites for glucocorticoid and progesterone receptors (Bonnaud et al. 2005), and including a trophoblast specific enhancer (TSE) with putative sequences for ubiquitous Ap-2, Sp-1 and placenta-specific GCMa binding sites (Prudhomme et al. 2004).

Glial cell missing (GCM) is a transcription factor family that has gradually attracted the attention of placenta researches. Originally isolated from a *Drosophila melanogaster* mutant line, two GCM homologues (GCMa and GCMb) have then been reported in mice, rats and humans (Keryer et al. 1998). GCMa is characterized by a zinc-coordinating DNA binding domain of β-sheets that recognizes an octomeric GCM binding motif 5′-ATGCGGGT-3′ (Cohen et al. 2003). GCMa is primarily expressed in placenta in humans and highly expressed in the labyrinthine trophoblast cells in mice (Basyuk et al. 1999). Two binding sites by which GCMa can specifically transactivate Syncytin-1 have been described (Yu et al. 2002). Moreover, GCMa regulation has been linked to the cyclic AMP (cAMP) and protein kinase A signaling pathways (Chang et al. 2005; Knerr et al. 2005). In agreement with these observations, the Syncytin-1 5′LTR core promoter is cAMP-inducible (Prudhomme et al. 2004). Interestingly, a recent microarray approach that aimed to identify GCMa target genes reported Syncytin-A to be downregulated in murine GCMa-deficient placenta (Schubert et al. 2008), and siRNA GCMa inhibition in BeWo cells led to a decrease in syncytialization upon fusion events (Baczyk et al. 2009). Altogether, these data argue that GCMa acts as a major regulator in the humans and mice Syncytins expression as well as in placenta maintenance and development.

To conclude with this regulation mechanisms overview, the imprinting hypothesis and the influence of the methylation level is briefly discussed. Genomic imprinting in mammals is though to be a rescue mechanism that maintains balanced growth and development through monoallelic expression of genes in placenta and embryo. Although very little is known about the regulation of most imprinted genes, in 2003 the observations that Syncytin-1 maps very closely to two neighboring maternally imprinted retroelements, SGCE and PEG10, and according to

their temporally coordinated regulation, the hypothesis that Syncytin-1 could be paternally expressed emerged (Smallwood et al. 2003). On the other hand, methylation pattern studies of the Syncytin-1 5′LTR revealed an inverse correlation between CpG methylation and locus expression indicating that demethylation of the promoter is a prerequisite for the Syncytin-1 expression in trophoblasts cells (Matouskova et al. 2006). In an attempt to generalize this epigenetic characterization, Gimenez and colleagues recently compared different HERV promoters methylation profiles and showed that Syncytin-1 and Syncytin-2 5′LTR are widely hypomethylated in cytotrophoblasts during pregnancy, although in a distinct and pregnancy-stage-dependant manner (Gimenez et al. 2009) (Fig. 4.6c). For instance, the Syncytin-2 locus remains unmethylated throughout pregnancy, whereas methylation of the Syncytin-1 locus appears increased in the last trimester. Thus, the selective and temporal unmethylation of the Syncytin-1 locus in placenta during the first trimester may allow Syncytin-1-mediated cell differentiation and fusion, while, in contrast, increased methylation at term may limit Syncytin-1 production and consequent cell fusion or putative anti-apoptotic protection (Knerr et al. 2007) in accordance with cytotrophoblast limited fusion and higher apoptosis rate.

4.3.3.3 Additional Factors

The complete inhibition of cytotrophoblast fusion can't be reached by blocking the Syncytin-1 protein (Mi et al. 2000; Frendo et al. 2003b), and as we mentioned above, extravillous cells that express Syncytin-1 and hASCT2 do not necessarily fuse. This indicates so far that Syncytin-1 plays a major role in fusion, but also strongly suggests that other elements may contribute to this event. We proposed here to review some of them, with special interest focuses on plasma membrane dynamics, cell–cell communication and immunity as a link between the various events leading to a coordinated tissue function. Human placenta only will be discussed.

Before many retroviral receptors were known, observations emerged that an infected cell could not be superinfected by the same retrovirus, and sometimes even not being superinfected by a different one (Kim et al. 2004). Thus, interference groups were defined as set of retroviruses that cannot infect a cell at the same time. Indeed, all the members of an interference group utilize the same receptor for cell entry, and when a cell is infected once, cellular receptors are blocked by the envelope proteins of the first retrovirus and no longer available for infection of the second one. It is obvious that the balance between the envelope and its local receptor availabilities needs to be taken under consideration before predicting anything, as illustrated by FLV (Sommerfelt and Weiss 1990). Yet, given that Syncytin-1 is an endogenous envelope protein, we can expect the same rules to be valid. Indeed, together with spleen necrosis virus (Ponferrada et al. 2003), Syncytin-1 belongs to the RDR/Type D mammalian receptors interference group (Blond et al. 2000). Consequently, the hASCT2 receptors would no longer be available for type D retroviruses infection as soon as Syncytin-1 is co-locally expressed in massive doses. This is the core hypothesis to discuss different models that integrate local and temporal expression data (for a complete review, see (Potgens et al. 2004)). Overall it is a clear indication that the

cytotrophoblast fusion is guided through local availabilities of both envelopes and their corresponding receptors.

We previously mentioned how precisely Syncytin-1 seems to be associated with a restricted part of the polarized cell membrane, i.e. mainly detected at the apical part of the syncytiotrophoblast with strong suspicions that turn around a membrane subdomain enriched in cholesterol and defined as DRMs/rafts. Plasma membranes are lipid bilayers with an asymmetrical distribution of phospholipids between the inner and the outer leaflet. They are fluidic structures that permanently reorganized themselves, through dynamic lateral diffusivity, rotations, and flippase-mediated flip-flop switches. Thus, membrane-anchored proteins are part of this flow. This would allow them to reach and stay at a plasma membrane sub-localization, as well as to contribute to their functional transmembrane stable insertion. Flippases are ATP-dependant translocase enzymes that assure the asymmetrical distribution of phospholipids between the inner and the outer leaflet of the membrane. Remarkably, the loss of asymmetry triggered by a redistribution of phosphatidylserine from the inner to the outer leaflet has been described to be a prerequisite for fusion in skeletal muscle (van den Eijnde et al. 2001) and in placenta derived BeWo cells (Lyden et al. 1993). In agreement, the loss of membrane asymmetry in a cell has been associated with early stages of the apoptosis cascade. Caspase 8 is a caspase initiator involved in early apoptosis that inactivates the flippases, resulting into the asymmetry loss. When antisense and peptide inhibition strategies against caspase 8 are used, fusion of trophoblast cells is inhibited (Black et al. 2004). In addition, cholesterol-enriched domains are associated with weak fluidic properties. Although the direct relation between membrane dynamics and fusogenic proteins still remains to be elucidated, we can speculate that the positioning of Syncytin-1 that leads to the fusion event, may occurs within a well-controlled membrane subdomain with physical properties that form a stringent envelope environment compatible with receptor binding and subsequent events.

The different points we just mentioned mainly focused on the syncytiotrophoblast. Syncytin-1 membrane localization, phospholipids dynamics and early stages of apoptosis were presented as part of the multinucleated cell life. The problem here is that the syncytiotrophoblast is presumed to have a very low transcriptional activity and likely depends on the input of RNA to avoid necrosis, as indicated by different in vitro experiments (Bernirschke and Kaufmann 2000; Huppertz et al. 1999). This suggests the importance of an effective cell–cell communication and material supply systems. Gap junctions are transmembrane channels composed of connexin that provide a diffusion system for small proteins such as cAMP, IP3 or Ca^{2+}. In primary trophoblast culture, the inhibition of connexin 43 (Cx43) resulted in a fusion inhibition (Frendo et al. 2003b) and a decrease of Syncytin-1 mRNA expression (Frendo et al. 2003a). Moreover, the hASCT2 receptor is a Na^{+}-dependent amino acid transporter that can carry amino acids such as L-glutamine, L-alanine, L-leucine and L glycine, through the membrane. The clear localization of the receptor within the membrane of cytotrophoblastic cells underlying the syncytiotrophoblast (Hayward et al. 2007) (Frappart, Cheynet, Mallet, unpublished data), suggests that hASCT2 could efficiently change

the amino acid balance between the cytotrophoblast and the syncytiotrophoblast. Altogether, material flow that involves small amino acids and molecules, along with numerous electric charges balance changes (Ca^{2+}through gap junctions, anionic phosphatidylserine *via* flip flop events) appear to play a direct or indirect role in the fusion regulation.

In such a model, the syncytiotrophoblast is indeed described as one element, part of a global turnover system which includes a generative pool of cytotrophoblast cells that can feed the multinuclear layer upon fusion events and thus maintain placental growth. Athough Syncytin-1 and Syncytin-2 and their receptors play major roles, they are probably not the only proteins involved in this cooperative mechanism. Given that the placenta is an extra-embryonic tissue, half paternal and half maternal genetically inherited, the past decades have gathered reproductive immunologists researches to solve the fetal allograft problem. The contact zone between mother uterus and fetus extravillous cells of spiral arterioles appears to be one of these predictive immunological conflict zones. In direct connection with our topic on retroviral fusogens, note that the simian retrovirus (SRV), that induces immunodeficiency, belongs to the same interference group that Syncytin-1, i.e. it binds to the ASCT2 amino acid receptor. The link between immune response and amino acid balance has been seriously explored, and interestingly the involvement of HERV in immune response has already been suggested (Espinosa and Villarreal 2000). We present here some points of discussion in such a way. During pregnancy, maternal tryptophan is required for the T lymphocytes activation and "immunosuppression by starvation" is the consequence of tryptophan depletion experiments (Mellor et al. 1999). Besides, a tryptophan-catabolizing enzyme, the indoleamine 2,3-dioxygenase (IDO), is particularly expressed in the syncytiotrophoblast. Thus, the lymphocyte regulation appears to be strongly mediated by the ability of the apical membrane to incorporate the tryptophan into the syncytiotrophoblast (Kudo and Boyd 2001). In other words, the tolerance towards the allograft is conditioned by the CD98/LAT1 tryptophan transporter and the resulting amino acid balance changes. Even if the Syncytin-1 and Ory-1 hASCT2 receptor only mediates the transport of small amino acids (and consequently probably not tryptophan), considerations about balance changes that could impact the immune system response are maybe not so far. Indeed, glutamine is a necessary substrate for the nucleotide synthesis of lymphocyte cells. In peripheral blood an optimal glutamine level is required to influence the switch within the sub-populations of T lymphocytes, Th1 and Th2, through a predominantly Th1 host response (Chang et al. 1999). In sepsis mice, glutamine supplementation changes the production of IL-6, IL-4 and IFN-γ, and thus may reverse or re-equilibrate the Th1/Th2 balance response during sepsis (Yeh et al. 2005). Although it is still in debate, the Th1/Th2 switch appears to play a critical role during pregnancy, especially through a Th1 bias in recurrent pregnancy loss (Chaouat 2007). This allows us to speculate a direct or indirect involvement of amino acids balance changes as a consequence of the envelope protein fixation on its receptor. Yet, another answer to the allograft tolerance during pregnancy emerged after it was reported that MSRV particles (related to the HERV-W family) induce T lymphocyte response (Perron et al. 2001), the analysis of the putative

immunosuppressive domain in the TM subunit of different Syncytins has revealed a immunosuppressive activity for Syncytins-2 and -B but surprisingly not for the Syncytins-1 and -A (Mangeney et al. 2007)

We can conclude this discussion by briefly reporting different abnormalities that occur in pathological contexts in relation with some elements just mentioned above. Pre-eclampsia (PE) and HELLP syndrome (hemolysis, elevated liver enzymes and low platelets) are disorders associated with abnormal placentation, including defects in syncytiotrophoblast formation. Numerous studies have associated PE and HELLP with Syncytin-1 and Syncytin-2 significant reduction (Lee et al. 2001; Knerr et al. 2002; Chen and Olson 2005; Strick et al. 2007; Chen et al. 2008). Interestingly, a redistribution of the Syncytin-1 within the syncytiotrophoblast polarized cell layer was observed for patients with PE (Lee et al. 2001). Moreover, PE is associated with a predominant Th1 immunity type (Jianjun et al. 2010), that could hypothetically make the bridge with Syncytins defects, unbalanced amino acids flux and immunity. Hypoxia is overall important in the differentiation and fusion steps, since these conditions reduce the Syncytin-1 transcriptional level and inhibit cytotrophoblast fusion, whereas hASCT2 mRNA level remains unchanged (Kudo et al. 2003; Knerr et al. 2003; Chen and Olson 2005). Finally, higher apoptotic rates are observed in cultured cytotrophoblast cells from PE and HELLP (Strick et al. 2007).

