Rétrovirus endogènes humains et réponse immunitaire de l'hôte suite à une agression inflammatoire

Olivier Tabone

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

Ecole Doctorale N° 205
Ecole doctorale interdiscplinaire Sciences-Santé

Spécialité de doctorat : Biologie
Discipline : Bioinformatique

Soutenue publiquement le 31/01/2019, par :
Olivier Tabone

RETROVIRUS ENDOGENES HUMAINS ET REPONSE IMMUNITAIRE DE L'HOTE SUITE A UNE AGRESSION INFLAMMATOIRE.

Devant le jury composé de :

Dr Benkirane, Monsef, DR CNRS IGH Montpellier	Rapporteur
Dr Quesneville Hadi, DR INRA Versailles	Rapporteur
Dr Buseyne, Florence, CR Institut Pasteur Paris	Examinatrice
Pr Vieira-Heddi, Cristina, Université Claude Bernard	Examinatrice
Dr Textoris, Julien, Hôpital Edouard-Herriot	Directeur de thèse
Dr Mallet, François, LCR HCL-bioMérieux	Co-directeur de thèse

TABLE DES MATIERES

TABLE DES ILLUSTRATIONS

LISTE DES ABREVIATIONS

CLR	C-type Lectin Receptors
CMH	Complexe majeur d'histocompatibilité
CMV	Cytomegalovirus
DAMP	Danger-Associated Molecular Patterns
EBV	Epstein Barr Virus
eQTL	expression Quantitative Trait Loci
ESTs	Expressed Sequence Tags
ET	Jeu de données de la tolérance à l'endotoxine
FC	Fold Change
FDR	False Discovery Rate
HAI	Hospital Acquired Infection
HERV	Human Endogenous Retroviruses
HLA	Human Leukocyte Antigen
HMM	Hidden Markov Model
HV	Healthy Volunteer
IFN	Interféron
IS	Jeu de données ImmunoSepsis
LINE	Long interspersed nuclear elements
LPS	LipoPolySaccharide
LTR	Long Terminal Repeat
MaLR	Mammalian-Apparent Long-Terminal Repeat Retrotransposon
MIP	Jeu de données MIP
MLV	Murin Leukemia Virus
NGS	Next Generation Sequencing
PAMP	Pathogen-Associated Molecular Pattern
Pb	Paire de base
PBMC	Peripheral Blood Mononuclear Cell
PBS	Primer Binding Site
PMA	Phorbol Myristate Acetate
PRR	Pattern-Recognition Receptors
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
SEP	Sclérose en Plaques
SINE	Short interspersed nuclear elements
SNP	Single Nucleotide Polymorphism
SOFA	Sequential Organ Failure
TLR	Toll Like Receptors
UTR	Untranslated Region
VAF	Variant Allele Frequency
VIH	Virus de l'Immunodéficience Humain

1 CONTEXTE ET ETAT DE L'ART

1.1 SEPSIS ET AGRESSIONS INFLAMMATOIRES

La réanimation est une discipline médicale prenant en charge des patients graves, instables, avec un pronostic vital menacé par une ou plusieurs défaillances d'organe. Les causes d'admission sont multiples – Crise cardiaque, accident vasculaire cérébral, pneumonie, empoisonnement, complications suite à une chirurgie lourde, traumas majeurs et brulures graves, … - Une proportion considérable parmi ces patients sont admis à cause d'un sepsis ou en développent durant leur séjour : 30% des patients d'après l'audit international ICON (Intensive Care Over Nations) (Vincent et al., 2014). Le sepsis est aujourd'hui défini comme un dysfonctionnement d'organes mortel causé par une réponse de l'hôte dérégulée à une infection (Singer et al., 2016). Cette réponse inappropriée se traduit par des dérégulations du métabolisme (Singer et al., 2004), du système neuroendocrinien (Deutschman and Tracey, 2014) et du système immunitaire (Hotchkiss et al., 2013).

Cliniquement, elle se traduit par une augmentation d'au moins deux points du score « Sequential Organ Failure » (SOFA). Le patient doit avoir au moins deux des atteintes suivantes : une fréquence respiratoire supérieure ou égale à 22 respirations par minute, une altération de l'état mental ou une pression systolique inférieure ou égale à 100 mmHg. Le choc septique, forme la plus grave du sepsis, pour laquelle les altérations cellulaires et métaboliques et la défaillance circulatoire sont plus importantes et entrainent un risque plus élevé de mortalité que le sepsis. En clinique, le choc septique nécessite l'utilisation de vasopresseurs afin de maintenir une pression artérielle supérieure à 65 mm Hg et présente un niveau de lactate dans le sérum supérieur à 2mmol/L malgré un remplissage vasculaire suffisant (pas d'hypovolémie).

Le sepsis est un problème majeur de santé mondiale. L'incidence des syndromes septiques est élevée, on estime que plus de 30 millions de patients développent un sepsis et 19 millions un sepsis sévère chaque année dans le monde (Fleischmann et al., 2016). En France, la fréquence du choc septique est en augmentation croissante ces 20 dernières années, passant de 8,2 cas d'admission sur 100 en réanimation en 1993 à 15,4% en 2010

(Quenot et al., 2015). Bien qu'il existe une grande disparité entre les pays et les différents services de réanimation, en moyenne sur 730 centres répartis dans plus de 84 pays, le taux de mortalité à l'hôpital dans la population générale est estimé à 22.4%, alors qu'il est de 35.3% chez des patients avec un sepsis (Vincent et al., 2014). En plus de la mortalité observée dans les jours qui suivent la survenue du sepsis, des études ont évalué l'impact du sepsis sur la mortalité à long terme. Sur plus de 30 000 adultes, le taux de survie des patients ayant eu un sepsis était de deux à cinq fois plus faible que ceux n'ayant pas eu de sepsis dans les 6 ans suivant le sepsis (Wang et al., 2014). Les survivants souffrent souvent de morbidités comme un risque plus important d'être ré-hospitalisé, des maladies cardio-vasculaires, des déficiences cognitives (Iwashyna et al., 2010; Shankar-Hari and Rubenfeld, 2016). Les coûts engendrés sont très importants et variables selon la gravité du syndrome et la durée du séjour en service de réanimation. Aux États-Unis, le coût de prise en charge des patients septiques et en choc septique est estimé à 20 milliards de dollars, soit 5,2% des coût totaux des hôpitaux américains, faisant ainsi la condition la plus couteuse à traiter en hôpital (Torio and Andrews, 2006). En France en 2003, le coût par patient de sepsis était estimé à 25 000 €, et montait à plus de 35 000 euros pour un patient atteint de sepsis sévère (Brun-Buisson et al., 2003).

Les agents infectieux responsables d'un sepsis peuvent être d'origine bactérienne (Gram négatifs ou Gram positif), fongique ou encore virale. Les germes Gram négatifs sont les plus courants (Friedman et al., 1998). Pour lutter contre ces pathogènes, les antibiotiques sont utilisés, ceux ciblant spécifiquement le pathogène induit une diminution de la mortalité plus importante que l'utilisation d'antibiotiques à large spectre (Paul et al., 2010). L'identification de l'agent pathogène est donc primordiale pour prodiguer un traitement adapté au patient. Cependant, seulement 40 à 60% des patients en choc septique présentent une hémoculture positive (de Prost et al., 2013). Plusieurs raisons peuvent l'expliquer, notamment l'administration d'antibiotiques large spectre en amont de l'hémoculture pouvant fausser ses résultats, un bilan diagnostique incomplet ou encore une origine infectieuse par des organismes inhabituels et difficilement identifiables.

Bien que le sepsis puisse atteindre toutes classes de la population, l'âge et le sexe sont des facteurs de risque et de gravité du sepsis. Les hommes ont environ 1,3 fois plus de risque

de développer un sepsis que les femmes, et les personnes âgées de plus de 65 ans ont 13 fois plus de risque d'en développer un que les plus jeunes (Martin et al., 2006). Les enfants prématurés ont également un risque plus important de développer un sepsis néonatal (Hoogen et al., 2010). Les maladies chroniques ont aussi un impact sur le devenir du patient. Par exemple, les personnes atteintes du VIH, bien que n'ayant pas une prévalence supérieure au reste de la population, ont un taux de mortalité bien plus élevé que les patients septiques non séropositives (Mrus et al., 2005). On peut également citer les patients diabétiques, cirrhotiques ou cancéreux pour lesquels la prévalence du sepsis est plus élevée (Danai et al., 2006; Esper et al., 2009; Galbois et al., 2014).

Les consignes de prise en charge précoce des patients septiques, dont la dernière mise à jour date de janvier 2017, contiennent 93 préconisations, dont 32 sont des recommandations fortes (Rhodes et al., 2017). Dans les trois premières heures après l'arrivée du patient, le patient se voit prélevé pour les cultures bactériennes, mesuré son lactate et administré des antibiotiques à large spectre. L'administration d'antibiotiques précoce permet le contrôle de la source infectieuse et est associé à un risque diminué de mortalité par rapport à une administration plus tardive (Ferrer et al., 2014). Le sepsis induit une hypo-perfusion des tissus qui sont alors moins alimentés en oxygène et nutriments. Il faut alors apporter un soutien hémodynamique au patient, par l'administration de solutions cristalloïdes ou colloïdes de remplissage et éventuellement, si le remplissage ne suffit pas, par l'administration de vasopresseurs afin de maintenir une pression artérielle suffisante. L'hémofiltration est couramment utilisée afin d'éliminer les médiateurs de l'inflammation et toxines bactériennes réduisant ainsi la réponse pro-inflammatoire excessive (Rimmelé and Kellum, 2012), et est adaptée en cas d'insuffisance rénale. L'hyperglycémie est très fréquente chez ces patients, et doit être surveillée et maintenue à un taux inférieur à 180mg/dL , tout en maintenant une concentration minimale de glucose dans le sang. Pour compenser l'insuffisance rénale souvent observée chez les patients, certains réanimateurs administrent des corticostéroïdes à faible dose. Cependant, ce traitement reste controversé et des études récentes ont montré des résultats contradictoires de l'effet de ce traitement sur la mortalité à 90 jours, bien que des effets bénéfiques à court terme sont clairement identifiés dans les deux études (Annane et al., 2018; Venkatesh et al., 2018).

A travers l'exemple du sepsis, représentant l'un des syndromes les plus grave et les plus courant en réanimation et pourtant encore relativement méconnu du grand publique, on montre ici le besoin de mieux comprendre la physiopathologie de ces réponses exacerbées faisant suite à des agressions inflammatoires. Cela permettra de fournir au réanimateur tous les outils nécessaires pour un diagnostic adapté et une prise en charge efficace des patients.

1.1.1 PHYSIOPATHOLOGIE

1.1.2 REPONSE IMMUNITAIRE DE L'HOTE LORS DU SEPSIS

Les dérégulations du système immunitaire observées lors du sepsis, ou d'autres agressions inflammatoires comme des brulures, un traumatisme important ou à la suite d'une chirurgie lourde se traduit par une réponse pro-inflammatoire, responsable des défaillances d'organe et de la mortalité et dans un même temps, d'une réponse d'immunosuppression, expliquant la survenue d'infections secondaires et de la mortalité plus tardive (Figure 1-1) (Hotchkiss et al., 2013).

Rapidement, l'organisme de l'hôte répond à l'infection par la reconnaissance de l'agent pathogène (virus, bactérie, champignon), via son système immunitaire inné. Il reconnaît des PAMP (Pathogen-Associated Molecular Pattern), qui sont des motifs moléculaires conservés chez les micro-organismes, et absents chez l'hôte (Medzhitov, 2007). Parmi ceux-ci on trouve, entre beaucoup d'autres, le LPS (lipopolysaccharide), composant des bactéries Gram négatives, l'acide lipotéichoïque et le peptidoglycane qui composent les bactéries Gram positives. Pour les virus, c'est plutôt l'ADN ou l'ARN viral qui est reconnu (Kawai and Akira, 2010). Dans le cas d'atteintes non infectieuses, comme des dommage au niveau des tissus, les DAMP (Danger-Associated Molecular Patterns) ou alarmines sont relâchées par les cellules endommagées et peuvent initier la réponse immunitaire inflammatoire (Rubartelli and Lotze, 2007).

La reconnaissance des PAMP est médiée par les PRR (Pattern-Recognition Receptors), qui sont des récepteurs principalement exprimés par les cellules de l'immunité innée, dont les cellules dendritiques, les macrophages, les monocytes, les neutrophiles et les cellules épithéliales. Ces récepteurs peuvent être classés en deux groupes, ceux au niveau de la membrane cellulaire, comme les TLR (Toll Like Receptors) et les CLR (C-type Lectin

Receptors), et ceux présents dans le cytoplasme, comme les NLR (NOD-Like Receptors) et les RLR (RIG-I-Like-Receptors).

L'activation de ces récepteurs va enclencher en cascade des réponses inflammatoires, telles que la libération de cytokines (TNF-α, IFN type I, IL-6, IL1, IL18, …), de chimiokines et autres médiateurs de l'inflammation. On parle d'orage cytokinique quand leur sécrétion massive provoque un état hyper-inflammatoire. Ces médiateurs permettent le recrutement des cellules phagocytaires au niveau du site de l'infection et la présentation de l'antigène aux lymphocytes pour initier la réponse immunitaire adaptative. L'utilisation précoce d'anti-inflammatoires après la survenue du choc a montré un effet bénéfique chez les patients les plus graves, tout en étant un facteur aggravant chez ceux présentant un plus faible risque de mortalité (Eichacker et al., 2002). Cela montre ici l'importance pour le praticien de pouvoir identifier au mieux l'état de gravité du patient afin d'adapter son traitement. Si le patient survit à cette phase hyper-inflammatoire initialement dominante et ne se rétabli pas totalement, il doit ensuite faire face à l'immunosuppression qui s'installe et devient dominante.

1.1.3 IMMUNODEPRESSION INDUITE PAR L'AGRESSION INFLAMMATOIRE

Il n'existe pas de définition précise de la phase d'immunosuppression, bien que cette phase d'immunosuppression soit décrite depuis plusieurs années, et son importance dans la physiopathologie du sepsis n'a été admise que récemment (Hotchkiss et al., 2013). Cet état est aujourd'hui décrit par plusieurs phénomènes, montrant une incapacité globale du système immunitaire à répondre à de nouvelles agressions et est associé à une survenue d'infections secondaires (ou nosocomiales) (Grimaldi et al., 2014; Lukaszewicz et al., 2009) et à un pronostic péjoratif (van Dissel et al., 1998; Kollef et al., 2008).

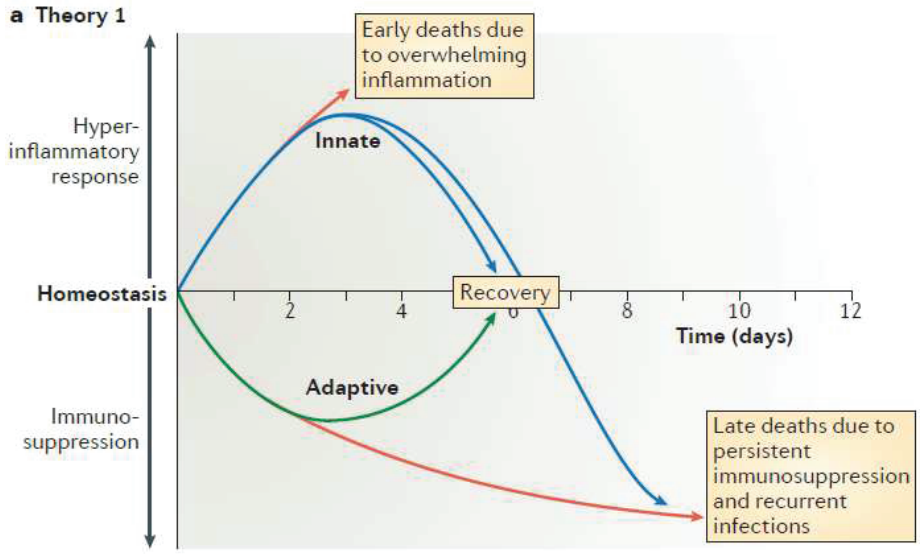


Figure 1-1: Modèle théorique de la réponse de l'hôte après un sepsis. Le patient répond d'une part par une réponse pro-inflammatoire responsable des défaillances d'organes et de la mortalité précoce. En parallèle, une réponse compensatrice anti-inflammatoire s'enclenche, favorisant les infections secondaires et des décès tardifs. D'après (Hotchkiss et al., 2013).

On observe lors de cette phase, des infections par des pathogènes peu virulents (Kollef et al., 2008; Otto et al., 2011) et une réactivation de virus latents (Papazian et al., 1996; Textoris and Mallet, 2017; Walton et al., 2014). Des altérations au niveau des cellules de l'immunité adaptative sont également décrites, telles qu'une diminution du nombres de lymphocytes (lymphopénie) et augmentation de polynucléaires neutrophiles immatures (Boomer et al., 2011; Demaret et al., 2015; Drewry et al., 2014; Pillay et al., 2012), une altération des lymphocytes jouant sur leur fonctionnalité (Venet et al., 2009), une perte de leur capacité de prolifération (Venet et al., 2012). On observe également une différence dans le ratio de cytokines pro-inflammatoire et anti-inflammatoire (IL10/TNF) (van Dissel et al., 1998) et l'expression de molécules de co-inhibition (ACTL-4, PD-1, PDL1, etc…) (Huang et al., 2009). Enfin, on retrouve une diminution des capacités de présentation de l'antigène, qui se traduit par une diminution de l'expression d'HLA-DR à la surface des monocytes (Monneret et al., 2004, 2006; Pachot et al., 2005). Le HLA-DR est un complexe majeur d'histocompatibilité de classe II faisant partie du système HLA, et présente les antigènes aux lymphocytes T. Cette diminution n'est pas retrouvée seulement dans le sepsis, mais aussi après des brulures graves (Venet et al., 2007) ou un traumatisme important (Meakins et al., 1977; Timmermans et al., 2016). Dans cette dernière étude, les auteurs montrent que des réponses anti-inflammatoires compensatrices de l'inflammation s'initient dans les minutes

qui suivent le traumatisme. Cela confirme qu'il existe très tôt une balance entre réponse pro et anti-inflammatoire, qui peut se dérégler, pouvant ainsi entrainer un déséquilibre dans la réponse de l'hôte. Ces réponses dérégulées empêchent un retour à l'homéostasie et font perdurer un état d'immunodépression.

Pour « rebooster » le système immunitaire, des approches thérapeutiques immunostimulantes paraissent donc prometteuses et plusieurs essais cliniques sont en cours, utilisant l'IFNγ, le GM-CSF, l'IL7 ou encore des inhibiteurs des molécules de co-inhibition.

Il apparait primordial de mieux comprendre la physiopathologie suivant ces agressions inflammatoires pour améliorer la prise en charge des patients. Pour cela, il y a un grand besoin en biomarqueurs permettant de diagnostiquer l'infection pour guider la prise en charge thérapeutique (notamment antibiotique), de pronostiquer le devenir du patient (défaillance d'organe, infections secondaires, mortalité) et de stratifier les patients selon l'évolution de leur réponse à l'infection et au traitement. A ce jour, malgré tous les efforts de la communauté scientifique, les biomarqueurs existants ne sont pas suffisants.

1.1.4 MODULATION DU TRANSCRIPTOME DANS LE SANG

Dans ces contextes, le transcriptome dans le sang est fortement modulé. Plusieurs études mettent en évidence des modulations après le sepsis (Scicluna et al., 2017), des brulures graves (Plassais et al., 2017) ou des blessures importantes (Laudanski et al., 2006). Ces études montrent une modulation globale du transcriptome sanguin, dont une partie peut persister plusieurs jours / semaines après la survenue de l'agression. On retrouve principalement des gènes impliqués dans la réponse immunitaire. De manière intéressante, dans leur étude, Scicluna et collègues ont séparé en 4 des cohortes de patients de sepsis (ou choc septique) selon leur état transcriptionnel global. Les réseaux fonctionnels de gènes modulés par rapport à des volontaires sains sont souvent les mêmes entre les 4 endotypes, mais dans certains groupes, ils peuvent être modulés positivement et dans d'autres ils sont modulés négativement. Cela confirme l'importance d'avoir à disposition des outils pour évaluer l'état du système immunitaire, de gravité, de risque de mortalité du patient.

Des modèles in-vitro mimant les réponses pro-inflammatoire et d'immunodépression de l'hôte sont également utilisés afin de mieux comprendre les mécanismes mis en jeu. Le

modèle de tolérance à l'endotoxine permet notamment de mimer l'immunodépression induite par le sepsis observée dans les monocytes de patients en choc septique (Allantaz-Frager et al., 2013). A partir de ces étude de transcriptome, différents gènes ont pu être identifiés comme potentiels biomarqueurs, voire potentiellement impliqués dans la réponse immunitaire disproportionnée. Entre autres, BPGM et TAP2 identifiant un des 4 endotypes de patients septiques dans l'étude de Scicluna et al. précédemment évoquée (Scicluna et al., 2017), diminution de l'expression de CX3CR1 comme prédicteur de mortalité chez les chocs septiques (Pachot et al., 2008), ou encore le ratio entre J3 et J1 de CD74, chaîne invariante du HLA-DR, comme prédicteur de la survenue d'infections secondaires. Cependant, aucun de ces marqueurs ne présente de performances suffisantes pour être utilisé par le clinicien. Il est en réalité très probable qu'une signature de plusieurs transcrits soit nécessaire pour évaluer assez finement l'état immunitaire du patient. D'autre part, des travaux ont récemment souligné l'importance des marques épigénétiques dans la mise en place et la persistance de cet état d'immunodépression, soit à partir d'un modèle de tolérance à l'endotoxine (Novakovic et al., 2016), soit en « entrainant » / « préparant » l'immunité (Saeed et al., 2014). Ces travaux suggèrent fortement qu'il existe un lien entre modifications épigénétiques et anergie des cellules du système immunitaire.

Une difficulté d'étudier le transcriptome dans le sang est la complexité de ce tissu. Il comporte de nombreux type cellulaires, beaucoup de molécules circulantes, et il n'est ainsi pas toujours aisé d'identifier dans quel type cellulaire sanguin la modulation d'un gène s'effectue. Certains travaux montrent que différents états de différenciation monocytes-macrophages peuvent être identifiés à l'aide d'une signature transcriptomique spécifique et pourrait être utilisé comme biomarqueurs dans divers contextes cliniques (Becker et al., 2015; Xue et al., 2014). Néanmoins en recherche translationnelle, il est indispensable que le traitement des échantillons ne nécessitent pas trop d'étapes et à partir desquels, un résultat interprétable par le médecin puisse être délivré rapidement. Pouvoir interpréter le plus directement possible un échantillon sanguin est donc important.

Toutes ces études s'intéressent exclusivement au transcriptome des gènes (exome), alors que la part du génome codante pour des protéines représente seulement 1.5% de celui-ci. Il apparait de plus en plus important de ne plus se restreindre à ce répertoire mais d'explorer l'ensemble du transcriptome. Grâce aux technologies de séquençage ou à des outils

spécialement dédiés, de plus en plus d'études vont dans ce sens et analysent les ARN non codant, dont les éléments répétés du génome, notamment dans le sepsis (Ho et al., 2016). D'autres études s'intéressent aux associations entre le génome, quel que soit le répertoire, et le transcriptome, par exemple, toujours dans le sepsis, par l'intermédiaire d'analyses eQTL (expression Quantitative Trai Loci, (Davenport et al., 2016)).

Parmi ces éléments du génome, on propose de s'intéresser aux rétrovirus endogènes humain (HERV pour « Human Endogenous Retroviruses »). On cherche à identifier si les HERV sont toujours présents dans les génomes humains, s'ils sont exprimés dans les cellules du sang dans des contextes d'agressions inflammatoires, et si certains jouent un rôle sur l'expression du transcriptome dans ces contextes.

1.2 RETROVIRUS ENDOGENES HUMAINS

1.2.1 GENERALITES

Les HERV sont d'anciens rétrovirus qui se sont intégrés à notre génome. Les rétrovirus sont des virus à ARN simple brin qui insèrent une copie de leur génome dans l'ADN de la cellule hôte, changeant ainsi son génome. Une de leur particularités est qu'ils codent pour la transcriptase inverse, une polymérase qui rétro-transcrit l'ARN simple brin en ADN complémentaire double brin. Une fois entré dans la cellule hôte (via attachement et fusion des membranes), le virus rétro-transcrit son ARN en ADN double brin. Cet ADN rentre dans le noyau et s'intègre au génome de l'hôte via une intégrase, il devient alors un provirus, et partie intégrante du génome de la cellule infectée (Figure 1-2).

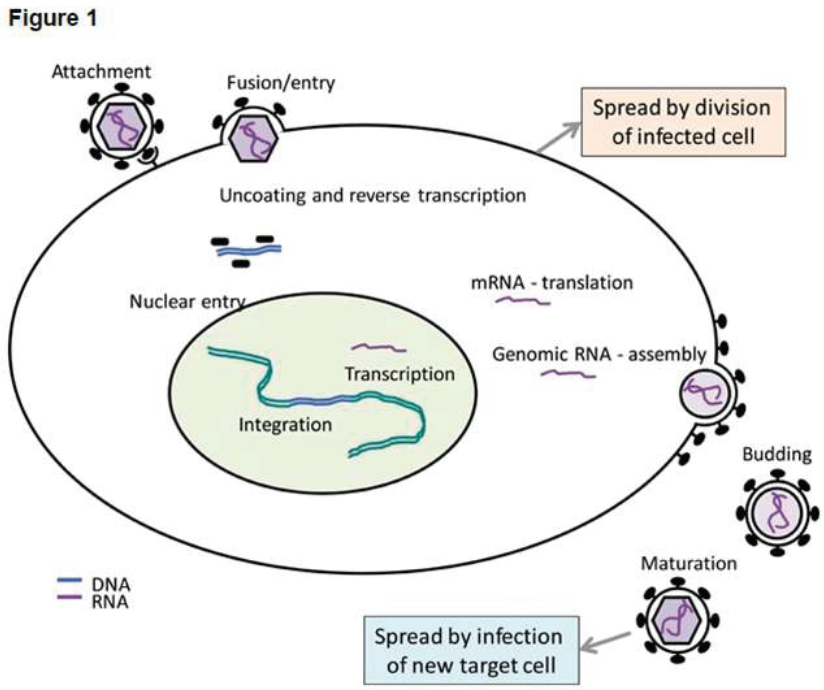


Figure 1-2: Cycle de vie du rétrovirus. Attachement et fusion du virion à la membrane. Rétrotranscription de l'ARN rétroviral dans le cytoplasme. Entrée dans le noyau et intégration du génome rétroviral dans le génome de l'hôte (provirus). Se comporte à ce moment-là de manière similaire à un gène endogène, avec expression d'ARN messager. Peut sortir de la cellule, forme un nouveau virion et après maturation peut infecter la prochaine cellule. D'après (Young et al., 2013).

Lorsque cet évènement se produit au niveau d'une cellule de la lignée germinale, et que son insertion n'est pas délétère pour la cellule, le matériel génétique du virus peut être transmis de manière verticale (ou mendélienne) à la génération suivante et peut ainsi se

fixer à la population (Figure 1-3). Ce processus est appelé endogénisation (Young et al., 2013). Les HERV partagent ainsi une structure similaires aux rétrovirus exogènes, avec des régions internes codantes pour des protéines rétrovirales (Gag, Pro, Pol, Env) entourées de 2 régions LTR identiques, allant de 300 à 1000 pb. Ce sont les régions de contrôle de l'expression des régions internes (Figure 1-6). Sous la forme de rétrovirus infectieux (ARN), la LTR 5' est composée des sous-régions R et U5, et la LTR 3' des sous-régions U3 et R, suivi d'une queue de polyadénylation (polyA). Sous sa forme provirale (ADN intégré au génome), les LTR, identiques, possèdent toutes deux les sous-régions U3, R et U5.

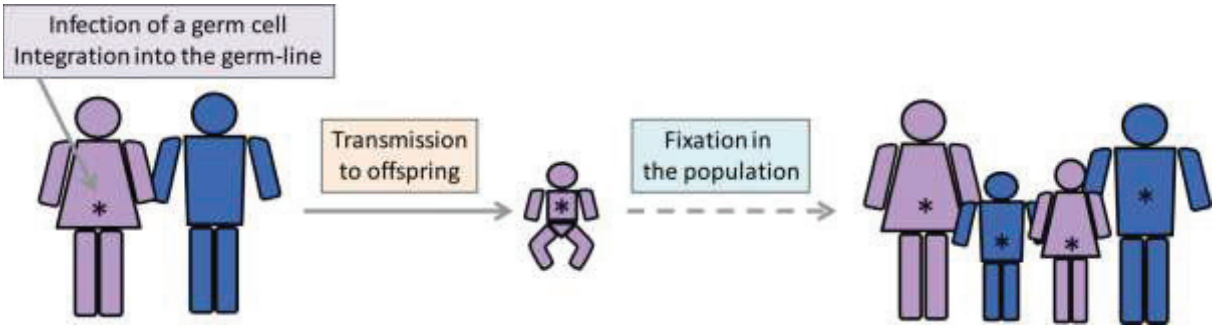


Figure 1-3 : Modèle de l'endogénisation rétrovirale. L'infection des cellules germinales par un rétrovirus peut induire son intégration au génome de l'hôte et être transmis de manière verticale à la descendance pour finalement être fixé dans la population. D'après (Young et al., 2013).

De tels évènements d'endogénisation sont arrivés de rares fois au cours de l'évolution, et ont donné naissance à plusieurs groupes (ou familles) de HERV (Bannert and Kurth, 2006). Il est cependant difficile d'identifier avec certitude le nombre d'évènements d'insertions qui se produit au cours de millions d'années d'évolution. Une fois intégrés, ces éléments se sont dupliqués dans le génome, en tant qu'élément transposable. Le génome humain en est composé à plus de 40%. Ils sont répartis en transposons de type I ou rétrotransposons, qui passent par un intermédiaire à ARN et se dupliquent via un mécanisme de « copier-coller », et les transposons de type II ou transposons à ADN, qui se déplacent par un mécanisme de « couper-coller ». Parmi les rétrotransposons, on distingue ceux qui possèdent une LTR et ceux qui n'en possèdent pas (SINE et LINE) (Figure 1-4). Les LINEs (« Long Interspersed Nuclear Elements ») sont des éléments qui codent pour la transcriptase inverse. Les SINEs (« Short Interspersed Nuclear Elements ») sont des éléments non autonomes qui utilisent la machinerie enzymatique des LINEs.

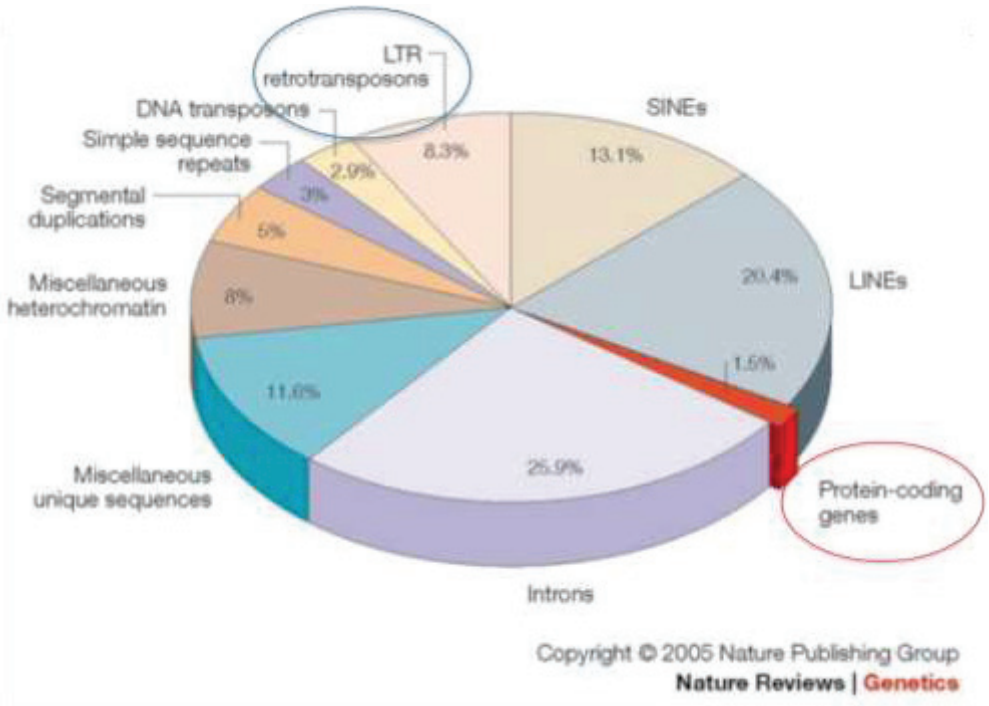


Figure 1-4: Composition du génome. D'après (Gregory, 2005).

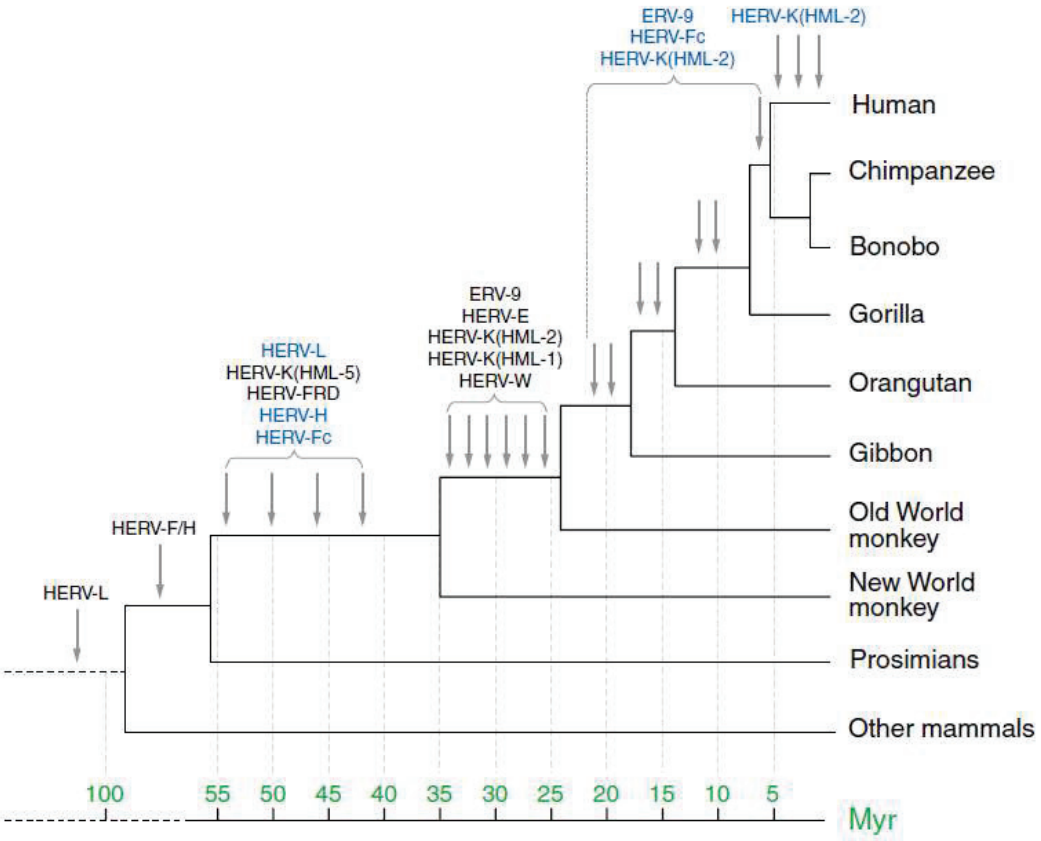


Figure 1-5: Age d'insertion de certains rétrovirus dans le génome humain. Date d'intégration de rétrovirus sur l'arbre phylogénétique de quelques primates. D'après (Bannert and Kurth, 2006).

Les rétrotransposons à LTR sont composés des HERV et des MaLR. Ces derniers ont une structure similaire aux rétrovirus, mais sans transcriptase inverse, sans site d'initiation de la transcription et sans région codante pour les protéines d'enveloppe (Bannert and Kurth, 2004), et donc sont non infectieux. Des chercheurs ont émis l'hypothèse que certains MaLR aient acquis un gènes codant pour une protéine d'enveloppe fusogène et ainsi donnés naissance aux rétrovirus infectieux (Cotton, 2001; Malik et al., 2000). Par simplicité, le terme HERV sera utilisé pour désigner l'ensemble des rétrotransposons à LTR (HERV + MaLR). Les HERV représentent donc plus de 8% de notre génome. A titre de comparaison, la partie codante pour les protéines représente moins de 2% (Figure 1-4). Depuis l'intégration il y a des millions d'années et leur duplication en de multiples copies, de nombreuses mutations / modifications se sont accumulées dans leur structure (Figure 1-6). Cela a pu être des mutations ponctuelles dans la séquences, des délétions de régions entières, des rétrotransposition de sous-régions provenant d'un autre HERV, ou encore des recombinaisons homologues. Ces dernières sont responsables de la formation de LTR solo, qui représentent aujourd'hui la très large majorité des HERV composant notre génome.

Ainsi, tous les HERV dupliqués à partir d'un même élément inséré dans le génome forment un même groupe. Due à ces nombreuses modifications et réarrangements dans leur séquence, il est cependant difficile de classifier les HERV. En 2016, Vargiu et al. (Vargiu et al., 2016) ont détecté 39 différents groupes homogènes et également 31 groupes hybrides (ou non canoniques). On peut toutefois estimer les dates d'insertions de certaines familles en comparant leur présence entre les génomes de primates (Figure 1-5). Plusieurs propositions de classification ce sont succédées, certains groupes ont été initialement défini à partir de leur similarité de séquence, d'autre à partir de leur séquence PBS (« Primer Binding Site », séquence située en aval de la LTR 5', reconnue par un ARN de transfert spécifique). La classification actuelle est ainsi un mélange de ces groupes définis de manières différentes, et pour lesquelles, les LTR et les régions internes peuvent avoir des noms différents, ce qui génère de la complexité pour les analyses et interprétations sur les groupes de HERV. Dans les bases de données publiques, comme Repbase (Bao et al., 2015), les HERV sont regroupés en plusieurs centaines de sous-groupes (normalement définis sur le groupe et la région du HERV correspondant), eux même regroupés en super-familles. Aucun de ces 2 niveaux ne correspond exactement aux groupes de HERV et rend ainsi difficile l'interprétation au niveau

famille. Une proposition très récente de classification permettrait d'unifier en partie ces nomenclatures existantes, toute en prenant en compte l'orthologie entre les différents groupes de HERV et les rétrovirus (Gifford et al., 2018).

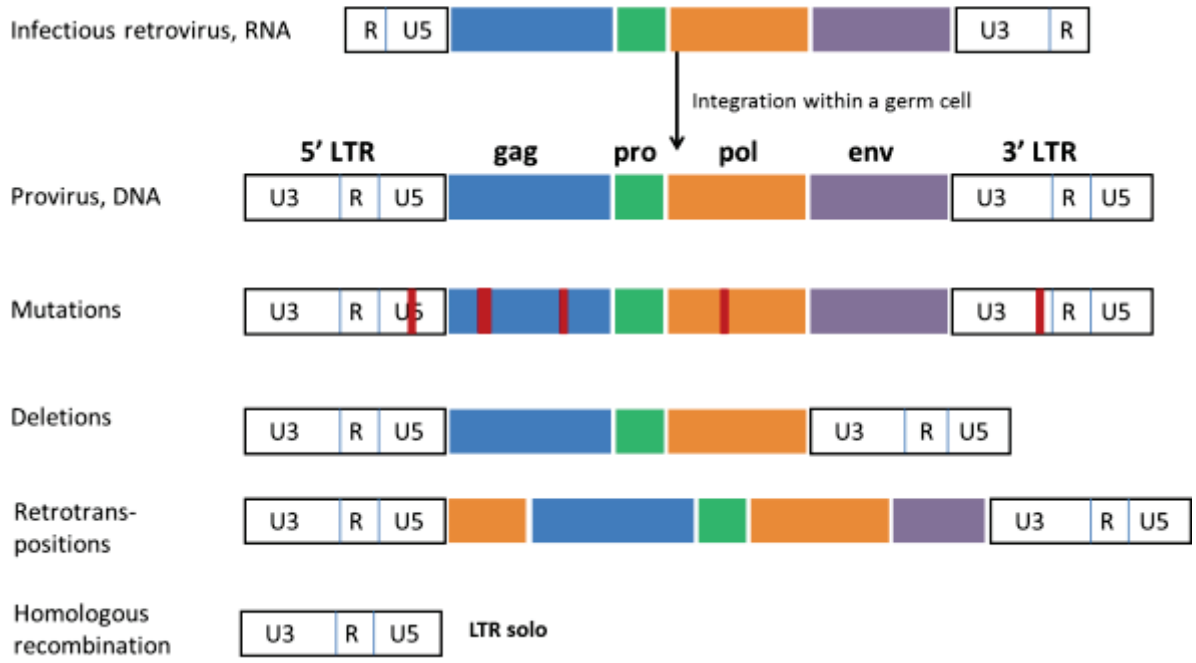


Figure 1-6: Structure et évolution des HERV au sein du génome. Initialement, après intégration du rétrovirus dans une cellule germinale, le HERV partage une structure identique au rétrovirus exogène. C'est-à-dire des régions internes codant pour les protéines rétrovirales (gag, pro, pol, env), flanquées de 2 régions LTR identiques (Long Terminal Repeat), régulatrices de l'expression rétrovirale. Au cours de l'évolution, l'accumulation de mutations et d'évènements de délétion / recombinaison ont donné des formes très diverses pour des éléments provenant du même évènement d'insertion, formant un groupe (ou famille) complexe. La majorité des HERV dans le génome sont sous la forme de LTR solo.

1.2.2 ROLES DES RETROTRANSPOSONS SUR LE GENOME

La conservation des rétrotransposons dans le génome des primates suggère qu'ils apportent un avantage sélectif à la survie de l'espèce (Cordaux and Batzer, 2009; Feschotte and Gilbert, 2012; Grow et al., 2015). Pour les HERV, ce serait le cas notamment via des effets sur la placentation (Simpson et al., 1996), le développement du cerveau (Mortelmans et al., 2016) ou du système immunitaire (Hurst and Magiorkinis, 2015).

L'ensemble des éléments répétés peuvent agir de multiples manières sur le génome et son expression. La réplication des rétrotransposons modifie fortement la structure du

génome et l'expression des gènes (Volff, 2006). Ils semblent qu'ils aient eu (ou ont) un impact important sur la plasticité et l'évolution des génomes, notamment par l'apport d'éléments régulateurs en cis « prêt à l'emploi » (Ellison and Bachtrog, 2013). L'insertion, au voisinage d'un gène, d'un rétrotransposon possédant un site de fixation de facteur de transcription peut présenter un avantage sélectif pour l'hôte. C'est le cas par exemple d'un HERV (du groupe MER41) qui entraine l'expression plus importante d'un gène (AIM2) impliqué dans la réponse inflammatoire innée à l'IFN (Chuong et al., 2016). Dans le génome humain, les éléments Alu (LINE) sont très actifs et se dupliquent à un taux élevé (Cordaux et al., 2006). Bien que la plupart de ces insertions soient neutres, certaines sont associées à des maladies génétiques (Hancks and Kazazian, 2012). Du fait de leur duplication, les rétrotransposons sont répétés dans le génome et peuvent favoriser des recombinaisons (Hughes and Coffin, 2005; Teixeira-Silva et al., 2013). Que ce soit pour la duplication ou la recombinaison, les rétrotransposons passent par un intermédiaire ARN, il y a donc transcription. L'expression de ces éléments transposables est généralement bien contrôlée par des modification épigénétiques. Cela peut être par la méthylation de l'ADN, avec par exemple le cas bien décrit de la dé-méthylation de loci HERV dans la région promotrice de la protéine Syncitine-1 exprimée dans les cellules du trophoblaste (Bolze et al., 2017; Gimenez et al., 2009). Cela peut aussi être via des marques histones, qu'elles soient répressives ou activatrices (Karimi et al., 2011; Macfarlan et al., 2011; Rowe et al., 2013).

1.2.3 HERV EN CONTEXTE PATHOLOGIQUE

Dans certains contextes pathologiques des modifications épigénétiques importantes peuvent avoir lieu et ainsi entraîner l'expression de HERV.

C'est initialement dans les maladies auto-immunes que l'expression de HERV a été décrite. Dans la sclérose en plaques (SEP), des particules rétrovirales possédant une activité de transcriptase inverse ont été détectées à partir de le liquide Céphalo-rachidien (LCR) de patients (Perron et al., 1989, 1991). Ces ARN ont été identifiés comme étant un rétrovirus endogène appartenant à la famille HERV-W et dénommé MSRV (multiple sclerosis associated retrovirus) (Blond et al., 1999; Perron et al., 1997). Une augmentation de l'expression de loci de la famille HERV-W a été retrouvée dans les cellules du cerveau de patients avec SEP (Antony et al., 2004; Mameli et al., 2007). D'autres familles de HERV sont retrouvées

exprimées dans le sang des patients SEP (Christensen et al., 1997; Laska et al., 2012). Des HERV sont retrouvés exprimés dans de nombreuses autres maladies auto-immune. Par exemple, dans le diabète de type I caractérisé par la destruction de cellules pancréatiques par les lymphocytes T, des ARN d'un HERV du groupe HERV-W ont été retrouvés dans les PBMC et conduisent à la formation d'une protéine d'enveloppe détectable dans le sérum de ces patients (Levet et al., 2017). Cette protéine est par ailleurs suspectée d'avoir un rôle dans la pathogénèse de cette maladie en inhibant la sécrétion d'insuline. On retrouve l'expression de protéines de HERV chez des patients atteints de Lupus érythémateux, notamment la protéine p30 de la région Gag d'un HERV-E (Hishikawa et al., 1997), ou encore des peptides rétroviraux provoquant la formation d'antigènes anti-rétroviraux (Bengtsson et al., 1996). Lors de l'infection par le virus de l'immunodéficience humaine (VIH), l'expression de HERV dans le sang a également été décrite (Contreras-Galindo et al., 2006; Vincendeau et al., 2015).