4.3.4 Syncytin-1 Expression Outside of Its Privileged Tissue

As illustrated above, the multi levels-control of expression of Syncytin-1 suggests that for all Syncytins, expression is tightly regulated to be constrained to placenta. Among Syncytins, only the expression of Syncytin-1 has been so far described outside from its privileged tissue.

Syncytin-1 is expressed in astrocytes, glial cells and activated macrophages in brain regions affected by multiple sclerosis (MS). Syncytin-1 expression in astrocytes mediates neuroimmune activation and death of oligodendrocytes by inducing the release of cytotoxic redox reactants (Antony et al. 2004). In astrocytes, Syncytin-1 induces the expression of OASIS (old astrocytes specifically induced substance), an endoplasmic reticulum stress sensor, which in turn increases the expression of inducible NO synthetase and concurrent suppression of cognate hASCT1 receptor, resulting in diminished myelin protein production (Antony et al. 2007). What mechanisms reactivate Syncytin-1 in the brain in MS is still not clear. It could be the result of viral infection of the brain, such as herpes simplex virus, which has previously been shown to transactivate Syncytin-1 expression, or cytokine deregulation (Perron et al. 1993). Indeed it has been shown in astrocyte cultures that MS detrimental cytokines, IFN-γ and TNF-α are able to induce Syncytin-1 expression through NF-κB activation, while MS protective IFN-β inhibits its expression (Mameli et al. 2007). In addition Syncytin-1 induction by exogenous TNF-α into the corpus callosum, a region of the brain frequently exhibiting demyelination in MS, leads to neuroinflammation, reduction of myelin proteins level and neurobehavioural deficits in Syncytin-1-transgenic mice, as observed in MS (Antony et al. 2007).

Interestingly as a parallel between MS and cancers, NO production in tumor vessels correlates with an increase of the over-all survival as well as the decrease of metastatic potency in experimental systems (Mortensen et al. 2004). On line with this, the level of Syncytin-1 expression represented a positive prognostic indicator for recurrence-free survival of breast cancer patients (Larsson et al. 2007). Conversely, increased Syncytin-1 expression was associated with decreased overall survival in rectal but not in colonic cancer patients (Larsen et al. 2009). The situation appears unclear in endometrial carcinoma (EnCa) where Syncytin-1 expression increase in normal endometrium of patients may possibly influence the development of endometriosis (Oppelt et al. 2009). Thus, the prognostic impact of Syncytin-1 expression appears to vary with the tumor type potentially, due to different functions associated with different pathways of reactivation. In breast cancers, Syncytin-1 expression was observed for about one-third of patients, and additionally, neighbouring endothelial cells were shown to express hASCT2 receptor (Bjerregaard et al. 2006). In vitro studies confirmed the involvement of Syncytin-1 in the fusion process between breast cancer cell lines and endothelial cells (Bjerregaard et al. 2006). Syncytin-1 associated cell–cell fusion was also identified in EnCa tumors in vivo, but interestingly, in vitro studies showed the implication of Syncytin-1 in both the fusion and the proliferation of EnCa cells (Strick et al. 2007). Syncytin-1 up regulation *via* the cAMP pathway leads to cell–cell fusion while induction by steroid hormones (estradiol) leads to proliferation. This molecular switch is apparently controlled by TGF-β1 and TGF-β3 which are induced by steroid hormones and may override Syncytin-1 mediated cell–cell fusions (Strick et al. 2007).

4.4 Conclusion

Our life begins with fusion as a successful pregnancy in mammals appears to depend on Syncytin(s), retroviral members of a family of single-pass transmembrane proteins which contribute at least to cell–cell fusion necessary for placental syncytiotrophoblast morphogenesis. As we have seen, works of the last 10 years showed that Syncytins may represent extreme examples of foreign genes domestication as different elements were preserved during (parallel) evolution to assume (partly) similar roles in various species such as rodents, lagomorphs, sheep, and primates including human. These domesticated elements represent apparently a tremendous but very minor part of endogenous retroviruses (ERV) which colonized mammalian genomes. Can we consider that gift as a pay-back of retroviruses, as they emerged from our ancestors genome by transcomplementation of retrotransposons with viral envelopes (Xiong and Eickbush 1990; Malik et al. 2000), or is there a “price to pay”?

From studies in animal and human cancers, there is little doubt that tumor hybrids/fused cells are generated in vivo and that at least in animals they can be a source of metastasis (Pawelek and Chakraborty 2008). Interestingly, fusion between cancer and normal cells can lead to restoration of the apoptosis cascade or to cell differentiation, inducing a reduced tumorigenicity. However cancerous cells fusion may also lead on the contrary to a more aggressive phenotype, and, if fusion occurs

with vascular endothelial cells, to metastasis. Consequences of Syncytin-1 expression in various cancers may reflect such a diversity. Although retrovirology is a 100 years old discipline, it is less than 5 years ago that a role for viruses in cell fusion and its importance in the overall evolution of cancer was proposed (Duelli and Lazebnik 2007).

Altogether, these findings strongly support a comparative analysis of the modalities and consequences of infectious retroviruses and endogenous retroviral Syncytins expression on the cell–cell fusion processes. HERV expression/activation seems to be a common feature in cancers, a phenomenon that has been linked to deregulation of methylation (Schulz et al. 2006). Global hypomethylation of transposable elements may be a prerequisite of chromosomal instability. Similarly, viral induced fusion might result in the chromosomal instability observed in cancer cells (Duelli and Lazebnik 2007). Hypomethylation of the ERVWE1/Syncytin-1 in placenta and in seminoma (Gimenez et al. 2010) as compared to HERV-W family hypermethylation in placenta and hypomethylation in seminoma may reflect both situations (Gimenez et al. 2009). At the protein level, clarifying the interactions between Syncytin-1 and TGF-β may contribute to elucidate the regulation of cell–cell fusions occurring in development and in other syncytial cell tumors. Thus, the increase of cholesterol efflux from cellular membrane by TGF-β could modify membrane location and function(s) of Syncytin-1. Overall, fusion/differentiation, proliferation and suggested anti-apoptotic capacities of Syncytin-1 delineate the portrait of an oncogene.

Acknowledgments We are grateful to Sarah Prudhomme, Frederick Arnaud and Juliette Gimenez for their critical comments which contributed significantly to the improvement of this chapter. We thank Danièle Evain-Brion, Thierry Heidmann, Thomas E. Spencer, and François-Loïc Cosset for providing pictures and photographs.

References

Aguilar HC, Anderson WF, Cannon PM (2003) Cytoplasmic tail of Moloney murine leukemia virus envelope protein influences the conformation of the extracellular domain: implications for mechanism of action of the R Peptide. J Virol 77:1281–1291

An DS, Xie Y, Chen IS (2001) Envelope gene of the human endogenous retrovirus HERV-W encodes a functional retrovirus envelope. J Virol 75:3488–3489

Antony JM, Ellestad KK, Hammond R et al (2007) The human endogenous retrovirus envelope glycoprotein, syncytin-1, regulates neuroinflammation and its receptor expression in multiple sclerosis: a role for endoplasmic reticulum chaperones in astrocytes. J Immunol 179:1210–1224

Antony JM, van Marle G, Opii W et al (2004) Human endogenous retrovirus glycoprotein-mediated induction of redox reactants causes oligodendrocyte death and demyelination. Nat Neurosci 7:1088–1095

Arnaud F, Caporale M, Varela M et al (2007) A paradigm for virus-host coevolution: sequential counter-adaptations between endogenous and exogenous retroviruses. PLoS Pathog 3:e170

Baczyk D, Drewlo S, Proctor L et al (2009) Glial cell missing-1 transcription factor is required for the differentiation of the human trophoblast. Cell Death Differ 16:719–727

Barbulescu M, Turner G, Seaman MI et al (1999) Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. Curr Biol 9:861–868

Barre-Sinoussi F, Chermann JC, Rey F et al (1983) Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune dzficiency syndrome (AIDS). Science 220:868–871

Basyuk E, Cross JC, Corbin J et al (1999) Murine Gcm1 gene is expressed in a subset of placental trophoblast cells. Dev Dyn 214:303–311

Berlioz-Torrent C, Shacklett BL, Erdtmann L et al (1999) Interactions of the cytoplasmic domains of human and simian retroviral transmembrane proteins with components of the clathrin adaptor complexes modulate intracellular and cell surface expression of envelope glycoproteins. J Virol 73:1350–1361

Bernirschke K, Kaufmann P (2000) Pathology of the human placenta. Springer, New York, NY, pp 22–70

Bernstein HB, Tucker SP, Hunter E et al (1994) Human immunodeficiency virus type 1 envelope glycoprotein is modified by O-linked oligosaccharides. J Virol 68:463–468

Bertrand P, Cote M, Zheng YM et al (2008) Jaagsiekte sheep retrovirus utilizes a pH-dependent endocytosis pathway for entry. J Virol 82:2555–2559

Bhattacharya J, Peters PJ, Clapham PR (2004) Human immunodeficiency virus type 1 envelope glycoproteins that lack cytoplasmic domain cysteines: impact on association with membrane lipid rafts and incorporation onto budding virus particles. J Virol 78:5500–5506

Binley JM, Sanders RW, Clas B et al (2000) A recombinant human immunodeficiency virus type 1 envelope glycoprotein complex stabilized by an intermolecular disulfide bond between the gp120 and gp41 subunits is an antigenic mimic of the trimeric virion-associated structure. J Virol 74:627–643

Bjerregaard B, Holck S, Christensen IJ et al (2006) Syncytin is involved in breast cancer-endothelial cell fusions. Cell Mol Life Sci 63:1906–1911

Black S, Kadyrov M, Kaufmann P et al (2004) Syncytial fusion of human trophoblast depends on caspase 8. Cell Death Differ 11:90–98

Blaise S, de Parseval N, Benit L et al (2003) Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. Proc Natl Acad Sci USA 100:13013–13018

Blaise S, de Parseval N, Heidmann T (2005) Functional characterization of two newly identified Human Endogenous Retrovirus coding envelope genes. Retrovirology 2:19

Blaise S, Ruggieri A, Dewannieux M et al (2004) Identification of an envelope protein from the FRD family of human endogenous retroviruses (HERV-FRD) conferring infectivity and functional conservation among simians. J Virol 78:1050–1054

Blond JL, Beseme F, Duret L et al (1999) Molecular characterization and placental expression of HERV-W, a new human endogenous retrovirus family. J Virol 73:1175–1185

Blond JL, Lavillette D, Cheynet V et al (2000) An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. J Virol 74:3321–3329

Boller K, Konig H, Sauter M et al (1993) Evidence that HERV-K is the endogenous retrovirus sequence that codes for the human teratocarcinoma-derived retrovirus HTDV. Virology 196:349–353

Bolmstedt A, Hemming A, Flodby P et al (1991) Effects of mutations in glycosylation sites and disulphide bonds on processing, CD4-binding and fusion activity of human immunodeficiency virus envelope glycoproteins. J Gen Virol 72(Pt 6):1269–1277

Bonifacino JS, Traub LM (2003) Signals for sorting of transmembrane proteins to endosomes and lysosomes. Annu Rev Biochem 72:395–447

Bonnaud B, Beliaeff J, Bouton O et al (2005) Natural history of the ERVWE1 endogenous retroviral locus. Retrovirology 2:57

Bonnaud B, Bouton O, Oriol G et al (2004) Evidence of selection on the domesticated ERVWE1 env retroviral element involved in placentation. Mol Biol Evol 21:1895–1901

Bouard D, Sandrin V, Boson B et al (2007) An acidic cluster of the cytoplasmic tail of the RD114 virus glycoprotein controls assembly of retroviral envelopes. Traffic 8:835–847

Bour BA, Chakravarti M, West JM et al (2000) Drosophila SNS, a member of the immunoglobulin superfamily that is essential for myoblast fusion. Genes Dev 14:1498–1511

Braakman I, van Anken E (2000) Folding of viral envelope glycoproteins in the endoplasmic reticulum. Traffic 1:533–539

Brody BA, Rhee SS, Sommerfelt MA et al (1992) A viral protease-mediated cleavage of the transmembrane glycoprotein of Mason-Pfizer monkey virus can be suppressed by mutations within the matrix protein. Proc Natl Acad Sci USA 89:3443–3447

Brugger B, Glass B, Haberkant P et al (2006) The HIV lipidome: a raft with an unusual composition. Proc Natl Acad Sci USA 103:2641–2646

Bugelski PJ, Maleeff BE, Klinkner AM et al (1995) Ultrastructural evidence of an interaction between Env and Gag proteins during assembly of HIV type 1. AIDS Res Hum Retroviruses 11:55–64

Caceres M, Thomas JW (2006) The gene of retroviral origin Syncytin 1 is specific to hominoids and is inactive in Old World monkeys. J Hered 97:100–106

Chang CW, Chuang HC, Yu C et al (2005) Stimulation of GCMa transcriptional activity by cyclic AMP/protein kinase A signaling is attributed to CBP-mediated acetylation of GCMa. Mol Cell Biol 25:8401–8414

Chang WK, Yang KD, Shaio MF (1999) Effect of glutamine on Th1 and Th2 cytokine responses of human peripheral blood mononuclear cells. Clin Immunol 93:294–301

Chaouat G (2007) The Th1/Th2 paradigm: still important in pregnancy? Semin Immunopathol 29:95–113

Chen CP, Chen LF, Yang SR et al (2008) Functional characterization of the human placental fusogenic membrane protein syncytin 2. Biol Reprod 79:815–823

Chen EH, Grote E, Mohler W et al (2007) Cell–cell fusion. FEBS Lett 581:2181–2193

Chen EH, Olson EN (2005) Unveiling the mechanisms of cell–cell fusion. Science 308:369–373

Chernomordik LV, Kozlov MM (2003) Protein-lipid interplay in fusion and fission of biological membranes. Annu Rev Biochem 72:175–207

Chernomordik LV, Kozlov MM (2005) Membrane hemifusion: crossing a chasm in two leaps. Cell 123:375–382

Chernomordik LV, Kozlov MM (2008) Mechanics of membrane fusion. Nat Struct Mol Biol 15:675–683

Cheynet V, Oriol G, Mallet F (2006) Identification of the hASCT2-binding domain of the Env ERVWE1/syncytin-1 fusogenic glycoprotein. Retrovirology 3:41

Cheynet V, Ruggieri A, Oriol G et al (2005) Synthesis, assembly, and processing of the Env ERVWE1/syncytin human endogenous retroviral envelope. J Virol 79:5585–5593

Christensen T, Sorensen PD, Riemann H et al (1998) Expression of sequence variants of endogenous retrovirus RGH in particle form in multiple sclerosis. Lancet 352:1033

Christodoulopoulos I, Cannon PM (2001) Sequences in the cytoplasmic tail of the gibbon ape leukemia virus envelope protein that prevent its incorporation into lentivirus vectors. J Virol 75:4129–4138

Cianciolo GJ, Copeland TD, Oroszlan S et al (1985) Inhibition of lymphocyte proliferation by a synthetic peptide homologous to retroviral envelope protein. Science 230:453–455

Coffin JM (1992) Genetic diversity and evolution of retroviruses. Curr Top Microbiol Immunol 176:143–164

Cohen M, Powers M, O'Connell C et al (1985) The nucleotide sequence of the env gene from the human provirus ERV3 and isolation and characterization of an ERV3-specific cDNA. Virology 147:449–458

Cohen SX, Moulin M, Hashemolhosseini S et al (2003) Structure of the GCM domain-DNA complex: a DNA-binding domain with a novel fold and mode of target site recognition. Embo J 22:1835–1845

Contreras-Galindo R, Kaplan MH, Leissner P et al (2008) Human endogenous retrovirus K (HML-2) elements in the plasma of people with lymphoma and breast cancer. J Virol 82:9329–9336

Cote M, Zheng YM, Liu SL (2009) Receptor binding and low pH coactivate oncogenic retrovirus envelope-mediated fusion. J Virol 83:11447–11455

Crowell RC, Kiessling AA (2007) Endogenous retrovirus expression in testis and epididymis. Biochem Soc Trans 35:629–633

de Parseval N, Heidmann T (1998) Physiological knockout of the envelope gene of the single-copy ERV-3 human endogenous retrovirus in a fraction of the Caucasian population. J Virol 72:3442–3445

de Parseval N, Casella J, Gressin L, Heidmann T (2001) Characterization of the three HERV-H proviruses with an open envelope reading frame encompassing the immunosuppressive domain and evolutionary history in primates. Virology 279:558–569

de Parseval N, Lazar V, Casella JF et al (2003) Survey of human genes of retroviral origin: identification and transcriptome of the genes with coding capacity for complete envelope proteins. J Virol 77:10414–10422

Dedera D, Gu RL, Ratner L (1992) Conserved cysteine residues in the human immunodeficiency virus type 1 transmembrane envelope protein are essential for precursor envelope cleavage. J Virol 66:1207–1209

Dewannieux M, Harper F, Richaud A et al (2006) Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. Genome Res 16:1548–1556

Dolei A (2005) MSRV/HERV-W/syncytin and its linkage to multiple sclerosis: the usability and the hazard of a human endogenous retrovirus. J Neurovirol 11:232–235

Donner H, Tonjes RR, Bontrop RE et al (1999) Intronic sequence motifs of HLA-DQB1 are shared between humans, apes and Old World monkeys, but a retroviral LTR element (DQLTR3) is human specific. Tissue Antigens 53:551–558

Drewlo S, Leyting S, Kokozidou M et al (2006) C-Terminal truncations of syncytin-1 (ERVWE1 envelope) that increase its fusogenicity. Biol Chem 387:1113–1120

Duelli D, Lazebnik Y (2007) Cell-to-cell fusion as a link between viruses and cancer. Nat Rev Cancer 7:968–976

Dunlap KA, Palmarini M, Varela M et al (2006) Endogenous retroviruses regulate periimplantation placental growth and differentiation. Proc Natl Acad Sci USA 103:14390–14395

Dupressoir A, Marceau G, Vernochet C et al (2005) Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc Natl Acad Sci USA 102:725–730

Earl PL, Doms RW, Moss B (1990) Oligomeric structure of the human immunodeficiency virus type 1 envelope glycoprotein. Proc Natl Acad Sci USA 87:648–652

Earl PL, Moss B, Doms RW (1991) Folding, interaction with GRP78-BiP, assembly, and transport of the human immunodeficiency virus type 1 envelope protein. J Virol 65:2047–2055

Edwards TG, Wyss S, Reeves JD et al (2002) Truncation of the cytoplasmic domain induces exposure of conserved regions in the ectodomain of human immunodeficiency virus type 1 envelope protein. J Virol 76:2683–2691

Einfeld D, Hunter E (1988) Oligomeric structure of a prototype retrovirus glycoprotein. Proc Natl Acad Sci USA 85:8688–8692

Esnault C, Priet S, Ribet D et al (2008) A placenta-specific receptor for the fusogenic, endogenous retrovirus-derived, human syncytin-2. Proc Natl Acad Sci USA 105:17532–17537

Espinosa A, Villarreal LP (2000) T-Ag inhibits implantation by EC cell derived embryoid bodies. Virus Genes 20:195–200

Fang Y, Wu N, Gan X et al (2007) Higher-order oligomerization targets plasma membrane proteins and HIV gag to exosomes. PLoS Biol 5:e158

Fass D, Davey RA, Hamson CA et al (1997) Structure of a murine leukemia virus receptor-binding glycoprotein at 2.0 angstrom resolution. Science 277:1662–1666

Fass D, Harrison SC, Kim PS (1996) Retrovirus envelope domain at 1.7 angstrom resolution. Nat Struct Biol 3:465–469

Fenouillet E, Barbouche R, Jones IM (2007) Cell entry by enveloped viruses: redox considerations for HIV and SARS-coronavirus. Antioxid Redox Signal 9:1009–1034

Fenouillet E, Lavillette D, Loureiro S et al (2008) Contribution of redox status to hepatitis C virus E2 envelope protein function and antigenicity. J Biol Chem 283:26340–26348

Firouzi R, Rolland A, Michel M et al (2003) Multiple sclerosis-associated retrovirus particles cause T lymphocyte-dependent death with brain hemorrhage in humanized SCID mice model. J Neurovirol 9:79–93

Freed EO, Risser R (1987) The role of envelope glycoprotein processing in murine leukemia virus infection. J Virol 61:2852–2856
Frendo JL, Cronier L, Bertin G et al (2003a) Involvement of connexin 43 in human trophoblast cell fusion and differentiation. J Cell Sci 116:3413–3421
Frendo JL, Olivier D, Cheynet V et al (2003b) Direct involvement of HERV-W Env glycoprotein in human trophoblast cell fusion and differentiation. Mol Cell Biol 23:3566–3574
Gallaher WR, Ball JM, Garry RF et al (1989) A general model for the transmembrane proteins of HIV and other retroviruses. AIDS Res Hum Retroviruses 5:431–440
Gebhardt A, Bosch JV, Ziemiecki A et al (1984) Rous sarcoma virus p19 and gp35 can be chemically crosslinked to high molecular weight complexes. An insight into virus assembly. J Mol Biol 174:297–317
Gelderblom HR, Hausmann EH, Ozel M et al (1987) Fine structure of human immunodeficiency virus (HIV) and immunolocalization of structural proteins. Virology 156:171–176
Gimenez J, Montgiraud C, Oriol G et al (2009) Comparative methylation of ERVWE1/Syncytin-1 and other human endogenous retrovirus LTRs in placenta tissues. DNA Res 16:195–211
Gimenez J, Montgiraud C, Pichon JP et al (2010) Custom human endogenous retroviruses dedicated microarray identifies self-induced HERV-W family elements reactivated in testicular cancer upon methylation control. Nucleic Acids Res 38:2229–2246
Gong R, Huang L, Shi J (2007) Syncytin-A mediates the formation of syncytiotrophoblast involved in mouse placental development. Cell Physiol Biochem 20:517–526
Gould SJ, Booth AM, Hildreth JE (2003) The Trojan exosome hypothesis. Proc Natl Acad Sci USA 100:10592–10597
Grange MP, Blot V, Delamarre L et al (2000) Identification of two intracellular mechanisms leading to reduced expression of oncoretrovirus envelope glycoproteins at the cell surface. J Virol 74:11734–11743
Green BJ, Lee CS, Rasko JE (2004) Biodistribution of the RD114/mammalian type D retrovirus receptor, RDR. J Gene Med 6:249–259
Green N, Shinnick TM, Witte O et al (1981) Sequence-specific antibodies show that maturation of Moloney leukemia virus envelope polyprotein involves removal of a COOH-terminal peptide. Proc Natl Acad Sci USA 78:6023–6027
Grigorov B, Arcanger F, Roingeard P et al (2006) Assembly of infectious HIV-1 in human epithelial and T-lymphoblastic cell lines. J Mol Biol 359:848–862
Haahr S, Sommerlund M, Christensen T et al (1994) A putative new retrovirus associated with multiple sclerosis and the possible involvement of Epstein-Barr virus in this disease. Ann NY Acad Sci 724:148–156
Han X, Sterling H, Chen Y et al (2000) CD47, a ligand for the macrophage fusion receptor, participates in macrophage multinucleation. J Biol Chem 275:37984–37992
Hayward MD, Potgens AJ, Drewlo S et al (2007) Distribution of human endogenous retrovirus type W receptor in normal human villous placenta. Pathology 39:406–412
Heidmann O, Vernochet C, Dupressoir A et al (2009) Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new "syncytin" in a third order of mammals. Retrovirology 6:107
Herbst H, Kuhler-Obbarius C, Lauke H et al (1999) Human endogenous retrovirus (HERV)-K transcripts in gonadoblastomas and gonadoblastoma-derived germ cell tumours. Virchows Arch 434:11–15
Hoffman LH, Wooding FB (1993) Giant and binucleate trophoblast cells of mammals. J Exp Zool 266:559–577
Hull S, Fan H (2006) Mutational analysis of the cytoplasmic tail of jaagsiekte sheep retrovirus envelope protein. J Virol 80:8069–8080
Hunter E, Swanstrom R (1990) Retrovirus envelope glycoproteins. Curr Top Microbiol Immunol 157:187–253
Huppertz B, Frank HG, Reister F et al (1999) Apoptosis promoting caspases are sequentially activated during trophoblast differentiation: analysis of villous cytotrophoblast and syncytial fragments in vitro. Lab Invest 79 (12):1687–1702

Inabe K, Nishizawa M, Tajima S et al (1999) The YXXL sequences of a transmembrane protein of bovine leukemia virus are required for viral entry and incorporation of viral envelope protein into virions. J Virol 73:1293–1301

Inoue N, Ikawa M, Isotani A et al (2005) The immunoglobulin superfamily protein Izumo is required for sperm to fuse with eggs. Nature 434:234–238

Jianjun Z, Yali H, Zhiqun W et al (2010) Imbalance of T-cell transcription factors contributes to the Th1 type immunity predominant in pre-eclampsia. Am J Reprod Immunol 63:38–45

Joshi A, Ablan SD, Soheilian F et al (2009) Evidence that productive human immunodeficiency virus type 1 assembly can occur in an intracellular compartment. J Virol 83:5375–5387

Kavanaugh MP, Wang H, Zhang Z et al (1994) Control of cationic amino acid transport and retroviral receptor functions in a membrane protein family. J Biol Chem 269:15445–15450