L'expression de HERV a également été décrite dans plusieurs cancers, et dans certains cas, le HERV peut jouer un rôle sur le développement de la maladie. C'est le cas dans le lymphome de Hodgkin. Dans les lymphocytes B, une LTR de HERV est hypométhylée et joue un rôle de promoteur alternatif entrainant l'expression du gène CSF1R physiologiquement exprimé dans les macrophages (Lamprecht et al., 2010). Les différents rôles possibles que peuvent avoir les LTR sur l'expression des gènes conventionnels seront présentés dans la section 1.2.4. Dans beaucoup d'autres cancers, une expression de certains HERV non exprimés en conditions normales est observé, sans savoir si le HERV joue un rôle dans le développement du cancer. C'est le cas par exemple dans le mélanome (Humer et al., 2006), dans le cancer du pancréas (Schmitz-Winnenthal et al., 2007), cancer du sein (Wang-Johanning et al., 2001), leucémie myéloïde chronique (Brodsky et al., 1993).

Au-delà de leur expression, en tant qu'éléments répétés, les HERV peuvent être responsables de recombinaisons homologues de l'ADN (Feschotte and Gilbert, 2012). Dans le cancer de la prostate, des réarrangements chromosomiques entraînent la fusion entre un provirus HERV-K et un oncogène appelé ETS (Tomlins et al., 2007).

On retrouve aussi l'expression de HERV dans des contextes infectieux, tels que l'infection par l'Epstein-Barr virus (EBV), où une protéine est produite à partir d'un HERV dans les lymphocytes (Gross et al., 2011) ou encore dans la grippe A (Li et al., 2014).

Dans les contextes inflammatoires, encore peu d'études s'y sont intéressées chez l'Homme. Chez la souris, on retrouve la réactivation de MLV (Murin Leukemia Viruses) après stimulation par LPS (Isaak and Cerny, 1983; Kwon et al., 2011). *In vitro*, on retrouve également une expression plus importante de HERV dans les monocytes humains après stimulation au LPS ou PMA (Johnston et al., 2001). Plusieurs exemples montrent une modulation globale des HERV après infection par un pathogène (Assinger et al., 2013; Young et al., 2014). In vivo, dans le sang de patients brulés graves, une expression d'un dizaine de groupes de HERV a également été observée. Cependant, leur étude était basée sur des RT-PCR ciblant l'ensemble des éléments d'un groupe, et donc les auteurs ne pouvaient déterminer si l'ensemble des loci du groupe sont exprimés, ou si l'expression totale est représentée par un seul locus au sein d'un même groupe (Lee et al., 2014, 2013).

Dans tous ces exemples de pathologies liées à la réponse immunitaire, des HERV sont retrouvés exprimés, jouant un rôle ou non dans le développement de la maladie. Cela montre des liens étroits entre le dérèglement de la réponse immunitaire de l'hôte et l'expression des HERV. Bien que le rôle des HERV dans l'initiation de la réponse immunitaire ne soit pas toujours clairement défini, on peut imaginer que l'expression d'une protéine, ou même d'un ARN ayant une structure ressemblant à celle d'un rétrovirus puisse être reconnue par le système immunitaire comme un PAMP, comme cela peut être le cas pour les rétrovirus exogènes. Cela pourrait résulter en l'activation des PRR et donc de la réponse immunitaire (Hurst and Magiorkinis, 2015). Chez la souris, il a été démontré le rôle de certains récepteurs TLR (TLR3, 7 et 9) dans le contrôle de l'expression des HERV (Yu et al., 2012).

Comme vu dans le lymphome de Hodgkin, un rôle possible des HERV est d'influer en cis sur l'expression d'un gène, par ses LTR.

1.2.4 ROLES FONCTIONNELS DES LTR SUR L'EXPRESSION DES GENES

Comme évoqué dans les différents exemples précédents, les LTR des HERV, notamment en situation pathologique impliquant une réponse immunitaire, peuvent jouer un rôle en cis sur l'expression des gènes conventionnels. Les LTR sont le centre de contrôle de l'expression des gènes rétroviraux. Elles possèdent des sites de fixation de facteur de transcription, des sites promoteurs, des signaux de polyadénylation, voire des sites d'épissages. Des exemples montrent que des LTR, qui sont majoritairement sous forme de LTR solo dans le génome, peuvent donc influer sur l'expression de gènes conventionnels situés proche dans le génome (Figure 1-7).

Elle peuvent donc fournir des sites de fixation de facteurs de transcription pour augmenter (ou inhiber) l'expression d'un gène situé proche dans le génome. Un des premiers cas décrit de HERV pouvant avoir un rôle de fixation de facteur de transcription est celui de l'amylase (Meisler and Ting, 1993). Chez l'Homme, l'amylase est produite dans le pancréas et les glandes salivaires. En réalité, il y a 2 loci de l'amylase qui sont exprimés dans le pancréas, et 3 autres loci dans les glandes salivaires. Ces 3 derniers ont tous un HERV-E inséré en amont du site d'initiation de la transcription. C'est l'insertion de ce HERV qui permet l'expression de l'amylase dans les glandes salivaires. Un autre exemple intéressant et plus récent est celui du groupe MER41. Il possède des sites de fixation pour STAT1 et entraîne l'expression du gène AIM2, impliqué dans la création de l'inflammasome, en réponse à l'interféron (Chuong et al., 2016). Les auteurs émettent l'hypothèse que ces éléments HERV, anciennement des rétrovirus, réservoirs d'enhancers de la voie de l'interféron, auraient exploité les réseaux de l'immunité pour promouvoir leur transcription et leur réplication et que, d'un autre côté, l'hôte aurait pu les utiliser pour faciliter le « turn-over » immunitaire et ainsi s'adapter plus facilement.

Les LTR peuvent également avoir un rôle de promoteur du gène situé dans un voisinage génomique proche. L'exemple du cas évoqué précédemment, où la LTR joue un rôle promoteur du gène CSF1R dans le lymphome de Hodgkin (Lamprecht et al., 2010) est un des mieux décrit. Dans ce cas, la LTR n'est pas le promoteur naturel de ce gène. Dans d'autres cas, le rôle promoteur de la LTR peut engendrer un transcrit alternatif ou incomplet. C'est le cas par exemple dans les lymphocytes B, où certains patients expriment l'antigène CD5 et d'autres non. Il a été montré que l'absence de CD5 était directement liée à l'activité d'un

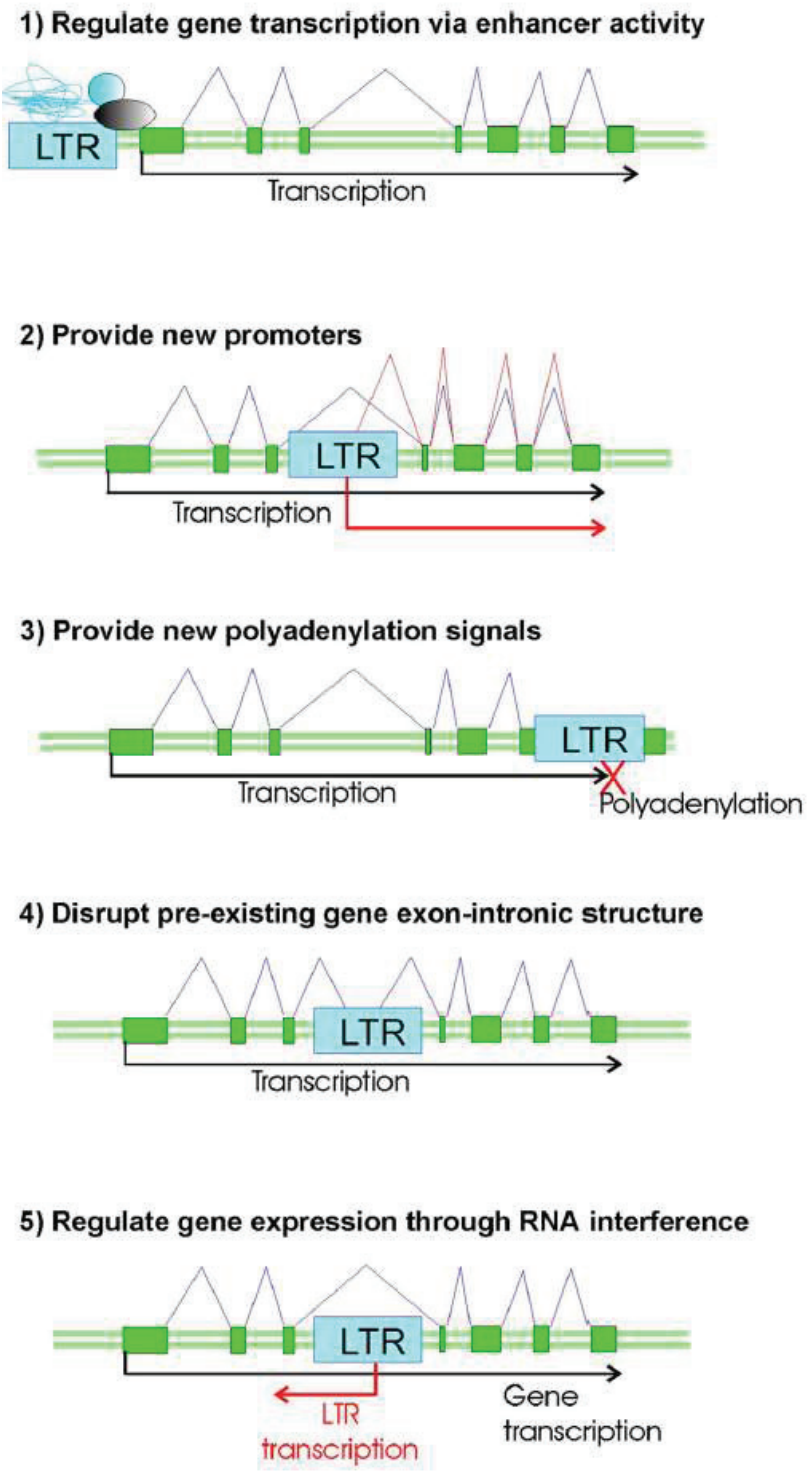


Figure 1-7: Rôles des LTR sur l'expression des gènes. 1) Les HERV peuvent agir comme enhancer ou silencer de la transcription d'un gène. **2)** Ils peuvent avoir un rôle de promoteur alternatif sur un gène proche , voire créer des transcrits nouveaux. **3)** Ils peuvent fournir des site de polyadénylation pouvant soit raccourcir ou allonger les transcrits des gènes de l'hôte. **4)** Ils peuvent modifier la structure exon-intron des gènes en fournissant de nouveaux sites d'épissage. **5**) Ils peuvent initier des transcrits anti-sens chevauchant les ARN du gène de l'hôte. D'après (Suntsova et al., 2015).

promoteur alternatif HERV, entraînant l'expression d'un transcrit CD5 différent et non codant pour la protéine (Renaudineau et al., 2005). De nombreux autres exemples sont documentés dans la littérature, notamment dans des contextes cancéreux (Nagai et al., 2012; Pérot et al., 2015; Scarfò et al., 2016; Wiesner et al., 2015). Cependant, certaines études montrent que dans la plupart du temps, la LTR joue un rôle promoteur mineur sur le gène, et souvent en ne changeant pas la nature des transcrits (Cohen et al., 2009).

Les LTR peuvent également apporter des signaux de terminaison de transcrits. La plupart des ARNm se terminent par une queue polyA, et le signal requis pour cette terminaison est la séquence AATAAA. Parmi les gènes humains possédant des UTR (Untranslated Region), 27% contiendrait au moins un transcrit contenant un élément transposable dans leur UTR (van de Lagemaat et al., 2003). Cette même équipe a mis en évidence deux gènes humains dont le signal de polyadénylation est une LTR (Mager et al., 1999).

Enfin, les LTR peuvent agir sur le gène proche en générant des transcrit anti-sens des transcrit du gène. Des LTR de HERV-K, localisées dans un intron des gènes SLC4A8 et IFT172, expriment des transcrits qui sont complémentaires de certains exons des transcrits des gènes. Un expression importante de ces transcrits anti-sens étaient associés à une diminution significative de l'ARNm de ces transcrits (Gogvadze et al., 2009).

1.2.5 POLYMORPHISME HERV ET REPONSE DE L'HOTE

Des liens ont été mis en évidence entre le génome et la réponse de l'hôte à l'infection. Dans le sepsis, les gènes des TLR1 et TLR4, qui sont polymorphiques, sont associés à un risque augmenté de développer un choc septique d'origine bactérienne (Agnese et al., 2002; Lorenz et al., 2002). Dans une autre étude, un SNP situé dans le gène SVEP1 est associé à une probabilité plus grande de mortalité à 28 jours après le diagnostic du sepsis (Nakada et al., 2015). Le système HLA (Human Leukocyte Antigen), localisé sur le chromosome 6 et comportant tous les gènes codants pour les CMH, est connu pour son polymorphisme très important. Ce polymorphisme pose notamment des problèmes de rejets lors des greffes, car les protéines produites par l'organe du donneur peuvent être reconnues comme des antigènes par le système immunitaire du receveur (Morishima et al., 2015; Petersdorf et al., 1995; Williams, 2001). Ce polymorphisme peut aussi avoir des conséquences sur la réponse

de l'hôte à l'infection, comme par exemple celui observé dans la 3'UTR de HLA-G, qui peut avoir des conséquences bénéfiques ou néfastes sur la réponse immune de l'hôte, selon le contexte pathologique (Donadi et al., 2011). Le polymorphisme du système HLA est aussi considéré comme un des facteurs les plus important concernant le devenir des patients avec hépatite B ou C (Tamori and Kawada, 2013). Ce sont ici des exemples avec des SNP, où il est clair que l'état du génome à un endroit donné peut influencer la capacité de l'hôte à combattre une infection. Les HERV dans le génome peuvent aussi avoir une influence sur l'état de l'organisme hôte (Bannert and Kurth, 2004). Nous pouvons alors nous demander si l'absence ou la présence d'un locus HERV peut avoir une influence sur la réponse de l'hôte.

A l'heure actuelle, aucun HERV n'a été retrouvé infectieux chez l'Homme. Ils est également admis que les HERV ont perdu la capacité de se dupliquer. Cependant, une série de recombinaisons entre éléments d'un même groupe peut restaurer la capacité de réplication du HERV (Young et al., 2013). Chez des souris immunodéficientes, qui ne semblent pas avoir de MLV (Meurine Leukemia Virus) capables de se répliquer, un rétrovirus endogène acquiert spontanément la capacité à être infectieux (Young et al., 2012). Chez l'Homme, ce phénomène est notamment suspecté pour la famille HERV-K, et a pu être reproduit *in vitro* (Dewannieux et al., 2006; Lee and Bieniasz, 2007). De plus, il semble que les HML-2 peuvent recombiner naturellement (Hughes and Coffin, 2005).

Par ailleurs, des études ont montré des insertions de HERV du groupe HML-2 qui ne sont pas fixées dans les populations (Marchi et al., 2014; Wildschutte et al., 2016). Comme évoqué précédemment, les HERV peuvent générer des variations de loci en nombre de copie via des recombinaisons homologues (Campbell et al., 2014). Récemment, une équipe a récupéré toutes les insertions de HERV-K détectées dans les différentes études sur les 1000 génomes, et ont sélectionné parmi ces insertions, les HERV-K qui étaient associés à un SNP « adjacent ». Ensuite, ils ont fouillé dans les base de données d'eQTL, pour trouver des associations entre les SNP adjacent aux insertions des HERV-K et l'expression de gènes, associés à des maladies. Les auteurs ont détecté des associations significatives entre ces insertions de HERV-K ayant un SNP associé avec des maladies auto-immunes ou maladie de Parkinson (Wallace et al., 2018).

Pour résumer, les HERV-K, de par leur polymorphisme insertionnel, semblent avoir une influence fonctionnelle sur le génome, et plus particulièrement dans la région hautement polymorphique du système HLA. Il pourrait être intéressant de regarder d'une part s'il existe un polymorphisme sur l'ensemble des HERV annotés du génome et non pas seulement sur le groupe des HERV-K, bien qu'il semble qu'il soit admis que seul ce groupe est polymorphique. Et d'autre part, il serait intéressant d'étudier l'impact (HERV-K ou autre) de leur polymorphisme en nombre de copies sur l'expression du génome.

1.3 METHODOLOGIES D'EXPLORATION DU HERVOME

Avant le séquençage du génome humain, l'exploration du HERVome (l'ensemble des HERV présents dans le génome) était difficile à réaliser, que ce soit au niveau génomique ou transcriptomique. Plusieurs approches permettant l'exploration du transcriptome HERV par RT-PCR notamment, ont été successivement développées à partir des années 2000. Soit par ciblage de loci spécifiques, bien définis telles que les séquences d'enveloppes HERV possédant un cadre de lecture (Parseval et al., 2003). Soit par des approches au niveau des groupes de HERV par le développement d'amorces dégénérées exploitant la conservation évolutives des séquences *pol* des rétrovirus afin d'identifier des tendances d'expression au sein d'un même groupe (Forsman et al., 2005; Lee et al., 2014, 2013; Muradrasoli et al., 2006). Des approches alliant RT-PCR et puces à ADN rétroviral se développent en parallèle permettant une détection à plus large spectre de transcrits rétroviraux (Seifarth et al., 2000).

Les approches par ESTs (Expressed Sequence Tags) ont également permis l'étude de loci HERV individuels à l'origine d'expressions tissu-spécifiques (Oja et al., 2007; Stauffer et al., 2004). Ce sont des petites séquences d'ADN, généralement entre 200 et 500 nucléotides, générées par le séquençage de clones d'ADNc. Chaque fragment représente un évènement de transcription dans un contexte cellulaire donné, et peut être aligné au locus à l'origine de la transcription. Cependant, l'approche par EST comporte plusieurs limites telles que la difficulté d'obtenir des annotations fiables, le taux important d'erreur de séquençage (3%) des banques d'ADNc, problématique pour l'étude des éléments répétés du génome et le manque de concordance avec les approches par couplage entre RT-PCR et séquençage (Flockerzi et al., 2008). Cette dernière approche consiste à réaliser une RT-PCR spécifique sur tout un groupe de HERV puis le produit d'ADNc est cloné et séquencé et a permis d'étudier pour la première fois l'expression de l'ensemble des loci de la famille HML-2 (Flockerzi et al., 2008).

La volonté d'analyser le transcriptome HERV à grande échelle, tout en restant au niveau locus spécifique, ainsi que d'analyser leur fonctionnalité, a poussé le laboratoire à développer des puces à ADN dédiées à l'exploration de l'expression des HERV. Plusieurs générations de puces se sont ainsi succédées. La première (HERV-V1) cible les loci de 4 groupes de HERV (HERV-W, HML-2, HERV-H et HERV-E), avec des sondes ciblant

majoritairement les LTR (Gimenez et al., 2010). Cette première génération ne cible toutefois pas l'ensemble du transcriptome et a révélé des problèmes de spécificité dus à la nature répétées des éléments ciblés. La seconde génération (HERV-V2) a réglé les problèmes de spécificité et intègre les loci de 6 groupes (HERV-W, HERV-H, HERV-E, HERV-FRD, HML-2, HML-5, (Pérot et al., 2012)). Le répertoire exploré est encore incomplet et ne prend pas en compte les propriétés thermodynamiques des sondes s'hybridant avec leur cible. La troisième génération (puce HERV-V3) cible quant à elle la quasi-totalité du HERVome, les régions fonctionnelles de chaque locus et prend en compte les problèmes de réactions croisées. Une description détaillée de cet outil et de la base de données d'annotation associée se trouve dans la section 3.1.2.

Ainsi dans ce projet de thèse, nous employons principalement des puces à ADN, qu'elles soient commerciales (Affymétrix HG U133plus2) ou fabriquées au laboratoire (puce Affymétrix HERV-V3) pour explorer le transcriptome HERV. Certains loci identifiés à travers les différentes analyses ont ensuite pu être validés par RT-PCR quantitative. A l'heure du NGS (Next Generation Sequencing), la technologie des puces tend à disparaître. A l'inverse, les technologies NGS sont de plus en plus utilisées, de plus en plus performantes et de moins en moins couteuses. L'avantage majeur des technologies de séquençage à haut débit, est qu'elles ne nécessitent pas de connaissances des cibles à priori. Contrairement aux puces, qui nécessitent le design de sondes spécifiques de chaque élément ciblé. Il faut ainsi une connaissance *à priori* des séquences d'intérêt et par conséquent, elle ne permettent pas de découvrir de nouveaux transcrits.

Cependant, pour cibler les éléments répétés du génome avec les technologies de séquençage, il reste encore des améliorations à apporter (Ewing, 2015). Ces éléments ayant une très forte similarité entre eux, pour des reads de taille trop petite, il sera impossible de déterminer sur quel locus le read mappe, et un des risque est d'avoir une couverture de read plus importante sur les éléments répétés. Les algorithmes traitant le mapping des éléments répétés deviennent de plus en plus performants (Lengauer, 2008), la profondeur de séquençage et la taille des reads est de plus en plus élevée et peut maintenant aller jusqu'à 300pb pour les technologies NGS de technologie Illumina, voire 400pb pour les IonTorrent (Escalona et al., 2016), permettant ainsi de limiter cet effet. Des technologies de séquençage

de 3ème génération dites technologies « long reads » sont sorties plus récemment, et permettent d'avoir des tailles de reads allant jusqu'à plusieurs dizaine de kilo bases et donc de s'affranchir des problèmes de mapping sur les éléments répétés. Cependant, bien que ces technologies représentent très certainement l'avenir, elles sont encore très perfectibles. Elles comportent notamment un taux de fausses lectures très élevé, d'environ 15% pour la technologie long reads Minion nanopore (Jain et al., 2018). De plus, 15% du génome de référence n'est pas couvert par cette technologie. Et 80% des reads non assemblés appartiennent à ces éléments répétés, dont 16% des rétrotransposons à LTR. A cela s'ajoute que 60% des reads assemblés mais non mappés sont représentés par les éléments répétés. Pour la technologie PacBio (Pacific Biosciences), le taux d'erreurs est similaire, entre 11 et 15% (Korlach, 2015). Cela est bien évidemment encore trop important pour étudier des éléments répétés qui peuvent différer d'une seul base sur plusieurs centaines entre 2 loci d'un même groupe. De plus les couts sont encore très élevés. Une stratégie de plus en plus utilisée est de combiner des NGS et des technologies long reads (« hybrid sequencing »), cela améliore grandement les performances mais également les coûts (Rhoads and Au, 2015).

En autre argument en faveur du choix de la puce par rapport aux NGS pour explorer le transcriptome, est la quantité d'ARN nécessaire pour réaliser l'expérience. Pour la puce HERV-V3, le protocole utilisé nécessite seulement 12ng d'ARN au départ, ce qui est bien moindre que pour le RNAseq qui, pour Illumina par exemple, nécessite au moins 100ng d'ARN (kit « Illumina truseq stranded total rna »). Or, dans des contextes de recherche translationnelle, où nous analysons la plupart du temps directement des échantillons de sang de patients, les quantités d'ARN qu'on peut obtenir sont souvent limitées.

En prenant en compte qu'à l'époque du design de la puce HERV-V3 les technologies NGS étaient encore balbutiantes, qu'une expertise de réalisation de puces s'est développée au laboratoire depuis des décennies, que nous sommes dans un cadre de recherche translationnelle, et que nous nous intéressons spécifiquement aux éléments répétés, nous avons majoritairement utilisé la technologie des puces. Un travail très important sur le niveau de spécificité a été réalisé sur le design des millions de sondes que composent la puce HERV-V3 (section 3.1.2.1, article section 3.1.2.5). Ces sondes assurent une spécificité pour

chaque élément HERV ciblés et constituent une ressource importante, probablement également pour le développement de l'analyse des HERV par NGS.

Ces difficultés à mapper les éléments répétés du génome montrent aussi que les informations d'annotation du génome présentes dans les bases de données, spécialement pour les éléments répétés, sont probablement incomplètes voire erronées. Le génome de référence, bien que constamment amélioré/complété, provient initialement de seulement quelques individus. Le génome référence n'est probablement qu'un génome « consensus » et n'est pas représentatif du génome d'un individu. On se demande ainsi dans ce projet si les HERV annotés dans notre génome sont-ils toujours bien présents chez tous les individus, et si non, dans quelle mesure.

Bien que probablement imparfaites, les bases de données d'annotation des HERV du le génome sont indispensables pour pouvoir générer et interpréter les données génomiques ou transcriptomiques. Plusieurs bases existent, avec chacune leurs spécificités. La base de données RepBase regroupe les séquences répétées représentatives de plusieurs organismes eucaryotes. Elle n'est pas exhaustive, i.e. elle ne regroupe pas l'ensemble des loci des différents génomes, mais présente des séquences consensus, vérifiées, faisant une base de données fiable et correctement annotée (Bao et al., 2015). L'annotation des HERV la plus couramment utilisée aujourd'hui provient de cette base, où les séquences consensus sont numérotées et les régions LTR et internes d'un même locus portent des noms différents (ex : pour un locus du groupe HML-2, les régions internes se nomment HERVK-int et les LTR se nomment LTR5). La base de données Dfam est une base de données publique des familles des éléments répétés dans laquelle chaque famille est représentée par plusieurs alignements de séquences et des profils de chaine de Markov cachées (HMM, (Hubley et al., 2016)). La détection d'éléments répétés s'est réalisée *in silico* par l'utilisation de modèles HMM et elle contient un nombre d'entrées plus important que dans RepBase, avec 4150 modèles de familles. Enfin, les sites permettant l'exploration du génome (http://www.ensembl.org, https://genome.ucsc.edu/, https://www.ncbi.nlm.nih.gov/) annotant les éléments du génomes à différents niveaux, utilisent pour cela le logiciel RepeatMasker. Il permet la détection des éléments répétés ou de faible complexité dans le génome (http://www.repeatmasker.org/). Ainsi les « tracks » d'annotation des HERV (LTR)

dans les explorateurs du génome sont obtenues à partir de RepeatMasker, qui lui-même peut prendre en entrées les modèles HMM provenant de Dfam, et dont l'annotation est celle provenant de RepBase. Dans ce travail, nous exploitons également notre propre base de données (*hervgdb4*), utilisée pour le développement de la puce, et qui contient une annotation fine de l'ensemble des éléments HERV du génome, au niveau de loci entiers, et aussi au niveau de régions fonctionnelles (section 3.1.1.1).

2 PROJET DE THESE

L'objectif de ce projet de thèse est d'étudier la contribution des HERV au sein de la réponse immunitaire de l'hôte en conditions d'agression inflammatoire. Pour cela, nous proposons d'une part, de décrire le transcriptome HERV et sa modulation sur différentes cohortes de patients post-agression inflammatoires (choc septique, traumas, brulés), ou sur un modèle mimant la réponse immunitaire de l'hôte dans ces conditions. D'autre part, nous allons étudier le polymorphisme en nombre de copies des HERV sur plus de 2600 génomes, ainsi que les éventuelles associations sur leur propre expression et sur l'ensemble du transcriptome dans le sang. Enfin, nous allons tenter d'identifier des LTR jouant un rôle sur l'expression des composantes de la réponse immunitaire. Compte-tenu des connaissances actuelles sur le rôle des HERV dans des contextes liés à l'immunité, différentes de l'agression inflammatoire, il nous paraît pertinent d'étudier le génome et le transcriptome HERV dans ces contextes.

2.1 DEVELOPPEMENTS METHODOLOGIQUES

Une première partie consiste à présenter les développements méthodologiques effectués en amont et durant ce projet de thèse. Une partie importante pour la bonne réalisation du projet de thèse a été de participer à la validation de l'outil majoritairement utilisé pour explorer le transcriptome HERV, la puce HERV-V3, ainsi que la base de données d'annotation des HERV associée (hervgdb4). J'ai validé le pipeline d'analyse de la puce HERV-V3, puce non commerciale développée au sein du laboratoire. Afin de faciliter l'exploration de la base de donnée et l'interprétation des résultats de la puce par les membres de l'équipe, j'ai pu développer un outil de visualisation basé sur l'outil Shiny (RStudio, Inc, 2014).

2.2 TRANSCRIPTOME

Une seconde partie consiste à décrire le transcriptome HERV dans le sang, dans des contextes d'agression inflammatoire. Une telle étude n'ayant jamais été réalisée dans ces contextes et à si grande échelle. Pour cela, nous avons utilisé une approche intéressante développée par Reichmann et al. qui est de ré-annoter les sondes des puces à ADN commerciales, ayant fait l'observation que certaines sondes mappent sur un HERV proche du

gène ciblé. Nous montrons que certains HERV sont exprimés, qu'il existe de la modulation HERV commune entre 4 cohortes différentes, des chocs septiques (2 cohortes), des brulés sévères et des traumatisés graves, comparés à des volontaires sains (HV). La confirmation d'une modulation commune de HERV dans ces conditions et leur localisation proche de gènes impliqués dans la réponse immunitaire nous permet de suggérer un lien entre réponse spécifique du transcriptome HERV et réponse immunitaire après une agression inflammatoire.

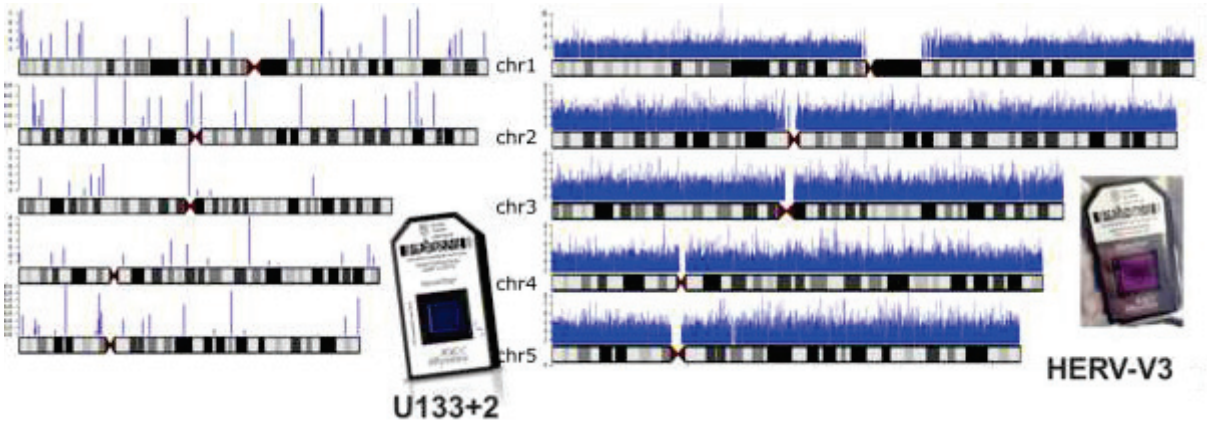


Figure 2-1: Comparaison des puces U133plus2 et HERV-V3. Représentation des probesets ciblant des HERV dans chacune des puces sur les 5 premiers chromosomes. Au total, la puce U133plus2 comporte 337 probesets ciblant des HERV, la puce HERV-V3 en comporte 1 279 090.

Cependant, cette analyse ne porte que sur 337 HERV, ciblant une infime partie du génome HERV. Le développement de la puce HERV-V3 dans le laboratoire, spécifiquement dédiée à l'analyse des HERV, nous permet aujourd'hui de cibler de manière quasi exhaustive leur transcriptome. La Figure 2-1 représente la fraction des HERV ciblés par chaque puce. A partir de la puce HERV-V3, et en collaboration avec Marine Mommert (une étudiante en thèse de biologie du laboratoire), nous avons décrit le transcriptome HERV dans le sang dans un modèle mimant la réponse immunitaire de l'hôte observée dans le sepsis, avec une condition inflammatoire et une condition d'immunodépression. Nous avons également décrit ce transcriptome dans le sang total de patients en choc septique. Enfin, nous avons débuté un travail de description de l'expression des HERV dans le sang par RNAseq, en collaboration avec Maria-Paola Pisano, une étudiante en thèse de biologie à l'université de Cagliari.

2.3 VARIATIONS EN NOMBRE DE COPIES DES HERV DANS LE GENOME

La troisième partie consiste en l'étude des variations en nombre de copies des HERV dans le génome. Pour cela, avec la collaboration de Maxime Bodinier un étudiant en bioinformatique du laboratoire, nous avons développé une méthode (HERVdel) permettant de décrire le polymorphisme de présence (absence ou présence) dans plus de 2000 génomes humains sains (cohorte 1000 génomes) de tous les HERV annotés dans le génome (annotation publique provenant de RepeatMasker et base de données *hervgdb4*). Pour cela, nous comparons dans chaque échantillon, le niveau de couverture des données de séquençage de chaque locus HERV avec son environnement génomique. Si la couverture au niveau d'un locus est significativement plus faible que son environnement, alors ce locus est considéré absent dans le génome étudié.

Cette étude a pour but initial de connaître l'ampleur de ce polymorphisme, connu pour le groupe HML-2, sur l'ensemble des HERV, ainsi que de le comparer entre les différentes populations humaines, réparties sur le globe. Elle a également pour objectif de comprendre la variabilité plus importante observée dans le transcriptome HERV et ainsi être capable de gagner en puissance d'interprétation. L'idée étant de vérifier, si les HERV exprimés dans différents contextes mais avec une grande variabilité d'expression inter-individuelle, sont globalement polymorphiques dans les populations étudiées.

J'exploite par la suite les résultats du génome HERV pour réaliser des analyses d'association avec le transcriptome, décrites dans la section 2.4.

2.4 ASSOCIATIONS DES HERV AVEC LE TRANSCRIPTOME

Afin d'avoir des pistes quant à un éventuel rôle des HERV dans la réponse immunitaire de l'hôte, dans cette dernière partie nous tentons d'associer les HERV (par leur polymorphisme de présence ou par leur expression) avec le transcriptome (HERV et gènes).

Afin d'identifier des HERV dont la présence (ou l'absence) peut potentiellement agir sur l'expression de gènes, j'ai exploité les résultats de l'étude du polymorphisme de présence des HERV (méthode *HERVdel*, section 0). Dans un premier temps je tente de corréler ce

polymorphisme à l'échelle des populations humaines (fréquence allélique ou VAF par HERV et par population), avec l'expression des HERVs dans le sang, dans les contextes d'agression inflammatoire. Dans un second temps, j'utilise les données de RNAseq, publiquement disponibles, provenant des mêmes volontaires sains des 1000 génomes, afin d'associer, individu par individu, l'expression des HERV avec leur génotype (ie. La présence d'un HERV est-elle corrélé avec une expression de ce même HERV ?). J'ai enfin tenté d'associer le polymorphisme de présence des HERV avec l'expression de gènes situés à proximité, dans le but d'identifier des loci pouvant potentiellement avoir un impact sur l'expression de gènes, avec un intérêt porté tout particulièrement dans les gènes impliqués dans la réponse immunitaire.

D'autre part, nous cherchons à identifier des LTR pouvant jouer un rôle en cis, promoteur ou de terminaison de transcrits de gènes de l'immunité dans le sang, à la suite d'une agression inflammatoire. Pour cela, j'exploite la base de données *hervgdb4* et les probesets de la puce HERV-V3, qui contiennent une annotation à différents degrés de précision, des LTR au niveau de leur sous-région U3, R et U5. Nous sommes parti du postulat qu'une LTR jouant un rôle promoteur sur un gène voisin verrait sa région U5 exprimée, compte-tenu de la position du promoteur chez les rétrovirus, normalement situé dans la région U3. Une LTR jouant un rôle de terminaison de transcrit, via la présence de signaux de polyadénylation, verrait sa région U3 exprimée, compte-tenu de la position dans la région U5 des signaux de polyadénylation chez les rétrovirus. Ainsi, nous avons sélectionné les LTR ciblées en totalité par la puce HERV-V3 et assigné des fonctions dans les différents échantillons analysés. La manière dont on exploite le signal de la puce HERV-V3 pour assigner une fonction est décrite sur la Figure 2-2.

Une fois des fonctions potentielles attribuées, nous mettons en place des modèles testant l'association (eQTL-like) ou la co-expression entre LTR et gène voisin. Le modèle d'association se base sur le même principe que les analyses d'eQTL (expression Quantitative Trait Loci), qui permettent d'associer un SNP dans le génome avec l'expression d'un transcrit à un autre endroit dans le génome (Figure 2-3 A). Cette méthode permet de mettre en évidence le possible rôle du polymorphisme d'un locus sur un gène. Dans notre cas, au lieu de considérer un SNP à un endroit du génome, nous considérons la fonction d'un HERV et sa

possible association avec l'expression d'un gène (Figure 2-3 B). En réalité, il est très probable de ne pas avoir plusieurs états de LTR au sein d'une même condition (Figure 2-3 C). Ce type de modèle peut également être utilisé pour associer la présence ou l'absence d'un HERV avec l'expression d'un gène. Afin d'avoir une approche différente et complémentaire de la méthode eQTL-like, nous cherchons également à identifier les LTR possédant une et une seule sous-région fonctionnelle par LTR très corrélée à l'expression d'un gène (modèle de co-expression). L'intérêt de n'avoir qu'une seule sous-région est de pouvoir lui inférer une fonction probable, soit Pr, si la région U5 est corrélée au gène, soit pA, si la région U3 est corrélée au gène.

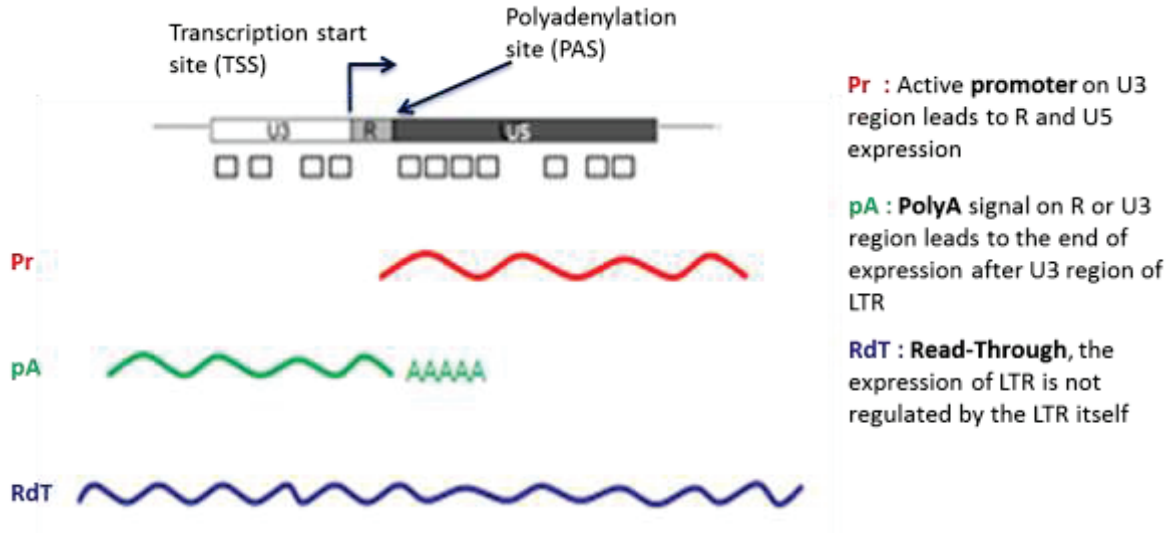


Figure 2-2: Assignation de fonction aux LTR à partir de la puce HERV-V3. Les sondes de la puce HERV-V3 (carrés blanc) sont regroupés en probeset ciblant chacun une sous-région fonctionnelle de la LTR (U3, R et U5). Une LTR sera assignée Promotrice (Pr, rouge) dans un échantillon si sa région U5 est exprimée mais pas sa région U3. Elle sera assignée comme signal de polyAdénylation (pA, vert) dans un échantillon si sa région U3 est exprimée mais pas sa région U5. Une LTR pourra être assignée « Read Through » (RdD, bleu), si ses 2 région U3 et U5 portent un signal, sans différence significative entre les 2 sous régions. Dans ce cas l'expression « passe » par la LTR, elle-même ne semblant pas jouer de rôle ni dans l'initiation, ni dans la terminaison du transcrit. Enfin si elle ne possède pas de signal, la LTR sera considérée comme silencieuse (No ou Silent).

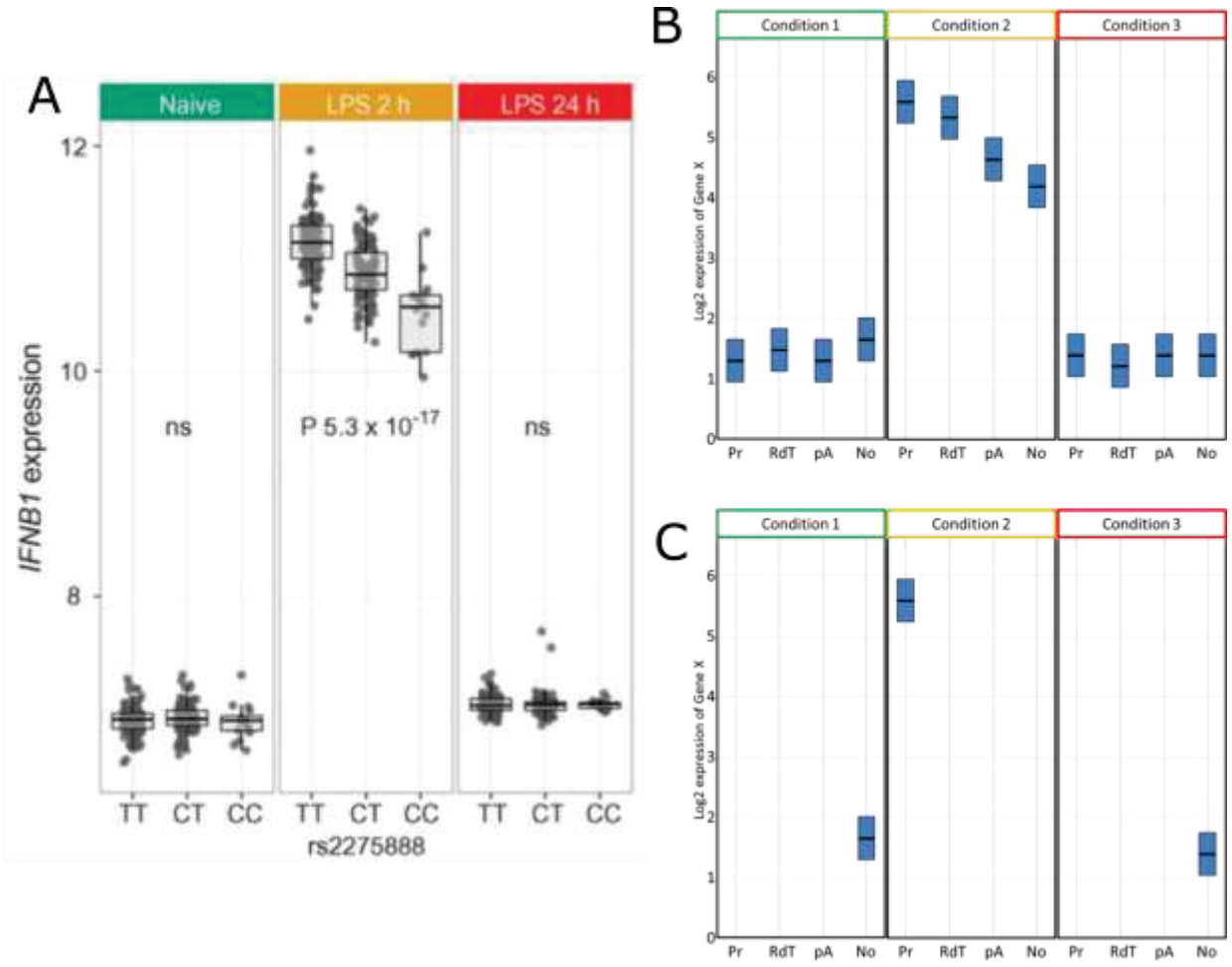


Figure 2-3: eQTL-like. A. Exemple d'un eQTL dans les monocytes. Deux heures après stimulation au LPS, l'expression du gène IFNB1 est augmentée par rapport aux autres conditions. Le SNP rs2275888 influence son expression, où la présence d'homozygotie pour TT entraîne une expression plus importante du gène que les individus CT et CC. D'après (Fairfax et al., 2014). **B. Transposition théorique du modèle eQTL en associant la fonction des LTR avec l'expression d'un gène.** Dans cette illustration, la fonction Promotrice de la LTR étudiée entrainerait une expression du gène plus importante que les autres état de la LTR. **C. Modèle eQTL like réaliste.** Il prend en compte la probable modulation de fonction de la LTR entre les conditions, ainsi que le fait qu'un LTR ne puisse être à la fois Pr et pA (déterminisme opérationnel). Pr : Promoteur, RdT : « Read Through », pA : polyAdénylation, No : LTR non exprimée.

Avec l'ensemble de ces méthodes, nous espérons identifier des HERV candidats, avec des LTR ayant soit un rôle promoteur ou de terminaison de l'expression de gène de l'immunité, soit des HERV dont leur absence ou présence dans le génome pourraient être associée à l'augmentation ou la diminution de l'expression d'un gène.