Keryer G, Alsat E, Tasken K et al (1998) Cyclic AMP-dependent protein kinases and human trophoblast cell differentiation in vitro. J Cell Sci 111(Pt 7):995–1004

Kiessling AA, Crowell R, Fox C (1989) Epididymis is a principal site of retrovirus expression in the mouse. Proc Natl Acad Sci USA 86:5109–5113

Kim FJ, Battini JL, Manel N et al (2004) Emergence of vertebrate retroviruses and envelope capture. Virology 318:183–191

Kim FJ, Manel N, Boublik Y et al (2003) Human T-cell leukemia virus type 1 envelope-mediated syncytium formation can be activated in resistant Mammalian cell lines by a carboxy-terminal truncation of the envelope cytoplasmic domain. J Virol 77:963–969

Kim HS, Takenaka O, Crow TJ (1999) Cloning and nucleotide sequence of retroposons specific to hominoid primates derived from an endogenous retrovirus (HERV-K). AIDS Res Hum Retroviruses 15(6):595–601

Knerr I, Beinder E, Rascher W (2002) Syncytin, a novel human endogenous retroviral gene in human placenta: evidence for its dysregulation in preeclampsia and HELLP syndrome. Am J Obstet Gynecol 186:210–213

Knerr I, Weigel C, Linnemann K et al (2003) Transcriptional effects of hypoxia on fusiogenic syncytin and its receptor ASCT2 in human cytotrophoblast BeWo cells and in ex vivo perfused placental cotyledons. Am J Obstet Gynecol 189:583–588

Knerr I, Schubert SW, Wich C (2005) Stimulation of GCMa and syncytin via cAMP mediated PKA signaling in human trophoblastic cells under normoxic and hypoxic conditions. FEBS Lett 579:3991–3998

Knerr I, Schnare M, Hermann K et al (2007) Fusiogenic endogenous-retroviral syncytin-1 exerts anti-apoptotic functions in staurosporine-challenged CHO cells. Apoptosis 12:37–43

Komurian-Pradel F, Paranhos-Baccala G, Bedin F et al (1999) Molecular cloning and characterization of MSRV-related sequences associated with retrovirus-like particles. Virology 260:1–9

Kowalski M, Potz J, Basiripour L et al (1987) Functional regions of the envelope glycoprotein of human immunodeficiency virus type 1. Science 237:1351–1355

Kudo Y, Boyd CA (2001) The physiology of immune evasion during pregnancy; the critical role of placental tryptophan metabolism and transport. Pflugers Arch 442:639–641

Kudo Y, Boyd CA, Sargent IL et al (2003) Hypoxia alters expression and function of syncytin and its receptor during trophoblast cell fusion of human placental BeWo cells: implications for impaired trophoblast syncytialisation in pre-eclampsia. Biochim Biophys Acta 1638:63–71

Lander ES, Linton LM, Birren B et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Langat DK, Johnson PM, Rote NS et al (1999) Characterization of antigens expressed in normal baboon trophoblast and cross-reactive with HIV/SIV antibodies. J Reprod Immunol 42:41–58

Larsen JM, Christensen IJ, Nielsen HJ et al (2009) Syncytin immunoreactivity in colorectal cancer: potential prognostic impact. Cancer Lett 280:44–49

Larsson LI, Holck S, Christensen IJ (2007) Prognostic role of syncytin expression in breast cancer. Hum Pathol 38:726–731

Laufer G, Mayer J, Mueller BF et al (2009) Analysis of transcribed human endogenous retrovirus W env loci clarifies the origin of multiple sclerosis-associated retrovirus env sequences. Retrovirology 6:37

Lavillette D, Marin M, Ruggieri A et al (2002) The envelope glycoprotein of human endogenous retrovirus type W uses a divergent family of amino acid transporters/cell surface receptors. J Virol 76:6442–6452

Lee X, Keith JC Jr, Stumm N et al (2001) Downregulation of placental syncytin expression and abnormal protein localization in pre-eclampsia. Placenta 22:808–812

Lee YN, Bieniasz PD (2007) Reconstitution of an infectious human endogenous retrovirus. PLoS Pathog 3:e10

Lepperdinger G, Strobl B, Kreil G (1998) HYAL2, a human gene expressed in many cells, encodes a lysosomal hyaluronidase with a novel type of specificity. J Biol Chem 273:22466–22470

Li H, Chien PC Jr, Tuen M et al (2008a) Identification of an N-linked glycosylation in the C4 region of HIV-1 envelope gp120 that is critical for recognition of neighboring CD4 T cell epitopes. J Immunol 180:4011–4021

Li K, Zhang S, Kronqvist M et al (2008b) Intersubunit disulfide isomerization controls membrane fusion of human T-cell leukemia virus Env. J Virol 82:7135–7143

Li M, Yang C, Tong S et al (2002) Palmitoylation of the murine leukemia virus envelope protein is critical for lipid raft association and surface expression. J Virol 76:11845–11852

Li Y, Bergeron JJ, Luo L et al (1996) Effects of inefficient cleavage of the signal sequence of HIV-1 gp 120 on its association with calnexin, folding, and intracellular transport. Proc Natl Acad Sci USA 93:9606–9611

Li Z, Pinter A, Kayman SC (1997) The critical N-linked glycan of murine leukemia virus envelope protein promotes both folding of the C-terminal domains of the precursor polyprotein and stability of the postcleavage envelope complex. J Virol 71:7012–7019

Lin L, Xu B, Rote NS (1999) Expression of endogenous retrovirus ERV-3 induces differentiation in BeWo, a choriocarcinoma model of human placental trophoblast. Placenta 20:109–118

Lin L, Xu B, Rote NS (2000) The cellular mechanism by which the human endogenous retrovirus ERV-3 env gene affects proliferation and differentiation in a human placental trophoblast model, BeWo. Placenta 21:73–78

Lindeskog M, Mager DL, Blomberg J (1999) Isolation of a human endogenous retroviral HERV-H element with an open *env* reading frame. Virology 258:441–450

Lodge R, Lalonde JP, Lemay G et al (1997) The membrane-proximal intracytoplasmic tyrosine residue of HIV-1 envelope glycoprotein is critical for basolateral targeting of viral budding in MDCK cells. EMBO J 16:695–705

Lopalco L, Longhi R, Ciccomascolo F et al (1993) Identification of human immunodeficiency virus type 1 glycoprotein gp120/gp41 interacting sites by the idiotypic mimicry of two monoclonal antibodies. AIDS Res Hum Retroviruses 9:33–39

Lopez-Verges S, Camus G, Blot G et al (2006) Tail-interacting protein TIP47 is a connector between Gag and Env and is required for Env incorporation into HIV-1 virions. Proc Natl Acad Sci USA 103:14947–14952

Lorizate M, Cruz A, Huarte N et al (2006) Recognition and blocking of HIV-1 gp41 pretransmembrane sequence by monoclonal 4E10 antibody in a Raft-like membrane environment. J Biol Chem 281:39598–39606

Loving R, Li K, Wallin M et al (2008) R-Peptide cleavage potentiates fusion-controlling isomerization of the intersubunit disulfide in Moloney murine leukemia virus Env. J Virol 82:2594–2597

Lyden TW, Johnson PM, Mwenda JM et al (1994) Ultrastructural characterization of endogenous retroviral particles isolated from normal human placentas. Biol Reprod 51:152–157

Lyden TW, Ng AK, Rote NS (1993) Modulation of phosphatidylserine epitope expression by BeWo cells during forskolin treatment. Placenta 14:177–186

MacKrell AJ, Soong NW, Curtis CM et al (1996) Identification of a subdomain in the Moloney murine leukemia virus envelope protein involved in receptor binding. J Virol 70:1768–1774

Malassine A, Blaise S, Handschuh K et al (2007) Expression of the fusogenic HERV-FRD Env glycoprotein (syncytin 2) in human placenta is restricted to villous cytotrophoblastic cells. Placenta 28:185–191

Malassine A, Handschuh K, Tsatsaris V et al (2005) Expression of HERV-W Env glycoprotein (syncytin) in the extravillous trophoblast of first trimester human placenta. Placenta 26:556–562

Malik HS, Henikoff S, Eickbush TH (2000) Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 10:1307–1318

Mallet F, Bouton O, Prudhomme S et al (2004) The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. Proc Natl Acad Sci USA 101:1731–1736

Mameli G, Astone V, Arru G et al (2007) Brains and peripheral blood mononuclear cells of multiple sclerosis (MS) patients hyperexpress MS-associated retrovirus/HERV-W endogenous retrovirus, but not Human herpesvirus 6. J Gen Virol 88:264–274

Mangeney M, Renard M, Schlecht-Louf G et al (2007) Placental syncytins: Genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins. Proc Natl Acad Sci USA 104:20534–20539

Marechal V, Prevost MC, Petit C et al (2001) Human immunodeficiency virus type 1 entry into macrophages mediated by macropinocytosis. J Virol 75:11166–11177

Marsh M, Helenius A (2006) Virus entry: open sesame. Cell 124:729–740

Martens S, McMahon HT (2008) Mechanisms of membrane fusion: disparate players and common principles. Nat Rev Mol Cell Biol 9:543–556

Matouskova M, Blazkova J, Pajer P et al (2006) CpG methylation suppresses transcriptional activity of human syncytin-1 in non-placental tissues. Exp Cell Res 312:1011–1020

Melikyan GB (2008) Common principles and intermediates of viral protein-mediated fusion: the HIV-1 paradigm. Retrovirology 5:111

Melikyan GB, Barnard RJ, Abrahamyan LG et al (2005) Imaging individual retroviral fusion events: from hemifusion to pore formation and growth. Proc Natl Acad Sci USA 102:8728–8733

Mellor A, Zhou M, Conway SJ et al (1999) HLA-G transgenic mice. J Reprod Immunol 43:253–261

Mercer J, Helenius A (2009) Virus entry by macropinocytosis. Nat Cell Biol 11:510–520

Mi S, Lee X, Li X et al (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403:785–789

Mizuochi T, Matthews TJ, Kato M et al (1990) Diversity of oligosaccharide structures on the envelope glycoprotein gp 120 of human immunodeficiency virus 1 from the lymphoblastoid cell line H9. Presence of complex-type oligosaccharides with bisecting N-acetylglucosamine residues. J Biol Chem 265:8519–8524

Mohler WA (2009) Cell–cell fusion: transient channels leading to plasma membrane merger http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=eurekah&part=A59047.

Moreno MR, Giudici M, Villalain J (2006) The membranotropic regions of the endo and ecto domains of HIV gp41 envelope glycoprotein. Biochim Biophys Acta 1758:111–123

Morita E, Sundquist WI (2004) Retrovirus budding. Annu Rev Cell Dev Biol 20:395–425

Mortensen K, Christensen IJ, Nielsen HJ et al (2004) High expression of endothelial cell nitric oxide synthase in peritumoral microvessels predicts increased disease-free survival in colorectal cancer. Cancer Lett 216:109–114

Muir A, Lever AM, Moffett A (2006) Human endogenous retrovirus-W envelope (syncytin) is expressed in both villous and extravillous trophoblast populations. J Gen Virol 87:2067–2071

Munoz-Barroso I, Durell S, Sakaguchi K et al (1998) Dilation of the human immunodeficiency virus-1 envelope glycoprotein fusion pore revealed by the inhibitory action of a synthetic peptide from gp41. J Cell Biol 140:315–323

Munoz-Barroso I, Salzwedel K, Hunter E et al (1999) Role of the membrane-proximal domain in the initial stages of human immunodeficiency virus type 1 envelope glycoprotein-mediated membrane fusion. J Virol 73:6089–6092

Novakovic S, Sawai ET, Radke K (2004) Dileucine and YXXL motifs in the cytoplasmic tail of the bovine leukemia virus transmembrane envelope protein affect protein expression on the cell surface. J Virol 78:8301–8311

Nydegger S, Foti M, Derdowski A et al (2003) HIV-1 egress is gated through late endosomal membranes. Traffic 4:902–910

Ochsenbauer-Jambor C, Miller DC, Roberts CR et al (2001) Palmitoylation of the Rous sarcoma virus transmembrane glycoprotein is required for protein stability and virus infectivity. J Virol 75:11544–11554

Ono A, Ablan SD, Lockett SJ et al (2004) Phosphatidylinositol (4,5) bisphosphate regulates HIV-1 Gag targeting to the plasma membrane. Proc Natl Acad Sci USA 101:14889–14894

Oppelt P, Strick R, Strissel PL et al (2009) Expression of the human endogenous retroviruse-W envelope gene syncytin in endometriosis lesions. Gynecol Endocrinol 25:741–747

Otteken A, Moss B (1996) Calreticulin interacts with newly synthesized human immunodeficiency virus type 1 envelope glycoprotein, suggesting a chaperone function similar to that of calnexin. J Biol Chem 271:97–103

Owens RJ, Burke C, Rose JK (1994) Mutations in the membrane-spanning domain of the human immunodeficiency virus envelope glycoprotein that affect fusion activity. J Virol 68:570–574

Owens RJ, Compans RW (1990) The human immunodeficiency virus type 1 envelope glycoprotein precursor acquires aberrant intermolecular disulfide bonds that may prevent normal proteolytic processing. Virology 179:827–833

Owens RJ, Dubay JW, Hunter E et al (1991) Human immunodeficiency virus envelope protein determines the site of virus release in polarized epithelial cells. Proc Natl Acad Sci USA 88:3987–3991

Palmarini M, Gray CA, Carpenter K et al (2001) Expression of endogenous betaretroviruses in the ovine uterus: effects of neonatal age, estrous cycle, pregnancy, and progesterone. J Virol 75:11319–11327

Palmarini M, Sharp JM, De las Heras M et al (1999) Jaagsiekte sheep retrovirus is necessary and sufficient to induce a contagious lung cancer in sheep. J Virol 73:6964–6972

Pawelek JM, Chakraborty AK (2008) The cancer cell – leukocyte fusion theory of metastasis. Adv Cancer Res 101:397–444

Pearce-Pratt R, Malamud D, Phillips DM (1994) Role of the cytoskeleton in cell-to-cell transmission of human immunodeficiency virus. J Virol 68:2898–2905

Peng X, Pan J, Gong R et al (2007) Functional characterization of syncytin-A, a newly murine endogenous virus envelope protein. Implication for its fusion mechanism. J Biol Chem 282:381–389

Pernet O, Pohl C, Ainouze M et al (2009) Nipah virus entry can occur by macropinocytosis. Virology 395:298–311

Perron H, Firouzi R, Tuke PW et al (1997) Cell cultures and associated retroviruses in multiple sclerosis. Acta Neurol Scand Suppl 169:22–31

Perron H, Geny C, Laurent A et al (1989) Leptomeningeal cell line from multiple sclerosis with reverse transcriptase activity and viral particles. Res Virol 140:551–561

Perron H, Jouvin-Marche E, Michel M et al (2001) Multiple sclerosis retrovirus particles and recombinant envelope trigger an abnormal immune response in vitro, by inducing polyclonal Vbeta16 T-lymphocyte activation. Virology 287:321–332

Perron H, Suh M, Lalande B et al (1993) Herpes simplex virus ICP0 and ICP4 immediate early proteins strongly enhance expression of a retrovirus harboured by a leptomeningeal cell line from a patient with multiple sclerosis. J Gen Virol 74:65–72

Pichon JP, Bonnaud B, Mallet F (2006) Quantitative multiplex degenerate PCR for human endogenous retrovirus expression profiling. Nat Protoc 1:2831–2838

Piller SC, Dubay JW, Derdeyn CA et al (2000) Mutational analysis of conserved domains within the cytoplasmic tail of gp41 from human immunodeficiency virus type 1: effects on glycoprotein incorporation and infectivity. J Virol 74:11717–11723

Pinter A, Honnen WJ (1988) O-linked glycosylation of retroviral envelope gene products. J Virol 62:1016–1021

Pinter A, Kopelman R, Li Z et al (1997) Localization of the labile disulfide bond between SU and TM of the murine leukemia virus envelope protein complex to a highly conserved CWLC motif in SU that resembles the active-site sequence of thiol-disulfide exchange enzymes. J Virol 71:8073–8077

Polonoff E, Machida CA, Kabat D (1982) Glycosylation and intracellular transport of membrane glycoproteins encoded by murine leukemia viruses. Inhibition by amino acid analogues and by tunicamycin. J Biol Chem 257:14023–14028

Ponferrada VG, Mauck BS, Wooley DP (2003) The envelope glycoprotein of human endogenous retrovirus HERV-W induces cellular resistance to spleen necrosis virus. Arch Virol 148:659–675

Potgens AJ, Drewlo S, Kokozidou M et al (2004) Syncytin: the major regulator of trophoblast fusion? Recent developments and hypotheses on its action. Hum Reprod Update 10:487–496

Poumbourios P, Wilson KA, Center RJ et al (1997) Human immunodeficiency virus type 1 envelope glycoprotein oligomerization requires the gp41 amphipathic alpha-helical/leucine zipper-like sequence. J Virol 71:2041–2049

Prudhomme S, Oriol G, Mallet F (2004) A retroviral promoter and a cellular enhancer define a bipartite element which controls env ERVWE1 placental expression. J Virol 78:12157–12168

Purtscher M, Trkola A, Gruber G et al (1994) A broadly neutralizing human monoclonal antibody against gp41 of human immunodeficiency virus type 1. AIDS Res Hum Retroviruses 10:1651–1658

Qian K, Morris-Natschke SL, Lee KH (2009) HIV entry inhibitors and their potential in HIV therapy. Med Res Rev 29:369–393

Rai SK, DeMartini JC, Miller AD (2000) Retrovirus vectors bearing jaagsiekte sheep retrovirus Env transduce human cells by using a new receptor localized to chromosome 3p21.3. J Virol 74:4698–4704

Rai SK, Duh FM, Vigdorovich V et al (2001) Candidate tumor suppressor HYAL2 is a glycosylphosphatidylinositol (GPI)-anchored cell-surface receptor for jaagsiekte sheep retrovirus, the envelope protein of which mediates oncogenic transformation. Proc Natl Acad Sci USA 98:4443–4448

Ratner L (1992) Glucosidase inhibitors for treatment of HIV-1 infection. AIDS Res Hum Retroviruses 8:165–173

Rein A, Mirro J, Haynes JG et al (1994) Function of the cytoplasmic domain of a retroviral transmembrane protein: p15E–p2E cleavage activates the membrane fusion capability of the murine leukemia virus Env protein. J Virol 68(3):1773–1781

Rolland A, Jouvin-Marche E, Viret C et al (2006) The envelope protein of a human endogenous retrovirus-W family activates innate immunity through CD14/TLR4 and promotes Th1-like responses. J Immunol 176:7636–7644

Rozenberg-Adler Y, Conner J, Guilar-Carreno H et al (2008) Membrane-proximal cytoplasmic domain of Moloney murine leukemia virus envelope tail facilitates fusion. Exp Mol Pathol 84:18–30

Salzwedel K, West JT, Hunter E (1999) A conserved tryptophan-rich motif in the membrane-proximal region of the human immunodeficiency virus type 1 gp41 ectodomain is important for Env-mediated fusion and virus infectivity. J Virol 73:2469–2480

Sanders DA (2000) Sulfhydryl involvement in fusion mechanisms. Subcell Biochem 34:483–514

Sandrin V, Cosset FL (2006) Intracellular versus cell surface assembly of retroviral pseudotypes is determined by the cellular localization of the viral glycoprotein, its capacity to interact with Gag, and the expression of the Nef protein. J Biol Chem 281:528–542

Sandrin V, Muriaux D, Darlix JL et al (2004) Intracellular trafficking of Gag and Env proteins and their interactions modulate pseudotyping of retroviruses. J Virol 78:7153–7164

Schubert SW, Lamoureux N, Kilian K et al (2008) Identification of integrin-alpha4, Rb1, and syncytin a as murine placental target genes of the transcription factor GCMa/Gcm1. J Biol Chem 283:5460–5465

Schulz TF, Jameson BA, Lopalco L et al (1992) Conserved structural features in the interaction between retroviral surface and transmembrane glycoproteins? AIDS Res Hum Retroviruses 8:1571–1580

Schulz WA, Steinhoff C, Florl AR (2006) Methylation of endogenous human retroelements in health and disease. Curr Top Microbiol Immunol 310:211–250

Sherer NM, Lehmann MJ, Jimenez-Soto LF et al (2003) Visualization of retroviral replication in living cells reveals budding into multivesicular bodies. Traffic 4:785–801

Sitbon M, d′Auriol L, Ellerbrok H et al (1991) Substitution of leucine for isoleucine in a sequence highly conserved among retroviral envelope surface glycoproteins attenuates the lytic effect of the Friend murine leukemia virus. Proc Natl Acad Sci USA 88:5932–5936

Smallwood A, Papageorghiou A, Nicolaides K et al (2003) Temporal regulation of the expression of syncytin (HERV-W), maternally imprinted PEG10, and SGCE in human placenta. Biol Reprod 69:286–293

Sommerfelt MA, Weiss RA (1990) Receptor interference groups of 20 retroviruses plating on human cells. Virology 176:58–69

Song C, Hunter E (2003) Variable sensitivity to substitutions in the N-terminal heptad repeat of Mason-Pfizer monkey virus transmembrane protein. J Virol 77:7779–7785

Song C, Micoli K, Bauerova H et al (2005) Amino acid residues in the cytoplasmic domain of the Mason-Pfizer monkey virus glycoprotein critical for its incorporation into virions. J Virol 79:11559–11568

Spies CP, Ritter GD Jr, Mulligan MJ et al (1994) Truncation of the cytoplasmic domain of the simian immunodeficiency virus envelope glycoprotein alters the conformation of the external domain. J Virol 68:585–591

Strick R, Ackermann S, Langbein M et al (2007) Proliferation and cell–cell fusion of endometrial carcinoma are induced by the human endogenous retroviral Syncytin-1 and regulated by TGF-beta. J Mol Med 85:23–38

Suarez T, Gallaher WR, Agirre A et al (2000) Membrane interface-interacting sequences within the ectodomain of the human immunodeficiency virus type 1 envelope glycoprotein: putative role during viral fusion. J Virol 74:8038–8047

Suomalainen M (2002) Lipid rafts and assembly of enveloped viruses. Traffic 3:705–709

Tailor CS, Nouri A, Kabat D (2000) A comprehensive approach to mapping the interacting surfaces of murine amphotropic and feline subgroup B leukemia viruses with their cell surface receptors. J Virol 74:237–244

Takeda M, Leser GP, Russell CJ, Lamb RA (2003) Influenza virus hemagglutinin concentrates in lipid raft microdomains for efficient viral fusion. Proc Natl Acad Sci USA 100:14610–14617

Tarlinton R, Meers J, Young P (2008) Biology and evolution of the endogenous koala retrovirus. Cell Mol Life Sci 65:3413–3421

Taylor GM, Sanders DA (1999) The role of the membrane-spanning domain sequence in glycoprotein-mediated membrane fusion. Mol Biol Cell 10:2803–2815

Thomas A, Roth MJ (1995) Analysis of cysteine mutations on the transmembrane protein of Moloney murine leukemia virus. Virology 211:285–289

Trujillo JR, Rogers R, Molina RM et al (2007) Noninfectious entry of HIV-1 into peripheral and brain macrophages mediated by the mannose receptor. Proc Natl Acad Sci USA 104:5097–5102

Tucker SP, Srinivas RV, Compans RW (1991) Molecular domains involved in oligomerization of the Friend murine leukemia virus envelope glycoprotein. Virology 185:710–720

Turner G, Barbulescu M, Su M et al (2001) Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr Biol 11:1531–1535

van den Eijnde SM, van den Hoff MJ, Reutelingsperger CP et al (2001) Transient expression of phosphatidylserine at cell–cell contact areas is required for myotube formation. J Cell Sci 114:3631–3642

van Regenmortel MH, Mayo MA, Fauquet CM et al (2000) Virus nomenclature: consensus versus chaos. Arch Virol 145:2227–2232

Vargas A, Moreau J, Landry S et al (2009) Syncytin-2 plays an important role in the fusion of human trophoblast cells. J Mol Biol 392:301–318

Venables PJ, Brookes SM, Griffiths D et al (1995) Abundance of an endogenous retroviral envelope protein in placental trophoblasts suggests a biological function. Virology 211:589–592

Voisset C, Blancher A, Perron H et al (1999) Phylogeny of a novel family of human endogenous retrovirus sequences, HERV-W, in humans and other primates. AIDS Res Hum Retroviruses 15:1529–1533

Voisset C, Bouton O, Bedin F et al (2000) Chromosomal distribution and coding capacity of the human endogenous retrovirus HERV-W family. AIDS Res Hum Retroviruses 16:731–740

von Schwedler UK, Stuchell M, Muller B et al (2003) The protein network of HIV budding. Cell 114:701–713

Wallin M, Ekstrom M, Garoff H (2004) Isomerization of the intersubunit disulphide-bond in Env controls retrovirus fusion. Embo J 23:54–65

Wang E, Obeng-Adjei N, Ying Q et al (2008) Mouse mammary tumor virus uses mouse but not human transferrin receptor 1 to reach a low pH compartment and infect cells. Virology 381:230–240