3 RESULTATS

3.1 DEVELOPPEMENTS METHODOLOGIQUES

3.1.1 BASE DE DONNEES *HERVGDB4*

3.1.1.1 DESCRIPTION

Une base de données des rétrotransposons à LTR, comprenant plus de 890 000 entrées annotant plus de 400 000 éléments a été générée dans le laboratoire (Table 3-1). Elle couvre la quasi-totalité des éléments HERV (HERV et MaLR) du génome. Elle a servi à développer la puce HERV-V3 (section 3.1.2). J'ai en tout premier lieu réalisé un travail d'annotation et de nettoyage de la base existante, appelée *Hervgdb4*. Cette base contient deux types d'annotation: L'annotation *prototype* et l'annotation *Dfam* (Figure 3-1).

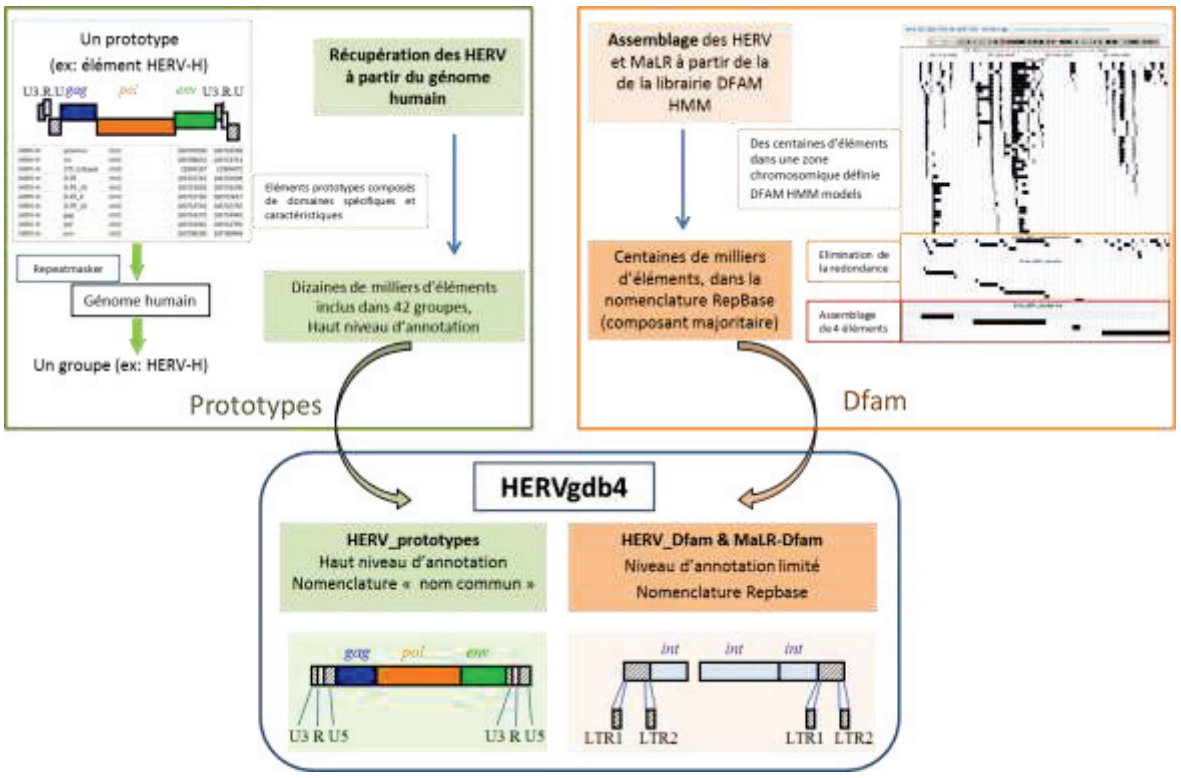


Figure 3-1: Création de la base de données HERVgdb4. Deux types de répertoire sont présents : Le répertoire *Prototypes*, qui contient relativement peu d'éléments HERV mais annotés avec précision et les répertoires *Dfam* qui contiennent beaucoup d'éléments (HERV et MaLR) mais moins précisément annotés. Les éléments prototypes ont été récupérés par homologie de séquences à partir d'un HERV Prototype pour chacun des 42 groupes de HERV. Les éléments *Dfam* ont été récupérés à partir de la librairie Dfam, qui les a identifié par une approche bioinformatique via des modèles de Markov Cachés.

	Base de données *Hervgdb4*			**Puce HERV-V3**			
Répertoire	***Nb de loci***	***Nb d'entrées***	***Nb de groupes***	***Nb de loci***	***Nb de probesets***	***Nb de sondes***	***Nb de groupes***
HERV_Dfam	168 278	340 134	450	150 540	566 722	2 029 484	450
MaLR_Dfam	227 174	453 582	45	153 746	621 980	2 075 943	45
HERV_proto	29 271	87 877	42	24 312	90 388	301 121	42
LINE1	998	4 610	1	661	2 814	11 583	1
lncRNA	3 795	3802	1	3757	7 528	43 448	1
Gènes U133	/	/	/	1 559	7 704	84 904	/
Gènes HTA	/	/	/	1 559	70 632	686 938	/
Gènes Opti	/	/	/	1 559	3 032	17 472	/
Total	**429 516**	**890005**	**539**	**337 693**	**1 370 800**	**5 250 893**	**539**

Table 3-1: Composition de la base de données *Hervgdb4* et de la puce HERV-V3, par répertoire.

L'annotation *prototype* regroupe des loci homologues à des séquences de « prototypes » de HERV connus. Cela a conduit à la formation du répertoire HERV prototypes, qui comporte 29 859 loci HERV répartis au sein de 42 familles ou groupes. Ces éléments sont annotés en détail, au niveau des sous-régions fonctionnelles. Le second type consiste à assembler des éléments HERV et MaLR à partir des éléments fragmentés contenus dans la base de données DFAM (Hubley et al., 2016) et identifiés par des modèles de Markov caché ou « Hidden Markov Models » (HMM). Cela a conduit à la formation des répertoires HERV Dfam et MALR Dfam, qui comportent 169 821 et 228 429 éléments respectivement. Ces éléments Dfam sont nommés avec la nomenclature RepBase (Bao et al., 2015), avec un niveau d'annotation limité. La base de données contient également quelques éléments LINE et longs ARN non-codants.

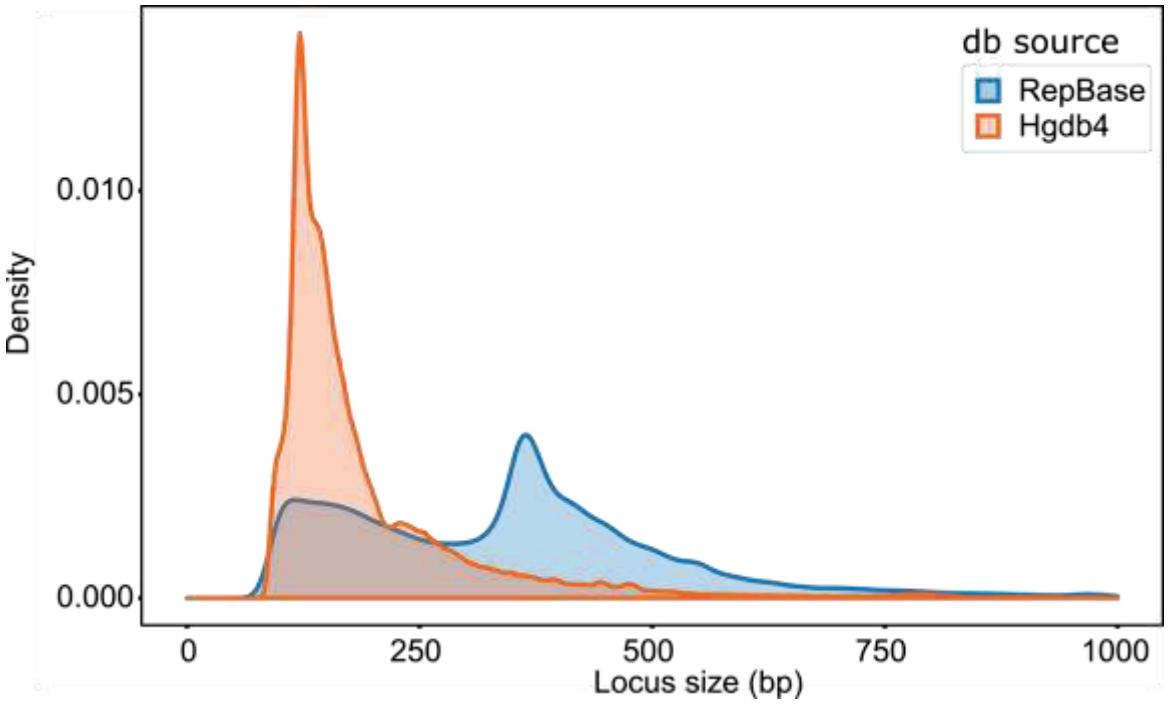


Figure 3-2 : Distribution des tailles de loci selon la base de donnée. RepBase : Base de données d'annotation publique. Hgdb4 : base de données d'annotation du laboratoire.

Un de mes premiers travaux a consisté à transformer les coordonnées des entrées de *hervgdb4* de la version Grch37 vers la dernière version du génome, Grch38. J'ai également extrait des informations du site Dfam, afin de récupérer le groupe d'appartenance des HERV et MaLR du répertoire *Dfam*. *Hervgdb4* est annoté au niveau des sous-régions, c'est-à-dire que chaque entrée de la base correspond à une sous-région fonctionnelle du HERV ou MaLR. J'ai ajouté les attributs dans la bases permettant de retrouver l'identité de la région complète du HERV (LTR solo, LTR 5', région interne ou LTR 3'), ainsi que la totalité du locus HERV (identifiant de 7 chiffres). Comme expliqué précédemment, l'annotation pour les répertoires *Dfam* est limitée. Les LTR ont été découpées en 2, LTR1 et LTR2, dans l'ordre de leur position génomique. Si le HERV est sur le brin +, la LTR1 correspond donc à la sous-région U3 (~U3), la LTR2 aux sous régions R et U5 (~RU5), et vice versa pour le brin -. Le découpage des LTR en sous-régions pour ces répertoires est approximatif. Ces annotations sont indispensable pour analyser et interpréter les études génomiques ou transcriptomiques sur les HERV.

La base de données publique, provenant de *RepBase,* regroupe l'annotation des HERV et MaLR et a été obtenue en sélectionnant la track LTR dans *RepeatMasker* (Smit et al., 2013). Elle contient 720 177 entrées. Le niveau d'annotation est différent de celui de *Hervgdb4*.

Ceci est illustré par des distributions de taille de locus différentes entre les deux bases de données (Figure 3-2). *Hervgdb4*, fragmenté au niveau des sous-région fonctionnelles, contient des entrées de taille plus petites, et qui ont été utilisées pour le développement des probesets de la puce HERV-V3.

3.1.1.2 OUTIL DE VISUALISATION

Afin d'explorer facilement la base de données, j'ai développé une application web en R via le package Shiny (v1.0.5). Cette application a été mise à disposition dans l'intranet bioMérieux, et toute personne du laboratoire a la possibilité de la consulter. L'utilisateur peut, de manière interactive, représenter et résumer graphiquement les données de *Hervgdb4*, (ex: représentation des comptes par chromosome de tous les HERV *prototypes*). Il est également possible d'afficher toute la base de données sous forme de tableau interactif. Cela permet de filtrer un ensemble d'entrées rapidement (ex : tous les HERV-H situés sur le chromosome 6) ou de ne sélectionner qu'un seul locus en particulier. Pour chaque locus, un lien vers le site web Ensembl (Aken et al., 2017) permet d'afficher le contexte génomique du locus. L'utilisateur peut également afficher les résultats d'expression sur des jeux de données de la puce HERV-V3 (Figure 3-3). Cette fonctionnalité permet à tout utilisateur de visualiser facilement les profils d'expression des loci de son choix, ciblés par la puce. De plus, il peut sélectionner des groupes d'échantillons, basés sur plusieurs variables expérimentales (conditions de stimulation, réplicats techniques, batchs, …) ou cliniques (jour de prélèvement, sexe des patients, infections secondaires, …)

3.1.2 PUCE HERV-V3

Avant mon arrivée, une puce à ADN ciblant spécifiquement les rétrotransposons à LTR a été développée à partir de la base *hervgdb4*. J'ai participé à la validation de cet outil et à la définition du pipeline d'analyses. C'est une puce Affymetrix de haute densité à technologie « oligo courts » (25pb), basée sur l'hybridation complémentaire et spécifique d'une sonde avec un ARN cible. Cette puce cible plus de 420 000 loci HERV et MaLR. Trois grandes étapes ont été nécessaires au développement de la puce. Premièrement, la récupération des HERV du génome et leur annotation, comme expliqué dans la section 3.1.1.1, afin de créer la base

de données *Hervgdb4*. Deuxièmement, une étape de mise en place du modèle d'hybridation. Troisièmement, le design et la définition des sondes.

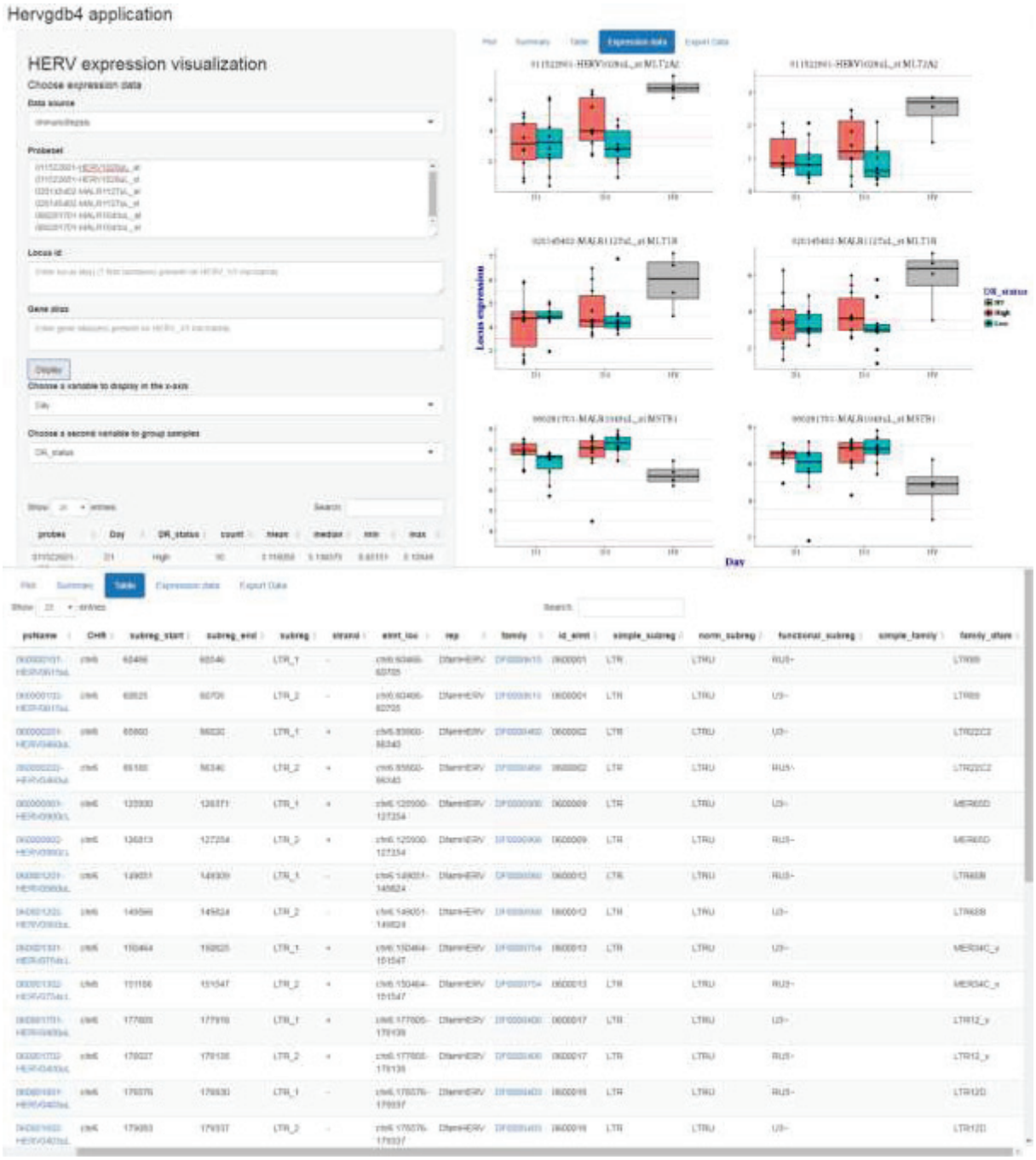


Figure 3-3 : Captures d'écran de l'application permettant l'exploration de la base de données HERVgdb4. L'application développée en Shiny, permet de visualiser sous forme de graphiques descriptifs et interactifs la base de données HERVgdb4 - d'afficher sous forme d'un d'un tableau interactif l'ensemble des entrées de la base et de trier et filtrer sur plusieurs paramètres – et de représenter graphiquement les données d'expression de puces HERV-V3 réalisées sur 3 jeux de données.

3.1.2.1 DEVELOPPEMENT

Comme expliqué dans l'introduction (section 1.3), la répétition multiple d'éléments de séquence identique ou proche dans le génome complique le design de sondes spécifiques, et donc l'analyse de l'expression de ces éléments.

Un modèle d'hybridation appelé PEHM pour « Pentamer rEgion-dependent Hybridization Model » a ainsi été développé. Brièvement, ce modèle a pour but de déterminer les possibilités d'hybridation croisées pour chaque sonde potentielle ciblant chaque HERV et donc d'éliminer les sondes qui ne sont pas assez spécifiques ou dont les hybrides sonde / cible ne sont pas assez stables. Pour déterminer le potentiel d'hybridation croisée, le modèle calcule des affinités d'appariement des brins d'ADN de la sonde et de sa cible. Ces affinités sont calculées comme étant la somme des effets de la sonde découpée en k-mers. Les effets tiennent compte des propriétés structurelles (flexibilité et stabilité, qui dépendent des interactions avec les paires de bases voisines), de l'impact de la position du k-mer dans la sonde (qui a été découpée en 3 parties de tailles égales), et enfin de l'impact des mismatchs, et gaps. Enfin, la taille des k-mers permettant la meilleure prédiction du modèle est de 5. Cette taille a été déterminée à partir de 20 probesets au format commercial U133, connus et validés, et pour lesquelles les intensités des sondes prédites par le modèle PEHM ressemblent le plus aux intensités observées (Figure 1B, Figure S3 de l'article section 3.1.2.5). Une fois le modèle d'hybridation des sondes défini, un seuil d'affinité a été sélectionné pour distinguer les hybrides stables et instables. Il a été défini tel que 90% des sondes avec une affinité sous ce seuil ont une intensité inférieure au bruit de fond.

La dernière étape a consisté au design des sondes. Chaque entrée de *Hervgdb4* a été décomposée en potentielles sondes de 25 pb, décalé à chaque fois de 1 à 4 pb. Le modèle PEHM est utilisé pour chacune des sondes potentielles, et l'affinité pour sa cible est calculée. Si l'affinité excède le seuil, la sonde est alignée au génome de référence. S'il en ressort un seul « hit », alors la sonde est classifiée « spécifique ». S'il en ressort plusieurs, les affinités pour chaque hit sont calculées et si moins de 4 hits ont une affinité supérieure au seuil, alors la sonde est classifiée comme « potentiellement cross-hybridante ». Si la sonde a plus de 4 hits avec une affinité supérieure au seuil, la sonde est classifiée « non-spécifique » et est exclue. Ensuite, les sondes pour chaque entrée de *Hervgdb4* ont été regroupées en jeux de

sondes ou probesets. Un probeset doit regrouper entre 3 et 6 sondes, est restreint à une région de 400pb maximum et au moins une sonde classifiée « spécifique » doit être présente.

3.1.2.2 DESCRIPTION

La puce cible 328 598 loci HERV ou MaLR, à l'aide de 1 279 090 probesets et 4 406 548 sondes (Table 3-1). La puce contient 77% des loci de *Hervgdb4*. Elle cible des éléments provenant des répertoires HERV_prototypes, HERV_Dfam et MALR_Dfam de *Hervgdb4*, mais aussi des Lines et lncRNA. De plus, elle cible 1559 gènes liés à l'immunité. En détail, la puce HERV-V3 comporte 3 répertoires différents ciblant les même 1559 gènes: le répertoire U133, qui contient les sondes de la puce commerciale Affymetrix hgU133plus2 – le répertoire HTA, qui contient les sondes de la puce commerciale Affymetrix HTA 2.0 (Human Transcriptome Array) – le répertoire Opti, qui se base sur l'utilisation du modèle d'hybridation PEHM appliqué sur ces gènes.

3.1.2.3 VALIDATION DE LA PUCE

Pour valider ce nouvel outil, plusieurs critères ont été pris en compte, basés sur la répétabilité et la précision. Comme illustré dans la figure 2 de l'article de la puce HERV-V3 3.1.2.5, la répétabilité est meilleure pour les signaux d'intensité élevée. La figure 3 du même papier montre une très bonne corrélation entre la puce U133plus2 ou HTA et le répertoire U133 ou HTA de la puce HERV-V3.

Pour évaluer la reproductibilité de la puce, j'ai proposé une expérience à partir des 2 échantillons du MAQC (Microarray/Sequencing Quality Control consortium (MAQC Consortium et al., 2006)) : A, Stratagene Universal RNA, et B, Ambion Human Brain RNA. Nous avons répété 2 fois la même expérience, de manière indépendante. Des triplicats biologiques ont été réalisés sur l'échantillon A et sur l'échantillon B, qui ont été hybridés sur les puces HERV-V3 et U133. Pour chaque répétition, nous avons réalisé une analyse d'expression différentielle entre les échantillons A et B. Puis nous avons comparé les listes de loci différentiellement exprimés entre les deux répétitions (Figure 3-4 A). La concordance des deux listes de loci différentiellement exprimés a été évaluée par le POG (Percentage of Overlapping Genes). Le principe étant de séparer les éléments de chaque liste en 2 : ceux qui ont un log2 Fold Change (log2FC) négatif (ratio expression d'un loci A/B inférieur à 1), et

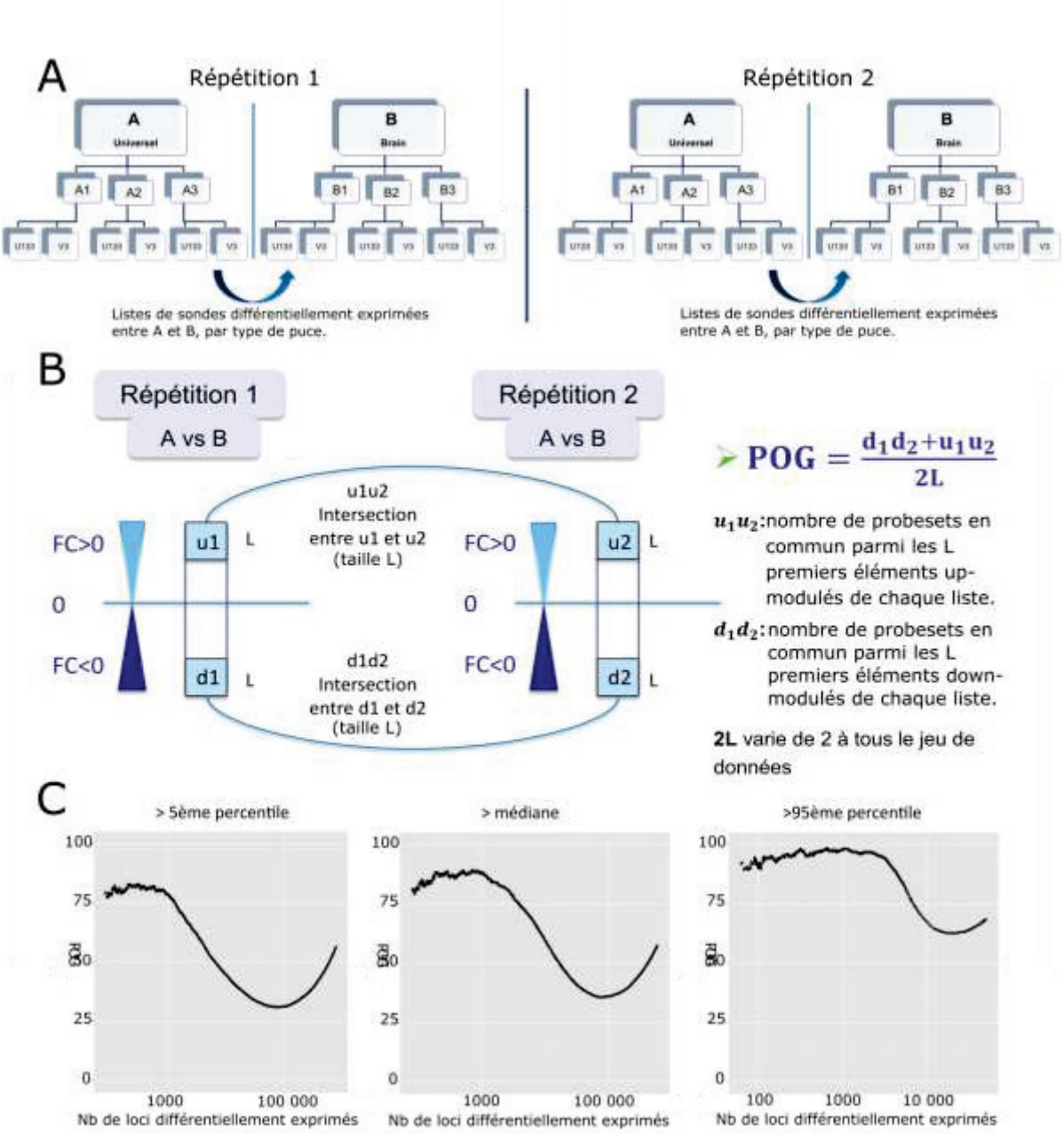


Figure 3-4: Etude de la reproductibilité de la puce. A. Design de l'expérience. **B.** Principe du calcul de concordance, ou POG (Percentage of Overlapping Genes). **C.** Résultats de concordance et influence du filtre sur l'intensité des sondes.

ceux qui ont un log2FC positif (ratio d'expression A/B supérieur à 1); De les classer par ordre de log2FC décroissant en valeur absolue ; Puis à tailles de liste égales, de comparer les listes entre A et B, deux à deux (Figure 3-4B). On fait varier les tailles de liste de 2 à la taille de la puce et on représente les POG en fonction de la taille de liste (Figure 3-4 C). Cette concordance relativement faible entre les listes d'éléments différentiellement exprimés des deux répétitions peut être améliorée en filtrant dès le départ le jeu de données, en ne

gardant que les probesets les plus exprimés. La figure montre de gauche à droite, avec un filtre de plus en plus strict, une amélioration de la concordance. Pour les 5% des probesets les plus exprimés, la concordance ne chute pas en dessous de 65%.

On peut conclure que la puce HERV_V3 est répétable et reproductible, à condition de ne sélectionner, en amont des analyses, que les HERV les plus exprimés dans les différents jeux de données.

3.1.2.4 PIPELINE D'ANALYSE

Plusieurs étapes sont ainsi nécessaires avant d'analyser des données. Ces étapes dites de « pre-processing » ont pour but de corriger les biais techniques, les biais liés à la technologie, de rendre comparables les puces entre elles et de ne sélectionner que les signaux interprétables.

3.1.2.4.1 PREPROCESSING

Correction du bruit de fond : Dans les puces à ADN, les valeurs de bruit de fond sont souvent importantes (Klebanov and Yakovlev, 2007; Marshall, 2004; Tu et al., 2002). Pour rappel, le bruit de fond est la valeur de fluorescence moyenne due à une hybridation non spécifique. Pour chaque intensité de signal mesurée, une part de l'intensité est due au bruit de fond et doit être corrigée pour ne garder que l'intensité due à l'effet biologique. Pour analyser la puce HERV-V3, nous avons choisi la méthode de correction RMA (Robust Multi-array Average, (Irizarry et al., 2003)). D'après cette méthode, l'intensité observée est composée de l'intensité réelle et d'un bruit aléatoire, qui suit une loi normale. On ne connait que l'intensité observée et on veut retrouver l'intensité réelle. Les paramètres de la loi normale du bruit aléatoire sont estimés séparément pour chaque puce à partir de la distribution des intensités observées. Les intensités réelles sont ainsi estimées.

Normalisation par les quantiles : Afin de pouvoir rendre comparable les données provenant de plusieurs échantillons, il est nécessaire que les distributions soient identiques. Le principe pour réaliser cela est assez simple: Considérons notre jeu de données avec, en ligne toutes les sondes de la puce et en colonne, leurs valeurs d'expression pour chaque échantillon (pour chaque puce). On classe dans l'ordre croissant les valeurs d'intensité, pour

chaque colonne séparément. Ainsi la première ligne de chaque colonne contiendra les valeurs des sondes ayant la plus faible intensité et la dernière celles qui ont la valeur la plus élevée. Ensuite, sur chaque ligne, on fait la moyenne de ces valeurs. On remplace chaque élément du tableau de données par la moyenne de sa ligne. Enfin on remet les valeurs dans l'ordre initial. Ainsi, les puces deviennent comparables entre elles.

Réduction du jeu de données (probes en probesets) : Le but de cette étape est de combiner les valeurs d'intensité des sondes qui ciblent le même élément en une seule valeur (jeu de sondes ou probeset). Pour cela on utilise un lissage par médiane. Pour cela, on considère des sous-tableaux contenant en colonne les puces et en lignes toutes les sondes qu'ils vont composer le même probeset. L'idée est de calculer, à chaque itération, la médiane de chaque colonne (la valeur médiane des sondes pour chaque probeset), on soustrait aux valeurs d'intensité la valeur de médiane de leur colonne correspondante. Ensuite, on calcule les médianes de chaque ligne et on soustrait aux valeurs d'intensité la valeur de médiane de leur ligne correspondante. On répète tant que les médianes des colonnes et des lignes sont toutes différentes de 0 (pour éviter des boucles infinies, le nombre de tour de boucle est en général limité à 5). La matrice résultante, de même taille que le tableau de données original est appelé la matrice résiduelle. On soustrait la matrice initiale par la matrice résiduelle pour obtenir une matrice des valeurs ajustées. La moyenne de chaque colonne (chaque puce) correspondra à la valeur d'intensité du probeset pour chaque puce.

3.1.2.4.2 CONTROLE QUALITE ET CORRECTION DES EFFETS BATCH

Le contrôle qualité des puces est basé sur plusieurs critères. Si une puce ne remplit pas assez de critères, elle sera retirée de l'analyse. Les critères évaluent : (1) la qualité des échantillons et de leur préparation, mesurée par le RIN, les contrôles d'amplification et de fragmentation de l'ARN, (2) la qualité des signaux bruts mesurée par les RLE plots (comparaison de l'expression de chaque probeset de chaque puce avec la médiane d'expression du probeset sur tout le jeu de données), les NUSE plots (Erreur standard normalisée qui permet d'identifier d'éventuels problèmes d'hybridation), ou des Analyses en Composante Principale (ACP) et (3) la qualité des signaux normalisés où les même critères que pour les signaux bruts sont évalués. L'évaluation de chacun des critères pour chaque

puce est résumé dans une table de décision, permettant de retirer une puce si elle n'a pas passé trois contrôles ou plus.

Il est ensuite nécessaire de corriger les effets techniques ou effets batch. Pour cela, nous employons, si nécessaire, la méthode COMBAT (COMbining BATches of gene expression microarray, (Johnson et al., 2007)).L' ANNEXE 1 (6.1) montre un exemple complet d'un contrôle qualité et de correction des effets batch réalisés sur un jeu de données (modèle ET) provenant de la puce HERV-V3.

3.1.2.4.3 FILTRE SUR L'INTENSITE DES SONDES

Face à de très grand jeux de données tels que les données de la puce HERV-V3 (matrice de 1 000 000 de lignes et un nombre de colonnes correspondant au nombre d'échantillons), et afin de réduire la complexité de calcul (en temps et en espace) et d'augmenter la puissance statistique, il est important de réduire leur taille. De plus, comme vu dans la section 3.1.2.3, la puce a une meilleure concordance quand on enlève les signaux les plus faibles, composant le bruit de fond.

On retire donc les signaux les plus faibles du jeu de données. Pour cela, nous devons choisir un seuil d'intensité en dessous duquel les signaux seront retirés. Pour nous affranchir de la variabilité observée sur les intensités entre différentes expériences, nous avons choisi de nous baser sur un seuil de coefficient de variation (CV). Nous représentons ainsi les CV en fonction des niveaux d'intensité. Nous choisissons le seuil de bruit de fond à partir duquel le troisième quartile de la distribution des probesets HERV est inférieur à un seuil de CV prédéterminé. Dans l'exemple de la Figure 3-5, le seuil de CV choisi est de 0,1, ce qui correspond à une intensité de 5,5 (en log base 2). Nous gardons ensuite les probesets qui ont une valeur supérieure à ce seuil dans x échantillons. Le nombre x choisi correspond en général au nombre minimal d'échantillons d'une même condition - 1. Par exemple, si notre jeu de données possède 45 échantillons, avec 3 différentes conditions comportant chacune 15 échantillons, on gardera un probeset si son signal est supérieur au seuil dans 14 échantillons ou plus. Une fois le filtre sur l'intensité réalisé, nous pouvons procéder à l'analyse des données. L'ensemble du pipeline d'analyse est résumé sur la Figure 3-6.

En conclusion, les développements méthodologiques réalisés dans cette section ont pour but de standardiser le processus d'analyse de la puce HERV-V3, afin d'obtenir des résultats interprétables, reproductibles, et de limiter le nombre de faux positifs. L'application web permettant l'exploration de *hervgdb4* et la visualisation des données d'expression de la puce HERV-V3 est quotidiennement utilisée par plusieurs personnes de l'équipe travaillant sur les HERV.

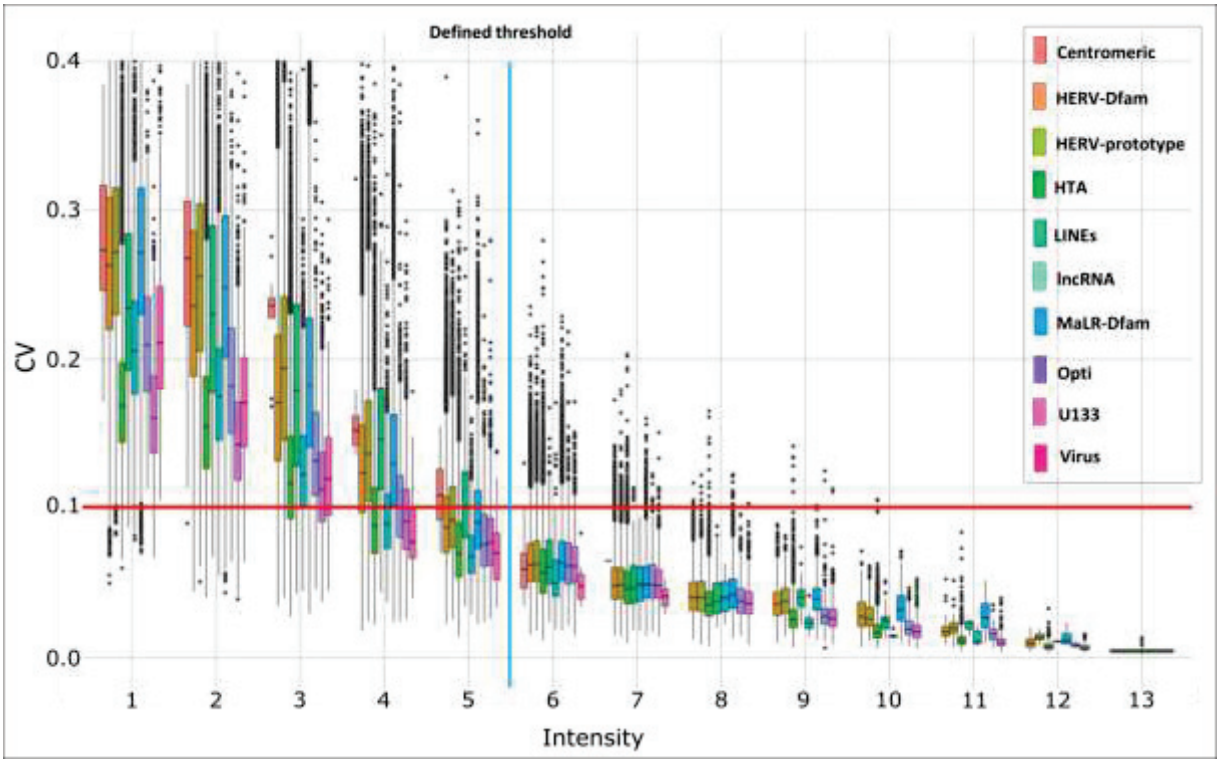


Figure 3-5 : Filtre sur l'intensité des probesets. Coefficients de variation (CV) en fonction du niveau d'intensité, par répertoire.

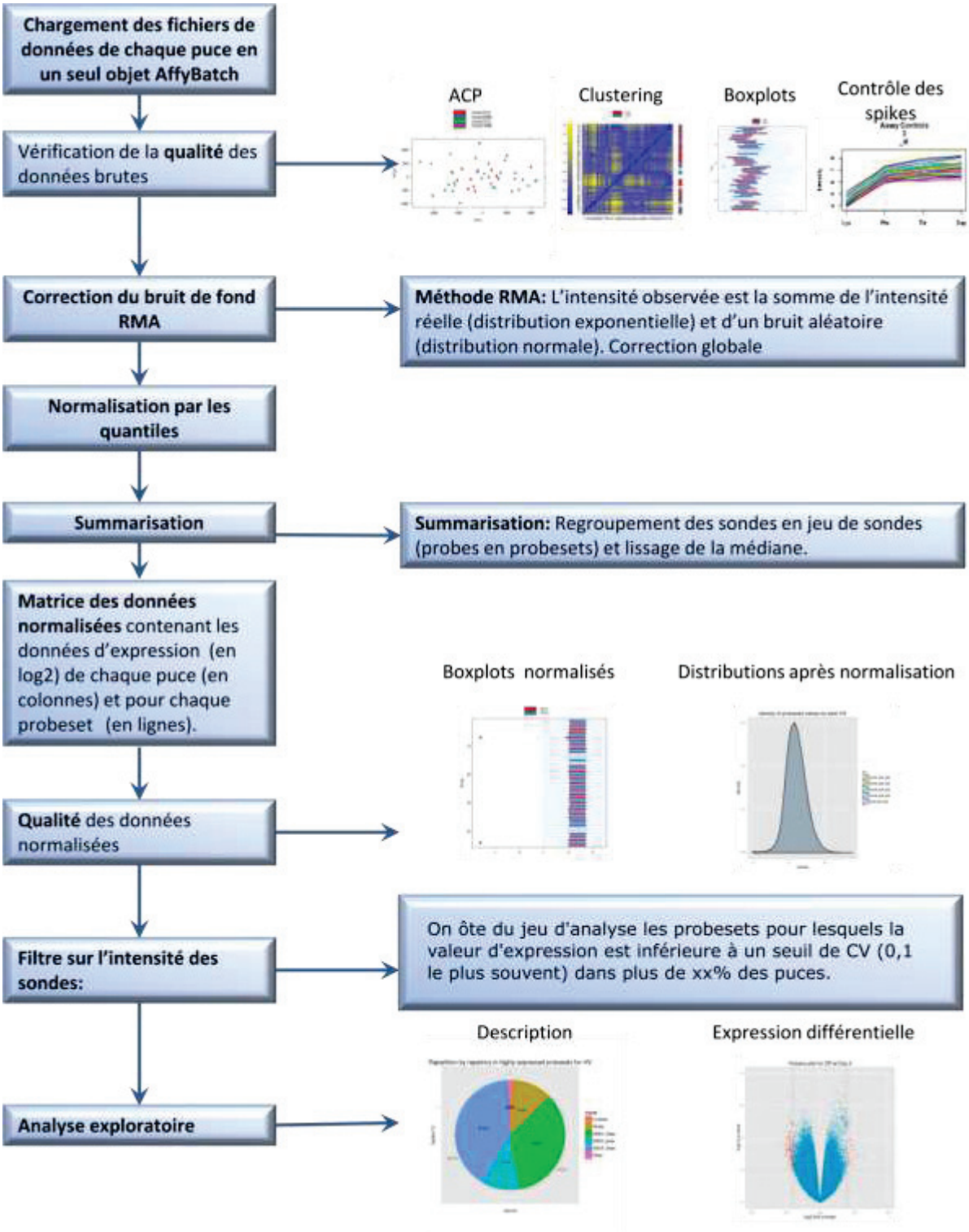


Figure 3-6 : Pipeline d'analyse de la puce HERV-V3.

3.1.2.5 ARTICLE

A COMPREHENSIVE HYBRIDIZATION MODEL ALLOWS WHOLE HERV TRANSCRIPTOME PROFILING USING HIGH DENSITY MICROARRAY

Jérémie Becker, Philippe Pérot, Valérie Cheynet, Guy Oriol, Nathalie Mugnier, Marine Mommert, **Olivier Tabone**, Julien Textoris, Jean-Baptiste Veyrieras and François Mallet

Année 2017

Publié dans BMC Genomics, (2017) 18:286

Becker et al. BMC Genomics (2017) 18:286
DOI 10.1186/s12864-017-3669-7

BMC Genomics

METHODOLOGY ARTICLE Open Access

A comprehensive hybridization model allows whole HERV transcriptome profiling using high density microarray

Jérémie Becker[1], Philippe Pérot[1], Valérie Cheynet[1], Guy Oriol[1], Nathalie Mugnier[2], Marine Mommert[1,3], Olivier Tabone[3], Julien Textoris[3], Jean-Baptiste Veyrieras[2] and François Mallet[1,3*]

Abstract

Background: Human endogenous retroviruses (HERVs) have received much attention for their implications in the etiology of many human diseases and their profound effect on evolution. Notably, recent studies have highlighted associations between HERVs expression and cancers (Yu et al., Int J Mol Med 32, 2013), autoimmunity (Balada et al., Int Rev Immunol 29:351–370, 2010) and neurological (Christensen, J Neuroimmune Pharmacol 5:326–335, 2010) conditions. Their repetitive nature makes their study particularly challenging, where expression studies have largely focused on individual loci (De Parseval et al., J Virol 77:10414–10422, 2003) or general trends within families (Forsman et al., J Virol Methods 129: 16–30, 2005; Seifarth et al., J Virol 79:341–352, 2005; Pichon et al., Nucleic Acids Res 34:e46, 2006).

Methods: To refine our understanding of HERVs activity, we introduce here a new microarray, HERV-V3. This work was made possible by the careful detection and annotation of genomic HERV/MaLR sequences as well as the development of a new hybridization model, allowing the optimization of probe performances and the control of cross-reactions.

Results: HERV-V3 offers an almost complete coverage of HERVs and their ancestors (mammalian apparent LTR-retrotransposons, MaLRs) at the locus level along with four other repertoires (active LINE-1 elements, lncRNA, a selection of 1559 human genes and common infectious viruses). We demonstrate that HERV-V3 analytical performances are comparable with commercial Affymetrix arrays, and that for a selection of tissue/pathological specific loci, the patterns of expression measured on HERV-V3 is consistent with those reported in the literature.

Conclusions: Given its large HERVs/MaLRs coverage and additional repertoires, HERV-V3 opens the door to multiple applications such as enhancers and alternative promoters identification, biomarkers identification as well as the characterization of genes and HERVs/MaLRs modulation caused by viral infection.

Keywords: Transcriptomics, Biostatistics, Microarray, Repetitive elements

* Correspondence: francois.mallet@biomerieux.com
[1]Joint research unit, Hospice Civils de Lyon, bioMerieux, Centre Hospitalier Lyon Sud, 165 Chemin du Grand Revoyet, 69310 Pierre-Benite, France
[3]EA 7426 Pathophysiology of Injury-induced Immunosuppression, University of Lyon1-Hospices Civils de Lyon-bioMérieux, Hôpital Edouard Herriot, 5 Place d'Arsonval, 69437 Lyon Cedex 3, France
Full list of author information is available at the end of the article

Background

The recent sequencing of model organisms unveiled the large proportion of repetitive elements (REs) in many species. In human, it is estimated that half of the genome is populated by REs and that retrovirus-like sequences amount for 8% of its coverage [1]. HERVs and MaLRs elements are organized into multi-copy families, for each of which, tens to thousands of distinct loci are scattered throughout the human genome, representing a pool of approximately 200,000 individual HERV loci. While bioinformatics approaches identified 103 HERV families and 1 MaLR family [1], only 40 HERV families were characterized in wet-lab studies [2–4]. Part of this genomic heritage is thought to originate from ancestral and independent retroviral infections within the germ line, before reinfection, retro-transposition and error-prone amplification steps during the evolution, leading to the formation of multi-copy families [5]. To date, no infectious endogenous virus has been detected in human, however 30% of the whole retrovirome is estimated to have a transcriptional activity [6]. Multiple functions have been assigned to these elements: HERVs have been demonstrated to act as canonical and alternative transcription start sites [7] (up to 30% of human and mouse TSSs are located in REs [8]), transcription termination sites [9] as well as splice donor and splice acceptor sites [10]. REs have further been suggested to be instrumental in the long intergenic non-coding RNA (lincRNA) regulatory system, where a majority of lincRNAs have been found to contain REs [11]. HERVs are increasingly associated with distinct physiological and pathological processes. One notable example is provided by the two syncytins genes that have been co-opted in human (and other mammals) to mediate placentation [12]. More recently, HERV-H loci have been shown to be instrumental in the maintenance of pluripotency [13]. Other investigations have further described associations between HERVs reactivation and multiple sclerosis [14–16], solid [17, 18] and hematological [19] tumors. Taken together, these studies show that REs provide binding sites for mammalian TFs and that they have rewired a number of developmental regulatory networks.