Waterston RH, Lindblad-Toh K, Birney E et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Weissenhorn W, Dessen A, Harrison SC et al (1997) Atomic structure of the ectodomain from HIV-1 gp41. Nature 387:426–430

Weissenhorn W, Hinz A, Gaudin Y (2007) Virus membrane fusion. FEBS Lett 581:2150–2155

Welsch S, Keppler OT, Habermann A et al (2007) HIV-1 buds predominantly at the plasma membrane of primary human macrophages. PLoS Pathog 3:e36

Wooding FB (1984) Role of binucleate cells in fetomaternal cell fusion at implantation in the sheep. Am J Anat 170:233-250

Wyss S, Berlioz-Torrent C, Boge M et al (2001) The highly conserved C-terminal dileucine motif in the cytosolic domain of the human immunodeficiency virus type 1 envelope glycoprotein is critical for its association with the AP-1 clathrin adaptor [correction of adapter]. J Virol 75:2982–2992

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9:3353–3362

Xu Y, Lu H, Kennedy JP et al (2006) Evaluation of "credit card" libraries for inhibition of HIV-1 gp41 fusogenic core formation. J Comb Chem 8:531–539

Yang C, Compans RW (1996) Analysis of the cell fusion activities of chimeric simian immunodeficiency virus-murine leukemia virus envelope proteins: inhibitory effects of the R peptide. J Virol 70:248–254

Yang C, Spies CP, Compans RW (1995) The human and simian immunodeficiency virus envelope glycoprotein transmembrane subunits are palmitoylated. Proc Natl Acad Sci USA 92:9871–9875

Yeh CL, Hsu CS, Yeh SL et al (2005) Dietary glutamine supplementation modulates Th1/Th2 cytokine and interleukin-6 expressions in septic mice. Cytokine 31:329–334

Yu C, Shen K, Lin M et al (2002) GCMa regulates the syncytin-mediated trophoblastic fusion. J Biol Chem 277:50062–50068

Zhao Y, Zhu L, Benedict CA et al (1998) Functional domains in the retroviral transmembrane protein. J Virol 72:5392–5398

Zhou Z, Shen T, Zhang BH et al (2009) The proprotein convertase furin in human trophoblast: Possible role in promoting trophoblast cell migration and invasion. Placenta 30:929–938

IV.2. Description d'une méthode d'analyse des puces à ADN Affymetrix HG-U133-PLUS2

- **Lecture des puces et acquisition du signal brut**

Le scanner conçu pour LIRE les puces Affymetrix permet d'acquérir leur fluorescence en balayant la surface en microscopie confocale. Un laser à argon excite les fluorochromes liés aux cibles hybridées. L'émission de fluorescence à 570 nm qui en résulte est captée par l'objectif du microscope et passe à travers une série de filtres jusqu'à un tube photomultiplicateur. La fluorescence est ainsi transformée en courant électrique qui est enregistré et analysé par le logiciel de traitement. L'image de la puce donne un fichier au format DAT. Une grille sera apposée virtuellement à l'image obtenue, ce qui permet d'attribuer des coordonnées pour chaque point fluorescent de l'image (Figure IV-1).

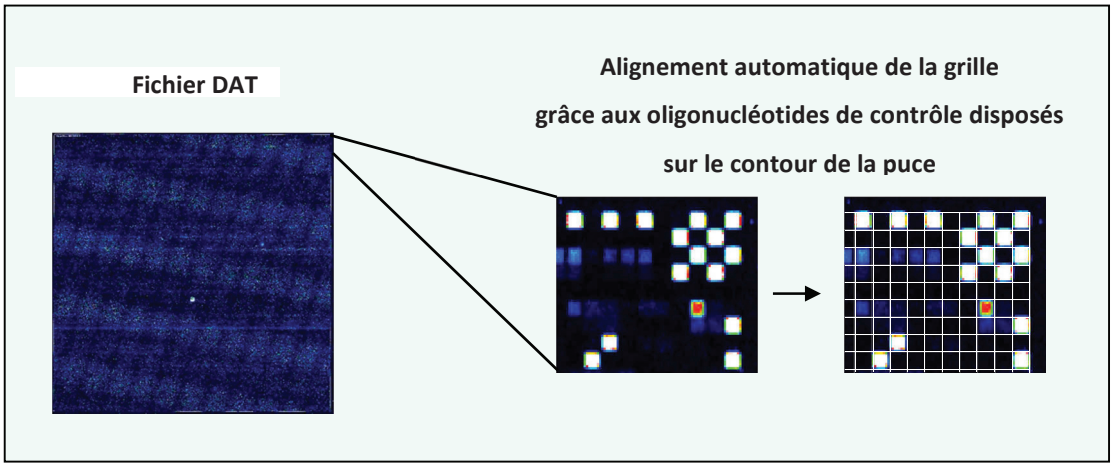


Figure IV-1 Lecture d'une puce HG-U133-PLUS2. Après acquisition de la fluorescence, une grille virtuelle est appliquée pour permettre une analyse des intensités par cellules.

Le traitement du signal se fait alors par l'utilisation de fichiers librairies, d'une part, et de fichiers d'analyse, d'autre part, afin de générer des données brutes exploitables.

Les fichiers librairies

Deux fichiers librairies constituent la base de données d'une puce :

Le fichier CDF (*Chip Description File*), qui contient le plan de conception de la puce. C'est le fichier qui sert à la synthèse du masque de fabrication. Il indique la position et la séquence des sondes.

Le fichier CIF (*Chip Information File*), qui est utilisé lors de la lecture de la puce. Il contient des informations telles que la longueur d'onde de lecture à utiliser, la taille du support (11 µm pour une puce HG-U133-PLUS2) ou encore les paramètres du scanner.

Les fichiers d'analyse

Le logiciel Affymetrix *GeneChip* génère différents types de fichiers :

Un fichier EXP qui contient les informations de l'expérience, telles que le type de puce, le numéro de lot, le rapport sur les éventuelles erreurs de machine, etc.

Un fichier DAT, comme expliqué précédemment, qui correspond à l'image scannée de la surface de la puce. Après alignement de la grille, toutes les cellules de la puce sont localisées et chaque cellule contient plusieurs dizaines de pixels ayant chacun une intensité.

Un fichier CEL qui donne des informations sur l'intensité de fluorescence par cellule à partir des intensités des pixels. Le passage du fichier DAT au fichier CEL se fait en moyennant les intensités des pixels de chaque cellule afin d'obtenir une intensité unique par cellule (Figure IV-2).

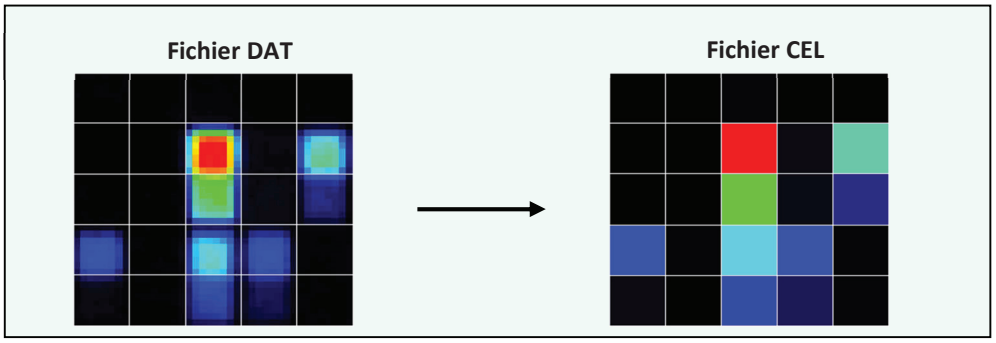


Figure IV-2 Transformation du fichier DAT en fichier CEL.

Un fichier CHP qui contient les informations analysées des données de chaque probeset.

Un fichier RPT, créé à partir du fichier CHP, qui contient les valeurs d'intensité, de bruit de fond, des contrôles d'hybridation (bioB, bioC, bioD and cre), de *scaling factor,* de normalisation et des contrôles internes (GAPDH et HSA) (sur tous ces termes, lire le paragraphe suivant).

- **Contrôles qualité des puces**

Différents paramètres sont contrôlés afin de savoir, premièrement, si la qualité d'une puce est satisfaisante et, deuxièmement, si cette puce peut être analysée avec d'autres puces. L'inspection visuelle de l'image générée par le scanner permet d'écarter les puces présentant des artefacts évidents comme de grandes zones sombres (pouvant résulter d'un défaut de fabrication) ou de grandes zones d'intensité saturée résultant du dépôt de corps insolubles pendant l'hybridation. Cette

inspection permet également de vérifier si le positionnement de la grille sur l'image, dont dépend la pertinence du fichier de données brutes généré, est correct.

D'autres paramètres sont contrôlés à partir du rapport généré par le logiciel *GeneChip* qui réalise une analyse « de première intention » de la puce. Premièrement, le bruit de fond de la puce, qui résulte de l'adsorption aspécifique de fluorophores et également d'un bruit de fond d'hybridation, ne doit pas dépasser 2^7. Deuxièmement, le bruit de fond électronique et optique du scanner ne doit pas subir de dérive au cours du temps. Troisièmement, le protocole d'amplification des ARN ne doit pas avoir introduit de biais (en dehors des biais connus et inhérents à la méthode). Ce critère est évalué à l'aide des ratios des intensités mesurées pour les régions 3'et 5' des gènes GAPDH et HSA. Quatrièmement, la qualité de l'hybridation est évaluée grâce à des cibles contrôles. Ces contrôles d'hybridation bioB, bioC, bioD et cre sont introduits dans le cocktail d'hybridation à raison de 1,5 pM, 5 pM, 25 pM et 100 pM, respectivement. BioB, qui correspond au seuil de détection d'une cible, doit être détecté dans 50 % des cas. BioC, bioD et cre doivent être détectés systématiquement et leurs intensités doivent refléter leurs concentrations relatives.

Enfin, la possibilité d'intégrer différentes puces dans une étude est déterminée par l'homogénéité des paramètres décrits ci-dessus et par un critère objectif d'homogénéité des intensités moyennes des puces. Un facteur *sf* pour *scaling factor* est calculé pour chaque puce comme étant le ratio entre une valeur cible et la moyenne des intensités de la puce. Au sein d'une étude, le *sf* le plus élevé ne doit pas excéder 3 fois le *sf* le plus petit.

$$sf_i = (\text{valeur cible})_i/(\text{moyenne des intensités de la puce})$$

- **Correction du bruit de fond**

L'existence d'un bruit de fond peut être mise en évidence simplement en marquant de l'eau stérile et en l'hybridant, avec du tampon, sur une puce. Ce bruit de fond a deux origines : il provient pour partie de l'adsorption de molécules fluorescentes sur la puce et pour une autre partie des artéfacts optiques liés au scanner. Dans le cas des puces à ADN utilisant des ADNc, le bruit de fond peut directement être apprécié par l'intensité du signal mesuré entre les *spots*. Dans le cas des puces Affymetrix, cette méthode n'est pas applicable. Le bruit de fond doit être estimé avant d'être pris en compte pour corriger les valeurs d'intensités mesurées. A cette fin, différentes méthodes ont été développées. Deux seront mentionnées ici : MAS 5 et RMA.

Correction du bruit de fond par la méthode MAS 5

La puce est divisée en 16 zones. Pour chaque zone, le $2^{ème}$ centile des valeurs d'intensités (à l'exception des contrôles) est retenu comme valeur du bruit de fond. Pour chaque cellule, un bruit de fond local est ensuite calculé en fonction des bruits de fond des 16 zones et de la distance séparant la

sonde en question de chacune de ces 16 zones (Figure IVIV-3). Ce bruit de fond local est ensuite déduit de la valeur d'intensité mesurée. Cette méthode corrige à la fois les mismatch (MM) et les perfect match (PM).

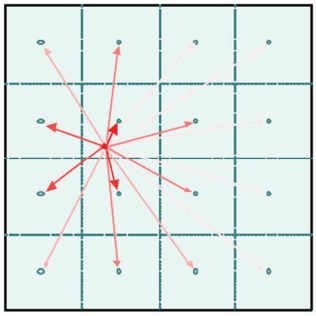


Figure IV-3 Correction du bruit de fond par la méthode MAS 5. La puce est divisée en 16 zones. Un bruit de fond est affecté au centre de chaque zone (points verts). Pour chaque cellule, un bruit de fond local est calculé comme la moyenne des bruits de fond des 16 zones pondérée par les distances séparant la cellule considérée de chacune des zones. L'intensité des flèches rouges indique le poids affecté à chaque zone.

Correction du bruit par la méthode RMA

RMA, pour Robust Multi-array Average, est un procédé de traitement des données brutes des puces à ADN alternatif à MAS 5, qui inclut une étape de traitement du bruit, mais également un procédé de normalisation des intensités d'un ensemble de puces et enfin une étape de résumé des intensités (Irizarry et al. 2003).

L'étape de correction du bruit est fondée sur l'utilisation d'un modèle statistique. Les valeurs mesurées pour chaque sonde (PM) sont décomposées en un signal plus un bruit de fond :

$PM_{ij} = B_{ij} + S_{ij}$ Avec : i = indice de la puce

j = indice de la sonde

B_{ij} composante bruit de fond de $PM_{ij} \sim N(\mu,\sigma^2)$

S_{ij} composante signal de $PM_{ij} \sim \exp(\alpha)$

Les valeurs S_{ij} sont ensuite estimées par ajustement du modèle sur les données générées. Cette méthode globale corrige uniquement les PM.

- **Normalisation des intensités par la méthode RMA**

L'étape de normalisation consiste à rendre identique la distribution des intensités des différentes puces (Figure IV-4). Elle peut être décrite par l'algorithme suivant :

1- Soit n puces comportant chacune un nombre p de PM, et soit la matrice X de dimensions pxn dans laquelle chaque puce correspond à une colonne.

2- Tri de chaque colonne de X pour donner $X_{triée}$.

3- Calcul de la valeur moyenne pour chaque ligne de $X_{triée}$ et affectation de cette moyenne à toute la ligne correspondante pour obtenir $X'_{triée}$.

4- Réorganisation de chaque colonne de manière à rétablir les rang initiaux dans X pour obtenir $X_{normalisée}$.

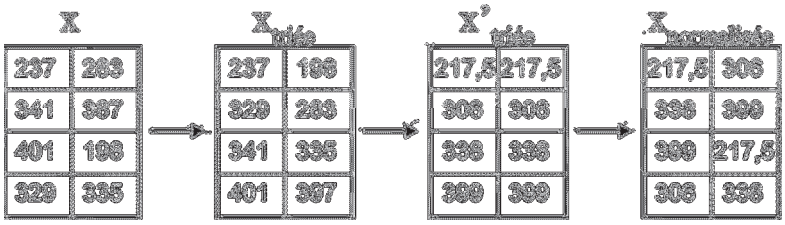


Figure IV-4 Exemple de normalisation par l'algorithme RMA dit des quantiles sur un jeu de données simplifié.

- **Résumé des intensités**

Il s'agit de l'étape, inclue dans RMA, qui regroupe les intensités des sondes en probesets. Elle est fondée sur l'utilisation d'un modèle statistique, prenant en compte l'effet sonde :

$$\log_2(PM_{ij}) = \theta_i + \phi_j + \varepsilon_{ij} \quad \text{avec} \quad \begin{cases} \theta_i \text{ représentant la mesure de l'expression sur la puce i} \\ \phi_j \text{ représentant l'effet de la sonde j} \\ \varepsilon_{ij} \text{ terme d'erreur} \end{cases}$$

L'ajustement du modèle aux données est réalisé à l'aide de la procédure *median polish.* Cet ajustement du modèle sur les données générées permet d'estimer une mesure de l'expression θ_i. Notons que le calcul des intensités résumées de chaque jeu de sonde prend en compte l'information de l'ensemble des jeux de sondes des puces étudiées.

- **Analyse exploratoire des données générées**

La visualisation des données permet d'avoir une première idée de leur structure et de la possibilité de définition expérimentale de différentes classes d'échantillons. Cela revient à appréhender de manière globale les échantillons testés à l'aide des données générées. Deux types d'approches permettent, notamment, une telle visualisation : la réduction de dimension et la classification non-supervisée (*clustering)* (Knudsen 2005).

Réduction de dimension ACP et MDS

Les expériences de puces à ADN génèrent des matrices de données de grandes tailles (quelques milliers de gènes *x* le nombre d'échantillons). L'inspection visuelle de telles matrices n'est pas réaliste et leur représentation graphique suppose un espace à plus de 3 dimensions. Cependant, il existe des méthodes mathématiques permettant de réduire le nombre de dimensions de manière à les rendre visualisables dans un espace réduit (à 2 ou 3 dimensions).

Analyse en Composantes Principales (ACP) : L'ACP crée des dimensions artificielles non corrélées telles que le maximum de la variabilité des données soit concentré dans un nombre réduit

de dimensions. L'ACP permet donc de regrouper les informations redondantes dans une seule composante. Ainsi, considérant que les premières dimensions de l'ACP traduisent un maximum d'informations sur les données, une représentation graphique de ces données pourra être obtenue en exprimant les valeurs de la première dimension (celle qui capture le maximum de variation) en fonction de la deuxième dimension (qui lui est orthogonale) et éventuellement d'une $3^{ième}$ dans le cas d'une représentation tridimensionnelle.

***Multi-Dimensional Scaling* (MDS) :** Le MDS est basé sur une matrice des distances (distance euclidienne ou distance fondée sur la corrélation) entre les objets, ici les puces. Il vise à représenter les objets dans un espace à dimension réduite tout en gardant un maximum d'informations sur les distances originales. Là encore, cette méthode permet de représenter les résultats obtenus dans un graphique à 2 ou 3 dimensions. Les graphiques ainsi obtenus peuvent faire apparaître des regroupements possibles entre les différentes puces (différents échantillons testés). Des méthodes de classification non-supervisée permettent ensuite de confirmer l'existence des classes suggérées.

Analyse non-supervisée (clustering)

Chaque puce peut être considérée comme un vecteur de dimension n (où n est le nombre de gènes étudiés). Dans un espace mathématique à n dimensions, il est possible de calculer la distance entre chaque vecteur. De la sorte, les puces (échantillons testés) peuvent être regroupées en fonction de leur proximité. Dans le cadre d'une approche exploratoire des données, ceci peut permettre de mettre en évidence la qualité de réplicats techniques ou de suggérer des regroupements de puces (correspondant à différents échantillons) au sein de classes. De manière analogue, il est possible de procéder à un regroupement sur les gènes. En effet, il est possible de définir pour chaque gène un vecteur de dimension p (où p est le nombre de puces analysées). Dans le nouvel espace ainsi crée, les regroupements traduisent alors des phénomènes de co-expression de gènes ou d'éventuelles réactions croisées.

Classification hiérarchique ascendante : le principe de l'algorithme hiérarchique ascendant consiste, à chaque étape, à agréger deux à deux les éléments les plus proches jusqu'à ce que tous les éléments aient été pris en compte. Ces algorithmes ne fournissent pas une partition en k groupes d'un ensemble de n individus mais une hiérarchie de groupes, se présentant sous la forme d'un arbre appelé également *dendrogramme*.

Classification par partitionnement : la finalité des méthodes par partitionnement consiste à construire k groupes d'éléments. Le nombre k de groupes est fixé *a priori* par l'utilisateur. La partition dont la qualité est la plus grande est celle qui minimise les distances intra-classes et qui maximise les distances inter-classes. Après une initialisation aléatoire des k centres de groupes, chaque individu est affecté au groupe dont il est le plus proche. Ensuite, les centres des k groupes

ainsi constitués sont calculés. De manière itérative, les individus sont réaffectés au groupe le plus proche et ainsi de suite jusqu'à stabilisation des groupes.

- **Tests statistiques et mise en évidence de gènes différentiellement exprimés**

Différents tests statistiques, paramétriques ou non, permettent de tester l'égalité des moyennes de deux groupes (Student, Welsh, Wilcoxon, etc.). Un test statistique est une méthode permettant un choix objectif entre deux hypothèses aux vues des résultats d'une expérience. Dans le cas de l'analyse des données d'expression, on définira :

H_{0i} « le gène i a le même niveau d'expression dans les deux groupes », c'est l'hypothèse nulle.

H_{1i} « le gène i n'a pas le même niveau d'expression dans les deux groupes », c'est l'hypothèse alternative.

Le risque α correspond à la probabilité de rejeter à tort H_{0i} (erreur de type I ou risque de première espèce), c'est donc le risque de faux positif. La valeur de 1-α définit la confiance du test. Le risque β correspond à la probabilité d'accepter à tort H_{0i} (erreur de type II ou risque de deuxième espèce). La valeur 1-β définit la puissance du test, c'est-à-dire sa capacité à identifier un gène différentiellement exprimé comme tel. Seul le risque α est contrôlable *a posteriori*, le contrôle du risque β intervient lors de la conception du plan d'expérience. Au terme de l'expérience, les quatre cas de figure suivants peuvent se présenter :

Décision / Réalité	H_{0i} non rejetée	H_{0i} rejetée
H_{0i} vraie	1-α	α
H_{1i} vraie	β	1-β

Tableau IV-1 Risques associés aux tests d'hypothèses.

Dans le cas de l'étude de l'expression différentielle de m gènes entre deux groupes, m tests statistiques doivent être effectués. Une procédure de test pourrait consister à effectuer m tests de seuil α pour chacune des hypothèses. Ce type de stratégie conduirait à rejeter à tort un nombre important d'hypothèses nulles, augmentant de façon proportionnelle au nombre d'hypothèses testées. Par exemple, avec 20 000 gènes testés vérifiant l'hypothèse nulle, 1 000 gènes seraient identifiés à tort comme différentiellement exprimés au seuil α = 5 %. Ceci n'est évidemment pas acceptable. Les tests statistiques multiples doivent donc être corrigés pour minimiser le nombre de faux positifs. Deux types d'approches principales ont été développées : les procédures contrôlant le FWER (*Family Wise Error Rate)* et les procédures contrôlant le FDR (*False Discovery Rate)* :

Le **FWER** mesure la probabilité de rejeter à tort au moins l'une des hypothèses nulles testées (parmi toutes les hypothèses testées). Le FWER est un critère assez conservatif permettant

d'identifier peu de gènes différentiellement exprimés. Contrôler le FWER au seuil α = 5 % permet d'être confiant à 95% de n'avoir aucun faux positif. On notera comme exemple la procédure de Bonferroni qui effectue m tests de seuil $\alpha^* = \alpha/m$, dont les p-valeurs ajustées sont alors $p_j^* = mp_j$.

Le **FDR** est l'espérance du taux de faux positifs. Il estime la proportion des erreurs de type I parmi les hypothèses rejetées. Le FDR est un critère moins conservatif que le FWER, il est donc plus adapté à une approche exploratoire. Une extension du FDR, le pFDR (*Positive False Discovery Rate*), est défini comme l'espérance du taux de faux positifs conditionné à l'existence d'au moins un rejet. Une procédure contrôlant le pFDR est la méthode **SAM** (*Significance Analysis of Microarrays*) (Tusher et al. 2001).

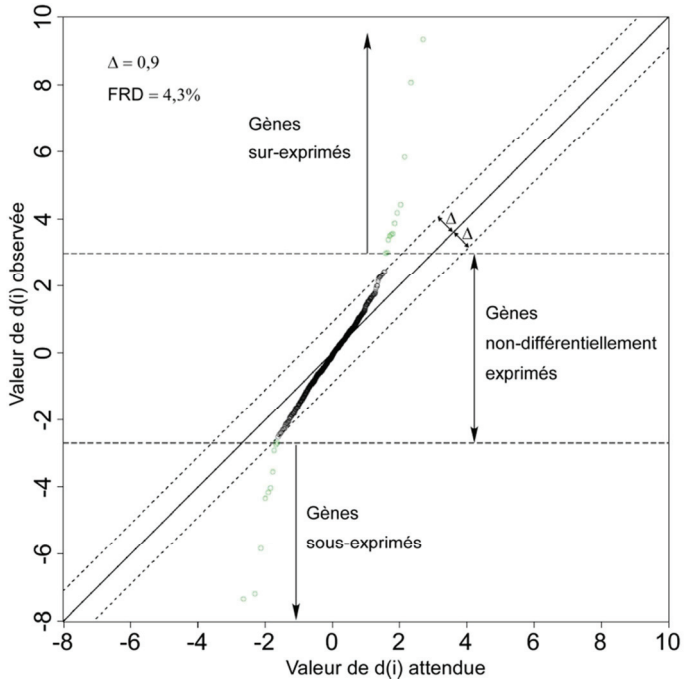


Figure IV-5 Graphique SAM. recherche de gènes différentiellement exprimés. Les cercles verts représentent les gènes différentiellement exprimés (au seuil Δ).

Cette méthode utilise un test de Student modifié comme test statistique et des jeux de données générés par permutations aléatoires des données d'origine entre les différents groupes. Pour chaque gène i, les données d'origine servent à calculer une valeur observée de la statistique de test $(d(i)_{obs})$ et les données générées par permutations servent à calculer une valeur moyenne attendue de la statistique de test $(d(i)_{att}$, sous-entendu la valeur que l'on obtiendrait si les différences observées étaient liées au hasard). Les gènes pour lesquels la différence entre la valeur de test observée et la valeur moyenne attendue dépasse un seuil Δ, sont considérés comme différentiellement exprimés (Storey and Tibshirani 2003) (Figure IV-5). C'est le choix du seuil Δ qui permet de contrôler le taux de faux positifs. Il est important de noter que la méthode SAM suppose une indépendance des événements (gènes) étudiés.