The central issue in the study of the HERV transcriptome arises from the phylogenetic proximity among the elements of a given HERV family, making the measure of each transcript technically challenging. Initially, RT-PCR techniques combined with degenerate primers [20] and low-density microarrays [18, 21] were developed to measure trends within families without, however, providing locus-specific information. Expressed sequence tags (ESTs) approaches gave a more comprehensive view of the HERV transcriptome but failed in many instances to identify the exact genomic source of expression [22]. Recent initiatives took advantage of probes targeting repetitive elements in commercial microarrays to monitor HERV behavior where, in addition to restricting their analysis to a small number of probes, the specificity of the probes was not evaluated [23]. More recently, HERVs transcription was also measured in various contexts using next generation sequencing (NGS) [24], which, while promising, remains difficult due to the ambiguity in assigning short reads mapping to more than one genomic location. For instance, in a study of HML-2 elements in teratocarcinoma cell line, Bhardwaj et al. showed that 47% of their reads had multiple alignments [25]. Two elegant initiatives sought to address this limitation by either using host surrounding sequences to anchor HERV copies [26] or by assigning multi-mapping reads probabilistically to specific locus based on the local genomic tag context [27]. However, in addition to assume that HERVs flanking regions are expressed, these approaches can probably not resolve multi-mapped reads for more than few hundred bases at the edges of HERV copies, leaving the ambiguity unchanged in the central regions.

Because HERV expression is globally low [28], very deep sequencing is required to capture the diversity of HERV transcripts among the many other and more abundant human transcripts, making unbiased NGS experiments costly and ineffective in this context. Targeted sequencing could alternatively be considered to reduce the experimental burden by specifically amplifying the transcripts of interest, as is typically applied in 16S metagenomic sequencing. This type of approach could either be performed at the family or locus level. The design of family-specific degenerate primers or locus-specific primers would however require an elaborate step of primer selection ensuring both family/locus specificity (as illustrated in Pichon et al. for PCR amplification of the Pol region [18]) and compatible annealing temperature for unbiased quantification. To our knowledge, no such systematic targeted sequencing approach has been proposed so far. The work presented in this study applies such methodology on microarray using a probe selection pipeline that aims to both maximize probe efficiency and mitigate non-specific reactions, minimizing thus the analysis step for the end-user. Microarrays platforms and in particular Affymetrix instruments are now deployed in many research laboratories and the cost per experiment makes microarrays affordable compared to NGS, with a reduced time-to-result.

Two custom microarrays were previously designed in the laboratory based on a unicity criterion and a specificity score. The first meant that only candidate probes with a single perfect match were selected [29], whereas the second estimated a cross-hybridization risk using the nature and position of mispairing (mismatches, MMs

and gaps) in probe-target hybrids [13]. Training sets consisting of PM and MM probes were introduced on both arrays to evaluate and refine these strategies of cross-hybridization control. Both platforms allowed the identification of cancer-specific loci (testis [29], prostate [13, 30], colon [13] subsequently validated by qRT-PCR on a large cohort [31]) and the assignment of LTR functions [13, 29], but did not prevent cross-reactions to occur, raising the need for an improved approach.

Building on these two experiences and leveraging the high-density Affymetrix format (5 micron feature size), we introduce here a new platform HERV-V3 which, like the previous versions, aims at measuring HERVs at the locus level. The two main improvements lie in the almost complete coverage of HERVs and their ancestors as well as the introduction of a specificity criterion based on a new hybridization model, named hereafter, the Pentamer rEgion-dependent Hybridization Model (PEHM). The aim of this model is to predict the affinity of any probe-target hybrid, and therefore, to evaluate the potential of cross-hybridization by determining whether a probe of interest hybridizes only with its target. Along HERVs elements, five additional repertoires were introduced on HERV-V3 that fall in three categories, repetitive elements (MaLRs and active LINE-1 elements), non-repetitive elements (lncRNA and a selection of 1559 human genes) and common infectious viruses. While the array design is primarily aimed at identifying HERVs and MaLRs implicated in physiological and pathological processes, broader applications can be envisioned with these repertoires, such as the detection of virus replication along with the monitoring of HERVs/MaLRs and genes modulation. In the following, we successively (i) describe the main steps of the array design, (ii) compare our probesets with those of Affymetrix on 1559 common genes according to the MAQC criteria and (iii) demonstrate that for a selection of loci characterized as tissue/pathology specific, the pattern of expression observed on HERV-V3 is consistent, illustrating the relevance of such platform as research tool.

Methods

The design of the HERV-V3 array followed three main steps: (i) the genomic detection and the annotation of HERVs/MaLRs elements presented here, (ii) the development of a hybridization model to prevent cross-reactions and (iii) the design of the probes. The hybridization model was fitted on the HERV-V2 training set, made of degenerated Affymetrix probesets (see below).

Database creation

The HERV-V3 array ambitions both to cover the whole human retrovirome and provides functional annotations when possible. These annotations are primarily meant to address hypotheses on LTRs functions (i.e. promoter or polyA) and to support data interpretation at the level of gag/pol/env regions and their putative ORFs. A first step of genomic detection and annotation was performed (Fig. 1a), step which is non-trivial given that HERV classification remains incomplete [32]. To this end, two different sources of information were used, a set of prototypes associated with 42 families described in the literature [3, 21, 33, 34] for which annotations were generated in the laboratory (Additional file 1: Supplementary Notes, section 1, and Additional file 2: Table S1), and 331 Repbase consensus for which no annotation could easily be generated [35]. In the first case, prototypes were aligned on the human genome (hg19) using RepeatMasker, leading to a set of annotated HERVs called hereafter "HERVs prototypes". In the second case, fragmented HERV elements were retrieved from Dfam, a database of repetitive elements detected by RepBase consensus [36], and subsequently reconstructed (Cf Additional file 1: Supplementary Notes, section 2). This two levels strategy was devised to generate accurate annotations on elements detected by prototypes and to recover as many HERVs as possible using the representativeness of Repbase consensus. All the detected and annotated elements were finally stored in a database named hereafter HERVgDB4.

Hybridization model

Once the database created, a hybridization model, PEHM, was developed to predict the probe cross-hybridization potential (see Fig. 1b). This was made possible by an explicit modeling of MMs and gaps, allowing thus a precise quantification of mispairing. Hybridization models have been explored in the past decade, where the focus was more on modeling perfect match hybridization to improve microarray design, data interpretation [37–39] and to detect cross-reactions [40]. Here, the goal of the model is to predict the affinity of DNA hybrids with possible MMs and gaps, from which the cross-reaction potential can be deduced. The model we introduce, PEHM, is along the same lines as Li & Wong and Zhang models [40, 41] that express the probe intensity as a product of the affinity for its target with the target concentration (otherwise called expression measure):

$$I_{ij} = \theta_i.\varphi_j + \varepsilon_{ij}; \varepsilon_{ij} \sim N\left(0,\ \sigma^2\right), \sum_n {\theta_n}^2 = N \qquad (1)$$

with I_{ij} the intensity of probe j on array i, φ_j the affinity of probe j, θ_i the expression level of the gene targeted by probe j on array i and ε_{ij}, an independent identically distributed error term centered on 0. Because of the product between affinity and target concentration, a constraint is

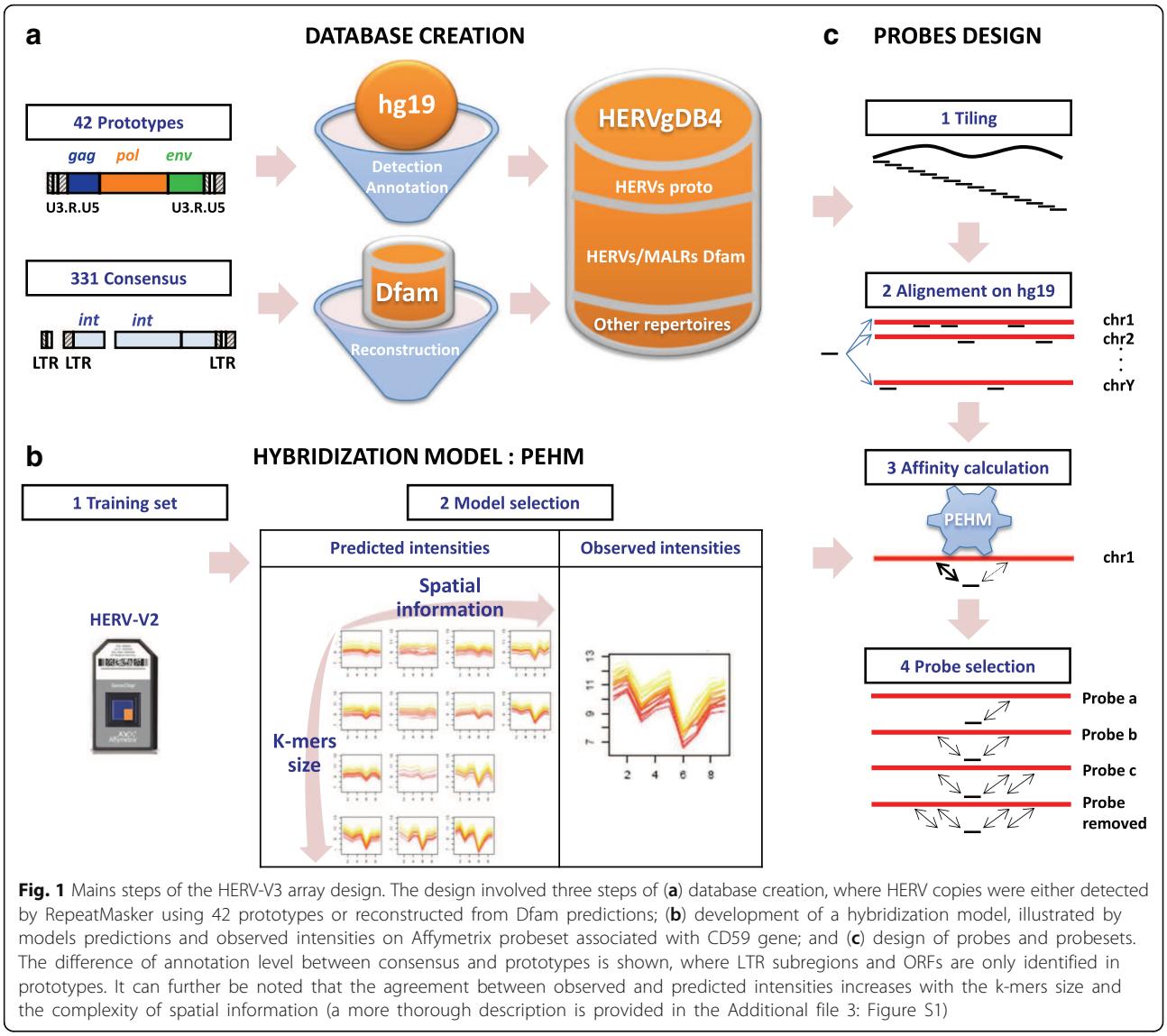


Fig. 1 Mains steps of the HERV-V3 array design. The design involved three steps of (**a**) database creation, where HERV copies were either detected by RepeatMasker using 42 prototypes or reconstructed from Dfam predictions; (**b**) development of a hybridization model, illustrated by models predictions and observed intensities on Affymetrix probeset associated with CD59 gene; and (**c**) design of probes and probesets. The difference of annotation level between consensus and prototypes is shown, where LTR subregions and ORFs are only identified in prototypes. It can further be noted that the agreement between observed and predicted intensities increases with the k-mers size and the complexity of spatial information (a more thorough description is provided in the Additional file 3: Figure S1)

required to allow parameter identifiability. An important difference between PEHM and Li & Wong is that the parameter of interest is the affinity in the first case, while it is the target concentration in the second. Consequently, instead of considering affinity as a nuisance parameter and imposing the identifiability constraint onto it [41, 42], the constraint here is imposed on the RNA quantity, where the sum squares of θ_n is set to N. Furthermore, PEHM links the probe-target affinity to the DNA sequence by modeling the affinity as a sum of k-mers effects, similarly to Zhang et al. This initial model is then extended in four ways: (i) given that DNA structural properties (i.e. flexibility, stability) depend on the interactions between neighboring base pairs, pentamers instead of dimers were used to improve affinity modeling (data not showed). (ii) While the spatial effect was previously modeled through position weights (modulating k-mers in function of their position, Zhang et al.) or by estimating k-mers at each position of the probe [37], an approximation of the latter is chosen here by considering three sub-regions of identical size in the probes. Although less precise, this approximation reduce by a factor 7 the number of parameters in comparison with Mei et al. approach. (iii) MM and gap 5-mers are taken into account as well as (iv) interactions between mismatches, following the idea that the k-mers additivity breaks down in presence of multiple MMs [43]. Overall, the affinity is expressed as follows:

$$\varphi_j = \sum_l \sum_k \beta_k^l X_{jk}^l + \sum_m \delta_m Z_{jm} \qquad (2)$$

With β_k^l the coefficient associated with k-mer k in subregion l, X_{jk}^l, the indicator matrix providing the number of k-mer k in region l of probe j, δ_m the coefficient

associated with interaction m and Z_{jm} the indicator matrix providing the presence or absence of interaction *m* in probe *j*. Although conceptually straightforward, the use of MM and gap 5-mers dramatically increased the number of parameters from 1024 to 113,664. Model parameters were estimated using the LASSO shrinkage method [44] to prevent overfitting and consequently improve the model predictions. The model training was performed in 10-folds cross-validation on the HERV-V2 training set that consists of 20 probesets derived from the Affymetrix U133 array. Each probeset contains the 10 original U133 PM probes along with 1800 degenerated MM/Gap probes including single, double MMs and single gaps, which represent a total of 37,200 probes. The data used in the model training arose from 36 microarray experiments performed on healthy and tumor tissues (colon, breast, ovary, uterus, prostate, testis, lung and placenta) carried out in a previous study [6]. Once the model defined, an "hybridization threshold" was determined on the affinity to distinguish stable from unstable hybrids in the probe design. This threshold was set such that 90% of the probes with an affinity under this threshold have intensity under the background noise. The model performances are illustrated on Additional file 3: Figure S1 (enlarged version of Fig. 1b) using Affymetrix probeset associated with CD59 gene.

Probes and probesets design

PEHM was used in the array design to select probes that are both specific and thermodynamically efficient. To do so, the number of hybridizing targets (specific and cross-hybridizing) was predicted for each candidate probe by PEHM, and only probes capable of hybridizing with one to three targets were retained. The array design involved three steps of tiling, probe selection and probeset generation (see Fig. 1c). Each region of interest was tiled into 25 bp candidate probes with a step size between 1 and 4 bp depending on the perimeter coverage and the quality of its annotation. For instance, a step of 1 bp was used for HERVs prototypes to ensure that all candidate probes were considered in this relatively small and well annotated perimeter. For each candidate probe, the affinity with its specific target was then computed to assess its thermodynamic performance. If the affinity exceeded the hybridization threshold, the probe was subsequently aligned against a reference library using BWA [45]. Two libraries were generated covering either the repetitive fraction of the genome (hg19 regions masked by RepeatMasker) or its complementary. The advantage of dividing the genome in two partitions was to reduce substantially the execution time of BWA whose complexity is in $l.n^{0.628}.m$ (l the number of probes, n the size of the reference library and m the probe size). Affinities were then calculated with PEHM for each hits, from which probes were classified into three categories: "specific", if only one hit was above the hybridization threshold, "potentially cross-hybridizing", if less than four hits exceeded the hybridization threshold and "non-specific" otherwise. In this latter case, the candidate probes were excluded. This relatively permissive strategy was designed to include as many loci as possible on HERV-V3, even those part of the most highly repetitive families. Also, given that a small proportion of HERV loci is generally expressed in a given biological context, the probability that two cross-hybridizing transcripts are simultaneously expressed is reduced.

In Mei et al., the generation of Affymetrix probesets was based on a score that maximizes probes responsiveness (quantity related to affinity), probes uniqueness (specificity) and inter-probes distance (spreadness) [37]. In HERV-V3 design, the affinity and specificity were controlled at the probe selection step, while the probeset size, the spreadness, and cross-reaction criteria were taken into account in the probeset generation step. More specifically, a probeset was required to contain between 3 and 6 probes to yield a robust estimation of gene-expression while keeping the probeset size low due to the large number of targeted elements. This relatively small lower bound was motivated by the high level of homology existing in certain families, preventing the definition of larger probesets. We therefore preferred smaller probesets than missing out loci. This point is further discussed in the evaluation of the platform performances. A probeset was restricted to a 400 bp region, in which, a maximum 30% overlap between two neighboring probes was allowed. This means that if two probes separated by less than 8 bp pass the specificity test described above, only one will be kept in the final probeset. Cross-hybridization was also mitigated at the probeset level where for a given probeset, cross-hybridizing probes had to cross-react with distinct loci and at least one probe had to be specific (with no cross-reaction). Approximately 2 weeks were necessary to run this three steps probe definition pipeline on a server (16 CPU, 128 GB of RAM).

RNA sources and ethical considerations

The technical performances were evaluated on the MAQC samples, composed of two independent samples (A, Stratagene Universal RNA, and B, Ambion Human Brain RNA) from which two titration samples were generated (C and D, consisting of 3:1 and 1:3 ratios of A to B, respectively). Each sample was performed in technical triplicate. The biological validation was, on the other hand, performed on three different tissues (colon, placenta and prostate) and two primary human cell lines (OSCAR and EBJ14). The colon (tumor and adjacent normal tissues in two

patients) and placenta RNA samples were purchased from Clinisciences and Ambion.

The prostate samples were isolated from post-surgery (radical prostatectomy) prostate cancer and prostate normal tissue, then treated by micro-dissection. Post-surgery prostate sample were provided by the Tumorothèque du Centre Hospitalier Lyon-Sud (Pierre Benite, France). The tissue samples conservation after prostate surgery in Centre Hospitalier Lyon-Sud was performed with the local ethics committee approval (Comité de Protection des Personnes de Lyon). All patients were informed through an individual notice during the hospital admission and then gave their verbal consent, as required by the French Loi de Bioéthique (2004), for the sample conservation and research use. Prostate RNAs were extracted following the Trizol protocol (Invitrogen) and purified on Rneasy columns (Qiagen). The quality of all RNA samples was assessed with the Bioanalyser 2100 capillary.

RNA extracted from the OSCAR and EBJ14 primary human cell lines were provided by the Brain Research Institute (INSERM U846, Université Lyon 1, Lyon, France). OSCAR cells consist of human embryonic stem cells (hESCs) cultured through the addition of FGF2 in the culture medium. EBJ14 (embryoid bodies) cells were obtained by culturing the OSCAR cells in non-adherent culture dishes without FGF2, environment in which cells form floating structures that spontaneously differentiate [46].

RNA amplification and labeling

The cDNA synthesis and amplification steps were performed from 16 ng of RNA using the Ovation Pico WTA System V2 kit (Nugen). Briefly, a first strand cDNA was generated from total RNA using a mixture of random and polydT DNA/RNA chimeric primers, followed by the synthesis of the complementary strand. The mRNA strand within the cDNA/mRNA complex was fragmented in order to create priming site to permit the DNA polymerase to synthesize the second cDNA strand. The double-stranded cDNA with a short DNA/RNA heteroduplex was amplified using the strand displacement based Single Primer Isothermal Amplification (SPIA) method. Schematically, RNase-H removed the RNA portion of the heteroduplex sequence and revealed a site for binding the DNA/RNA chimeric SPIA primer. DNA polymerase synthesized a new cDNA starting at the 3′ end of the primer, displacing the existing forward strand released as ssDNA. Priming with the chimeric SPIA primer recapitulated the heteroduplex creating a new substrate for RNase-H and the initiation of the next round of cDNA synthesis and ssDNA release.

The resulting amplified ssDNA was purified using the QIAquick purification kit (Qiagen), from which, total DNA concentration was measured using the NanoDrop 1000 spectrophotometer (Thermo Scientific) and the product quality was checked on the Bioanalyser 2100. Five micrograms of purified ssDNA were fragmented and labeled with the Encore Biotin Module kit (Nugen): the cDNA products were fragmented by enzymatic process into 50–100 bp fragments and subsequently labeled via enzymatic attachment of a biotin-labeled nucleotide to the 3-hydroxyl end of the fragmented cDNA. 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 min, incubated at 50 C for 5 min and centrifuged at 16,000 g for 5 min to pellet the residual salts. The HERV microarrays were pre-hybridized with 200 μL of hybridization buffer and placed under stirring (60 rpm) in an oven at 50 C for 10 min. The hybridization buffer was then replaced by the denatured hybridization cocktail. Hybridization was performed at 50 °C for 18 h 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 finally scanned using a fluorometric scanner (GeneChip scanner 3000 7G, Affymetrix).

Bioinformatics microarray analysis

Quality checks were systematically performed before microarray data analysis. The indicators examined were (i) the amplification and hybridization Affymetrix controls, (ii) the median absolute deviation versus the intensity median (MAD-Med) representation, (iii) the Normalized Unscaled Standard Error (NUSE) and (iv) the Relative Log Expression (RLE) [47].

Four pre-processing (background correction, normalization and summarization) approaches were compared, RMA [42], two alternatives to RMA and Li & Wong [41]. The two alternatives differ from RMA by their background correction step: the background noise is estimated either globally using the 15th percentiles of tryptophan probes or at the probe level using the median intensity of antigenomic probes with identical GC-content. The antigenomic probes have been introduced on exon arrays to estimate the non-specific hybridization effect related to probes GC content [48]. Their design is such that they do not match any location in the human genome and cover a wide range of GC content.

Lastly, the search for differentially expressed genes (DEG) was performed using LIMMA [49]. This method relies on a moderated t-stastistic, robust for small numbers of arrays. Q-value and fold-change thresholds of 0.01 and 2 respectively were used in the technical and biological validations. To ensure that probesets identified as differentially expressed were not in the background

noise, a threshold of 2^4 was set on the median of the technical replicates ($n = 3$), intensity for which CVs across technical replicates were under 15%.