IV.3. Compléments techniques relatifs à l'étude de faisabilité

- **Incidence de la quantité de cDNA hybridée sur les niveaux d'intensités détectés sur puce**

Nous avons cherché à évaluer l'incidence de la quantité de cDNA hybridée sur une puce à ADN dans le cadre des difficultés méthodologiques à obtenir systématiquement la quantité requise par le protocole standard (2 µg) lorsque l'on travaille à partir d'échantillons urinaires complexes. Cinq concentrations de cDNA ont été préparées à partir des produits d'amplification de l'échantillon 670 B (cf. II.4.3.1) : 250 ng, 500 ng, 1 µg, 1,5 µg et 2 µg. Ces différentes quantités d'acides nucléiques ont été fragmentées, marquées et hybridées sur des puces HG-U133-PLUS2. Il n' y a pas eu de normalisation des données brutes. Nous avons extrait certains des signaux correspondant à des gènes d'intérêt, et représenté les intensités de ces gènes en fonction de la quantité hybridée sur les puces (Figure IV-6).

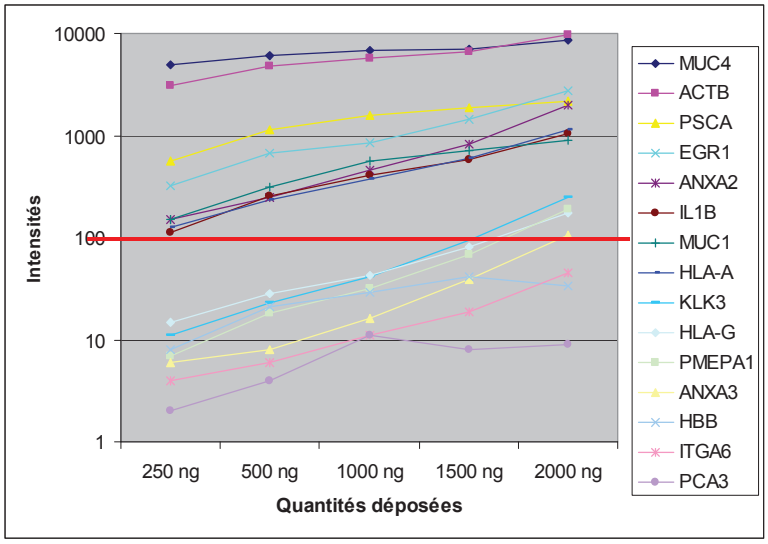


Figure IV-6 Intensités de détection de 15 gènes en fonction de la quantité de cDNA hybridée sur puce. Une échelle logarithmique est utilisée pour les intensités, en ordonnées. La barre horizontale rouge indique le seuil du bruit de fond.

On observe qu'une diminution des valeurs d'intensité suit la dilution de cDNA. Si l'on fixe la barre du bruit de fond à une valeur d'expression de 100, certains gènes ne sont alors plus détectés en hybridant moins de 2 µg (KLK3, HLA-G, PMEPA1), mais cela concerne des gènes qui présentaient un niveau d'expression peu élevé (entre 100 et 250) en conditions standard de 2 µg. Il semble donc envisageable d'hybrider moins de 2 µg sur une puce (en ce sens que l'expérience fonctionne), mais il faut s'attendre à ce que tous les signaux les plus faibles disparaissent, ce qui dans notre étude sur la prostate peut s'avérer particulièrement dommageable par la perte des signaux du PSA (KLK3). Nous avons donc cherché à assurer la quantité recommandée de 2 µg en réalisant, pour un échantillon donné, des amplifications en parallèle puis en les mélangeant.

- **Incidence du mélange de deux produits d'amplication WTO Nano sur la détection d'un signal sur puce**

Comme indiqué dans le paragraphe précédent, la disparité des rendements d'amplification obtenus à partir de prélèvements urinaires a soulevé la possibilité de conduire en parallèle deux amplifications WTO Nano à partir d'un même échantillon, dans le but de disposer d'une quantité suffisante pour être hybridée sur une ou plusieurs puces (par exemple une puce HG-U133-PLUS2 et une puce HERV-V2). Afin d'évaluer l'incidence d'un mélange de produits d'amplification, trois amplifications ont été conduites en parallèle, à partir du même échantillon biologique de départ. Puis deux puces HG-U133-PLUS2 ont été hybridées et lues. La puce 1 a été réalisée à partir d'une unique amplification, la puce 2 à partir du mélange de deux produits d'amplification ramené à la concentration de la puce 1. Aucune normalisation n'a été appliquée aux données brutes. Les écarts relatifs par probeset ont été calculés en prenant la puce 1 comme référence. La représentation des écarts relatifs en fonction de l'intensité du signal du référence est donnée sous forme de nuage de points (Figure IV-7).

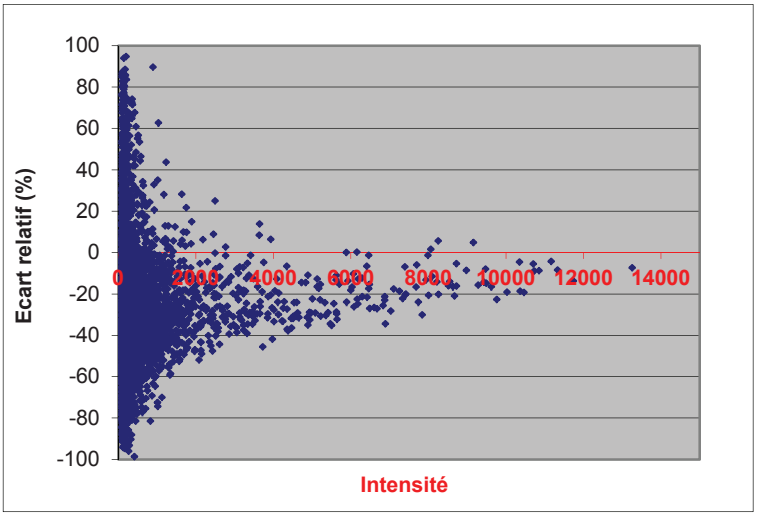


Figure IV-7 Nuage de points des écarts relatifs en fonction de l'intensité de signaux.

Sur cette expérience, on constate que faire un pool de deux produits d'amplification WTO Nano impacte la valeur des signaux par rapport au protocole de référence (puce 1). Par exemple, un signal à 2 000 (considéré comme fort) sur la puce de référence peut être modifié d'une amplitude de –50 % à + 30 %. On note également que la dispersion est orientée majoritairement vers des écarts négatifs par rapport à la référence, autrement dit vers une diminution du signal, quand on réalise un pool. D'une manière générale, les petites valeurs sont très touchées, pouvant varier du simple au double. Pour aller plus loin dans la description, la figure IV-8 donne la proportion de signaux par tranches de variations (0-10%, 10-20% etc) et tronçons d'intensités (100-200 , 200-500 etc.).

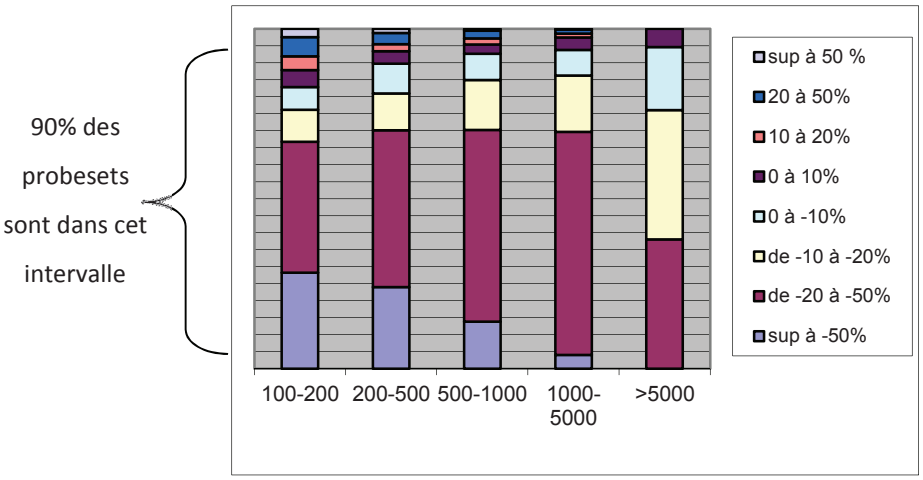


Figure IV-8 Répartition des variations de détection par tranches d'intensités de signaux. Les variations par tranches (eg : >50%, de 20% à 50% etc.) sont représentées pour 5 intervalles d'intensités (100-200 ; 200-500 ; 500-1000 ; 1000-5000 et > 5000). L'espace entre chaque ligne horizontale noire contient 5% des effectifs.

Les variations extrêmes (> +/- 50 %) se réduisent et disparaissent en allant vers des signaux forts. En revanche la convergence vers une variation nulle (tranches turquoise et bordeaux) reste partielle. La question a donc été de savoir si de telles variations sont problématiques ou non dans la détection qualitative de signaux d'expression. En considérant que ce qui est en-dessous du bruit de fond est systématiquement exclu des analyses de puces à ADN, nous avons évalué le risque de passer sous le seuil du bruit de fond dans le cas d'un mélange de produits d'amplification, par rapport à un protocole de référence (puce 1). Le graphique (Figure IV-9) suivant montre grossièrement (5 points de gamme seulement, échelle non linéaire) les amplitudes maximales possibles par rapport à la référence (mean, courbe bleue). Les points des courbes high (jaune) et low (rose) sont obtenus si l'on retient l'intervalle de 90% des probesets du graphique précédent.

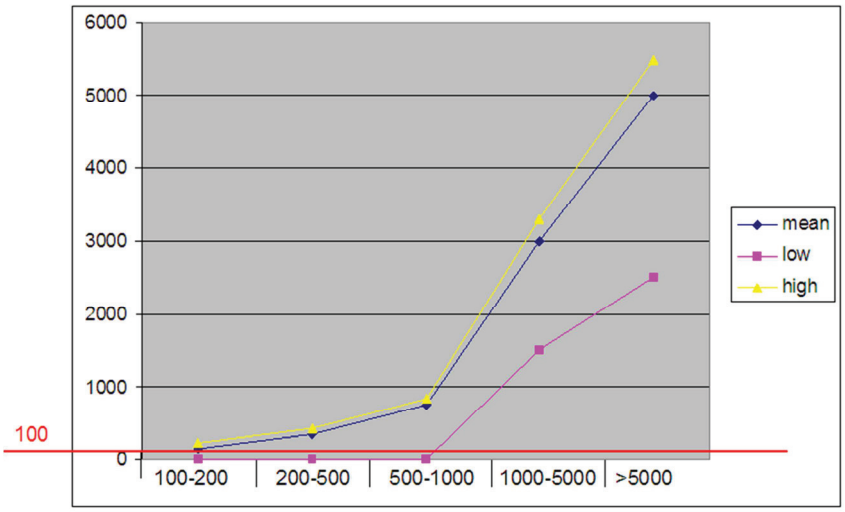


Figure IV-9 Estimation de l'amplitude d'erreur de détection induite par un mélange d'amplifications WTO Nano. En ordonnées : nombre de probesets ; en abscisse : intervalles d'intensités.

Les signaux «perdus» par l'action du mélange sont ceux qui passent sous la barre du bruit de fond (100). On peut donc évaluer le risque de perdre un signal par le rapport de deux surfaces : l'aire au-dessus de la ligne rouge et en dessous de la ligne bleue (les signaux détectés par la puce 1) et l'aire en dessous de la ligne rouge et au-dessus de la ligne rose (non détectés à tort par la puce 2). Ce rapport est de l'ordre de 5%. Ainsi, malgré une assez forte variabilité générale des données entre la puce 1 et la puce 2 utilisée en situation de mélange de produits d'amplification, seule une petite proportion de signaux (5 %) se perd, et cela concerne exclusivement des signaux de très faible à moyennement forte intensité (jusqu'à une intensité de 500). Un signal fortement détecté par le protocole de référence l'est toujours en réalisant un mélange d'amplifications.

En conclusion de ces deux compléments techniques, il apparait qu'en abaissant la quantité de cDNA hybridée sur une puce, de 2 µg à 1,5 µg (ou moins), tous les signaux de faible valeur se perdent, alors qu'en réalisant un mélange de deux amplifications WTO Nano pour assurer artificiellement une quantité de 2 µg, la perte des signaux de faible valeur se limite à 5 % environ. Nous avons donc fait le choix de réaliser plusieurs amplifications parralèles par échantillon dans l'étude transcriptomique clinique.