Results and discussion

Database and microarray contents

A total of 29,859 and 169,821 HERV prototypes and HERVs Dfam were collected and stored in HERVgDB4 (see Table 1). Six additional repertoires were added to this database, (i) 228,429 MaLRs (ancestors of HERVs) retrieved from Dfam and processed in the same way as the HERVs Dfam; (ii) 192 centromeric HERV elements (absent from hg19) shown to be reactivated in HIV infection [50]; (iii) a selection of 1072 putative active LINE-1 elements arising from the union of L1Base and dbRIP databases [51, 52]; (iv) 3777 long non-coding RNAs coming from two studies [53, 54], cleared of repetitive sequences with RepeatMasker (total coverage = 366.8 Mb); (v) 289 infectious viruses and (vi) 1559 genes involved in eight pathways (immunity, inflammation, cancer, central nervous system affections, differentiation, telomere maintenance, chromatin structure and gag-like genes, see Additional file 4: Table S2). Each of those 1559 genes are targeted by three probesets, two originating from commercial Affymetrix arrays (U133 and HTA v2), and one from our design. Put another way, the expression level of any of these 1559 genes is simultaneously measured by a U133 and HTA probeset as well as a probeset designed using the PEHM model. Their relative performances, presented in the following sections, provide a simple way to validate our probe design. For simplicity, we will call these probesets gU133, gHTA, and gPEHM. To ensure that we can rely on gU133 and gHTA as internal controls, we checked whether gU133 show a similar behaviour on HERV-V3 and HG-U133 Plus 2.0 array. A large correlation ($R^2 = 0.811$, probeset level) was found on gU133 probetsets between the two arrays, supporting thus the use of gU133 and gHTA as standard for comparison (Additional file 5: Figure S2). Overall, HERV-V3 contains 372,976 elements, represented by 2.7 million probes. Probes were synthesized in sense and antisense (5.3 million in total) to accommodate with any amplification protocols and retain transcripts strand, given that some LTRs were shown to exhibit bidirectional promoter activity [55].

Platform evaluation

Following on the MAQC consortium, the technical performances of the platform were first studied based on repeatability and accuracy, which have become standard in platform evaluation [56]. Accuracy has commonly been assessed either by comparing the estimated dilution mixtures from array intensities to their theoretical values, or by computing the titration response. The former relies on the assumption that in a titration sample, the signal of a given transcript is a linear combination of the signals measured in the two original samples ($C = \alpha_C A + \beta_C B$ and $D = \alpha_D A + \beta_D B$). If this assumption is satisfied, the fractions estimated on the array should be centered on the dilution mixtures $\beta_C = 0.25$ and $\beta_D = 0.75$. The latter measures the coherence between the abundance of the hybridized RNA and the intensity measured on the array using two samples A and B and their mixture C (75% A + 25% B) and D (25% A + 75% B). This titration implies that for any gene i, if the true expression level $A_i > B_i$, then the average intensities across triplicates are expected to follow $A_i > C_i > D_i > B_i$, and conversely, if $B_i > A_i$, then $B_i > D_i > C_i > A_i$.

Table 1 Number of elements and functional sub-regions contained in HERVgDB4 (left) and designed on HERV-V3 (right) where one probeset is defined by sub-region

Repertoire	HERVgDB4 (database)		HERV-V3 (array)		
	Number of elements	Number of sub-regions	Number of elements	Number of probesets	Number of elements
HERV prototypes	29,859	90,106	29,807	45,374	29,859
HERV centromeric	192	589	24	29	192
HERV Dfam	169,821	342,482	154,535	283,641	169,821
MaLR Dfam	228,429	45,543	179,323	311,286	22,8429
LINE1	1072	4627	664	1416	1072
lncRNA	3812	3819	3777	3777	3812
Viruses	291	386	289	368	2044
gPEHM	1559	1559	1559	1559	8743
gU133	1559	NA	1559	3884	42,964
gHTA	1559	NA	1559	35,398	344,002
Affymetrix Controls	NA	NA	NA	177	20,895
Total	435,040	898,998	372,976	686,869	2,651,585

The discrepancy between the number of elements in the database and on the array is due to cross-hybridizing elements discarded during the design

This quantity was first utilized to evaluate normalization procedures. Four methods were tested, Li-Wong [41], RMA [42] and two alternatives, RMA-TRPN and RMA-GCBG, that differed by their background correction (see the Bioinformatics microarray analysis section). The methods gave similar performances except RMA-GCBG whose titration curve showed a broader spread (see Fig. 2a). Inter-methods differences were quantified by measuring the B_i/A_i ratio at which 75% of the probesets show a monotonic titration. This ratio was reached at 1.45, 1.53, 1.6, 2.19 in RMA-TRPN, Li-Wong, RMA and RMA-GCBG, which prompted us to keep RMA-TRPN in the following. In theory, PEHM could also be used for data pre-processing. However, because affinities are likely to be inferred more accurately by direct data estimation (RMA) than sequence based prediction (PEHM) and because RMA has received a large consensus in the community [57], we chose RMA for normalizing our data in this study.

We then compared our probe design with Affymetrix's approach and checked whether the quality of measure was equivalent across repertoires (genes versus REs). The repeatability and the titration response were compared across the HERVs/MaLRs, gPEHM and gU133/gHTA compartments. Because the first two repertoires target two different sets of genomic elements while deriving from the same design method, their comparison reveals how our design approach performs on cellular genes and repetitive elements. The last two, on the other hand, target the same genes while deriving from two distinct design methods. Their comparison sheds light on the relative performances between Affymetrix design method and ours. Since gPEHM and gU133/gHTA gene repertoires presented higher intensity and larger probeset size (10 and 5.8 probes/probeset on average in gU133/gHTA and gPEHM, respectively) relatively to HERVs/MaLRs (3.5 probes/probeset on average, Fig. 2b, c), comparisons were carried out after stratification by intensity and probeset size. The low intensities observed in HERVs/MaLRs elements (Fig. 2a) are due to the fact that after embryonic development, a majority of retroelements are permanently repressed [28]. The reduced probesets size can, on the other hand, be attributed to the lack of large specific regions in HERVs/MaLRs loci that could allow the definition of bigger probesets.

gPEHM probesets were consequently regenerated such that the probeset size distribution in this new compartment, named "downsized gPEHM" (dgPEHM), matches this in HERVs/MaLRs. Repeatability and accuracy

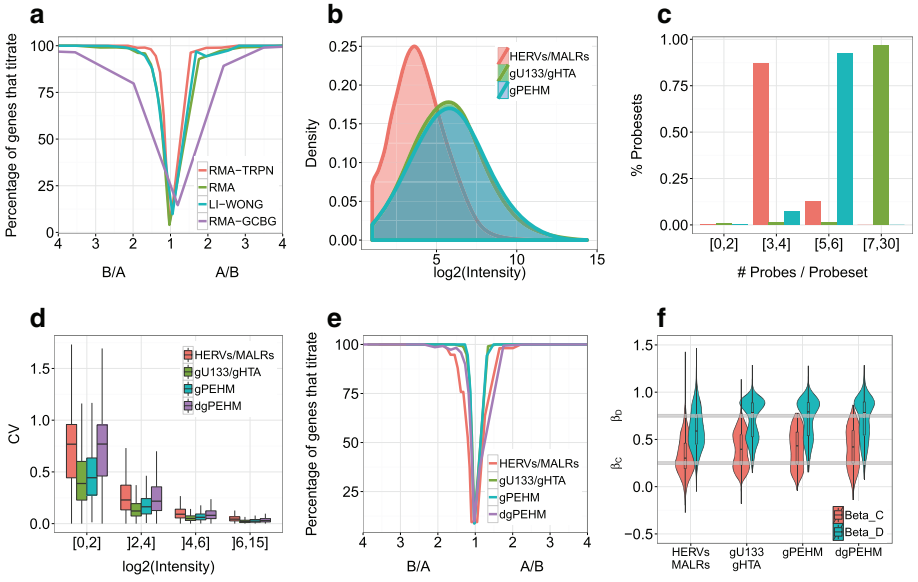


Fig. 2 Platform evaluation. **a** Pre-processing methods were evaluated on the whole array using the titration response as a function of the fold-change between samples A and B. Probesets were binned according to the fold-change values between A and B. Unlike GCBG-RMA, the three methods RMA-TPRN, RMA and Li-Wong present narrow titration curves, indicative of good performances. The two confounding factors (**b**) intensity and (**c**, same colour code as in 2**b**) probeset size distribution are represented in HERVs/MaLRs, gU133/gHTA and gPEHM compartments: the intensities are lower in HERVs/MaLRs than in genes (gPEHM, gU133/gHTA), reffecting a smaller proportion of expressed loci in the former. The three compartments, HERVs/MaLRs, gU133/gHTA, gPEHM, and downsized gPEHM (dgPEHM) are compared on (**d**) repeatability (CV) and accuracy measured both by (**e**) the titration response and (**f**) the estimated dilution mixture $(\hat{\beta}_C, \hat{\beta}_D)$. The grey horizontal lines in (**f**) symbolizes the theoretical mixture values β_C and β_D. Only probesets differentially expressed between samples A and B (fold-change A/B and B/A > 2, $P < 0.01$) were used to generate the boxplots in (**f**). The gene repertoires show similar level of repeatability and accuracy (similar median CVs, titration curves and $\hat{\beta}_C, \hat{\beta}_D$ distributions), whereas HERVs/MaLRs performances are slightly lower, due to smaller probesets

statistics were then computed. For a given intensity bin, the CVs were similar between gPEHM and gU133/gHTA, and dgPEHM and HERVs/MaLRs (see Fig. 2d), indicating that, after controlling for the confounding factors, the repeatability is similar across genomic elements and design methods. Nevertheless for a given intensity interval, HERVs/MaLRs and dgPEHM median CVs were approximately twice as large as gPEHM and gU133/gHTA due to probeset size heterogeneity. A similar trend was observed with the titration response curves (see Fig. 2e) built using probesets in the intensity bin]6; 15] : gPEHM and gU133/gHTA probesets reached the $y = 100\%$ asymptote at lower A/B and B/A ratios than HERVs/MaLRs and dgPEHM. More precisely, the ratio at which 75% of the probesets titrate is attained at $A_i/B_i = 1.43$ and 1.52 in HERVs/MaLRs and dgPEHM, whereas the same ratio was reached at 1.23 and 1.24 in gPEHM and gU133/gHTA. The evaluation of accuracy using the titration mixtures led to a different trend, the theoretical values being $\beta_C = 0.25$ and $\beta_D = 0.75$. While β_C was better estimated in HERVs/MaLRs compartments (median $\hat{\beta}_C = 0.30$) than in genes compartments (median $\hat{\beta}_C = 0.40$), the opposite was observed with D (median $\hat{\beta}_D = 0.78$ as compared to 0.59 in HERVs/MaLRs).

Overall, the observed differences in repeatability and titration response can essentially be attributed to the probeset size (restricted in HERVs/MaLRs owing to their repetitive nature) and not to the design method. The close examination of these results show that above a background noise of 2^6, the performances do not differ substantially between HERVs/MaLRs and gU133/gHTA, where the median CV is 4 and 2% respectively. Relating these performances to the probeset size, we can conclude that, in comparison with gHTA/gU133 probesets populated by 10 probes on average, (i) gPEHM show nearly identical performances while having an average probeset size of 5.8 probes, and (ii) HERVs/MaLRs have comparable performances with an average probeset size of 3.5 probes. These results are in line with Lu et al. [58] who estimated that probesets should contain at least 4 probes for reliable interpretation.

Consistency with Affymetrix design and model validation

Microarrays are generally used to measure the variation of transcript levels across two or more samples of interest. To assess the differential expression concordance among the gene repertoires, fold-changes and differentially expressed genes (DEG) were compared across the three gene repertoires. The log fold-changes between samples A and B were measured in the three gene compartments, leading to large R^2 values (see Fig. 3a–c). Although a higher correlation was obtained between the two Affymetrix repertoires ($R^2 = 0.86$), gPEHM showed a good coherence with Affymetrix fold changes ($R^2 = 0.75,\ 0.77$). These values are remarkably high given that gU133 and gPEHM probesets target genes 3′ UTR whereas gHTA covers all exons. Similarly, for a given repertoire, a large proportion of DEG are shared

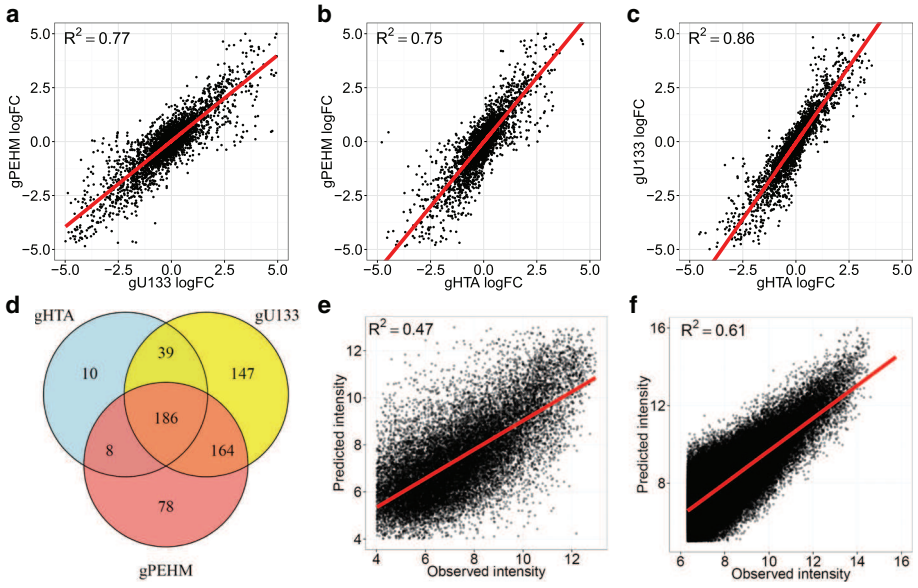


Fig. 3 Consistency with Affymetrix design and model validation. Gene expression variation is compared across the three gene compartments based on fold-change correlation (**a**–**c**) and intersections of genes differentially expressed in the gene repertoires (**d**). The hybridization model PEHM is evaluated by correlating predicted and observed intensities on gU133 probes (**e**) and HERV-V2 training set (**f**)

with the two others, these fractions being of 82.1, 75.4 and 95.9% in gPEHM, gU133 and gHTA respectively (Fig. 3b). Taken together, these results point toward a good concordance between Affymetrix and gPEHM probesets in the measure of gene expression variation, the smaller correlation with gPEHM being probably attributable to smaller probesets size in this compartment.

The last step in the platform evaluation consisted in the validation of PEHM. To this end, predicted intensities were generated from PEHM affinities and compared with those observed on the gU133 repertoire. For each gU133 probeset, the expression level was first estimated on two-third of the probes by regressing intensities onto PEHM affinities. Then, intensities were predicted on the last third of the probes by taking the product of PEHM affinities with the estimated expression level, leading to a $R^2 = 0.47$ between observed and predicted intensities (Fig. 3e). Although 0.14 lower than what was obtained on HERV-V2 ($R^2 = 0.61$, Fig. 3f), this value reflects a good ability of PEHM to model the probe-target affinity on HERV-V3, the discrepancy being probably due to the format change between HERV-V2 (11 micron, from which the model is trained), and HERV-V3 (5 micron) arrays.

When comparing the performances of PEHM ($R^2 = 0.61$) with the models proposed by Zhang et al. [40] ($R^2 = 0.98$) and Mei et al. [37] ($R^2 = 0.82$), our model may appears less predictive. This discrepancy probably reflects the differences in training set size (e.g. Zhang's model) and in whether the RNA abundance is accounted for (e.g. Mei's model). More precisely, while PEHM was evaluated on HERV-V2 training set consisting of 37,200 probes using total RNA from 15 different biological conditions, Zhang's model was evaluated on 14 probesets whose targets were spiked at 14 varying concentrations, Mei's model was, on the other hand, evaluated on all 25-mer probes spanning 90 human transcripts whose targets were spiked at 16 concentrations. Since their model was fitted for each concentration at a time, no abundance term θ was included. Of note, when testing Zhang's model and Mei's modified model (with the RNA abundance term θ added) on HERV-V2 training set, the performances found were $R^2 = 0.46$ and 0.54, respectively, that is 0.15 and 0.07 less than PEHM performance ($R^2 = 0.61$).

Validation on characterized HERV loci

Previous studies have revealed that certain HERV loci are expressed in a tissue, pathology and developmental stage specific manner and can potentially be used as biomarkers. In a perspective of biological validation, we sought to replicate these results on HERV-V3. We first evaluated whether HERV loci previously characterized by RT-PCR in placenta [29, 59, 60], prostate [61], Cheynet et al. unpublished data and colon tumor [6], showed similar expression patterns on HERV-V3. The heatmap Fig. 4a indicates that the intensities observed on the array are consistent with the expected patterns of expression: cancer and tissue specific loci are transcriptionally active only in their associated sample. The expression and tissue tropism of those loci were subsequently confirmed by RT-PCR (Additional file 6: Figure S3), with the same biological samples used in the microarray experiments. Cross-reactions were checked on the same loci by examining probesets targeting their closest paralogous sequences using blat [62]. For these probesets at risk of cross-hybridization, the intensity was under the background noise, pointing toward a high level of specificity of the array (Additional file 7: Figure S4).

Other works have shown the involvement of HERV-H in the maintenance of pluripotency, among which Wang et al. who found 550 HERV-H copies transcribed at higher level in human pluripotent stem cells (hPSCs) compared with embryoid bodies (early stage of hPSCs differentiation) [13]. To determine whether a similar enrichment in HERV-H elements was also found on HERV-V3, we searched for differentially expressed loci (DELs) between OSCAR and EBJ14, two primary human cell lines with differentiation levels similar to those in Wang et al. 563 loci were identified as differentially expressed, among which 122 belong to HERV-H family (see Fig. 4b). Given that HERV-H represents only 0.4% of the probesets on HERV-V3, this high proportion (21.7%) of HERV-H in the set of DELs argues in favor of non-random expression of HERV families (binomial test, $p < 2.10^{-16}$) and confirms the trend observed in NGS studies. It can be noted that the majority of the DELs are MaLRs, which is in line with Fort et al. who also observed the reactivation of these elements in human embryonic stem cells, although to a smaller extent than in mouse [24]. Finally, DELs positions were intersected with Wang loci, leading to 115 common regions spanning a total of 195 kb (Fig. 4c). While modest, this intersection represents 55.7% of the total DELs coverage and cannot be attributed to chance (binomial test, $p < 2.10^{-16}$). The discrepancy with Wang et al. is likely due to differences in sample (different cell lines) and assay (NGS versus microarray). Nevertheless, three HERV-H loci and one MaLR element identified as OSCAR specific on the microarray were validated by RT-PCR (Additional file 6: Figure S3), confirming thus the observed pattern on HERV-V3.

Conclusions

The recent development of high-throughput genomic approaches has enabled biologists to perform global analysis of gene expression. These technological advances have made possible to investigate disease mechanisms, identify

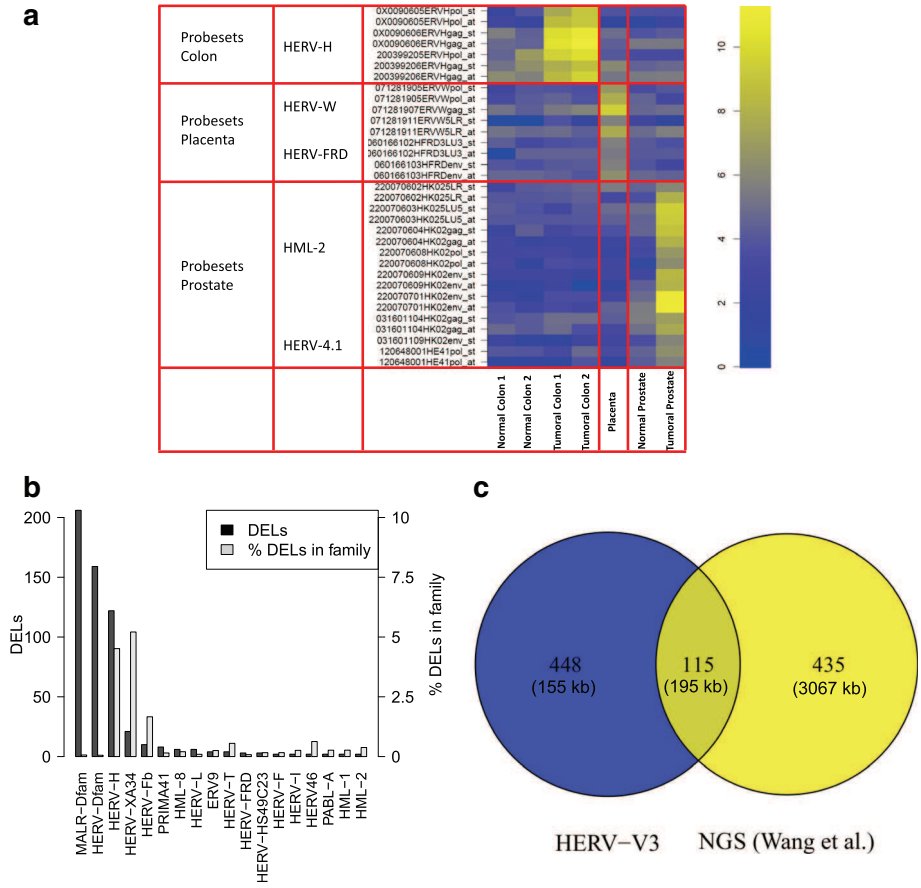


Fig. 4 Biological validation. **a** Intensity heatmap of tissue and pathology specific loci in seven HERV-V3 arrays: the observed intensities correlate well with the expected loci specificity. For each of the eight locus, the family and the probesets names are indicated (the family name and the sub-region annotation are abbreviated in the probeset name). **b** Distribution of differentially expressed loci (DELs) between hPSCs and embryoid bodies. While most of LDEs are found in MaLR-Dfam, HERV-Dfam and HERV-H, when normalized within family, the proportion of LDEs is higher in HERV-H and HERV-XA34, consistently with Wang et al. [13]. **c** Intersection between pluripotent loci identified by HERV-V3 and NGS (Wang et al.): despite a small number of shared loci (115), 55.7% of HERV-V3 loci coverage is contained in this intersection

biomarkers [63], group genes into functional pathways [64], assign function to previously unannotated genes, and evaluate the toxicity of candidate drugs [65]. Among those technologies, microarrays have been widely utilized in clinical studies for their cost-effectiveness, their rapid and mature turnaround, and their ability to provide high sensitivity and specificity results from limited biological materials (nanograms). In this work, we have presented a new high-density array allowing the examination of the whole HERVs/MaLRs transcriptome along with a selection of genes, LINE-1 elements and exogenous viruses. Such configuration opens the door to multiple applications such as the identification of enhancers and alternative promoters, the simultaneous detection of viruses and monitoring of genes and HERVs/MaLRs modulation, the identification of new biomarkers, etc. This was made possible by the careful detection and annotation of HERVs/MaLRs as well as the development of PEHM, allowing the optimization of probe performances and the control of cross-reactions. The evaluation of the platform showed that, (i) after controlling for confounding variables, similar levels of reproducibility and accuracy were obtained between Affymetrix and HERV-V3 arrays; (ii) a high consistency was found between gU133, gHTA and gPEHM probesets in term of GDE detection; (iii) for a selection of tissue/pathological loci specific, the pattern of expression reported in the literature was also observed on HERV-V3. In 2008, Mayer et al. highlighted the need for a HERV transcriptome project to study the contribution of HERVs as part of the human transcriptome [66]. Although previous works measured individual HERVs expression on a limited scale [6, 23], to our knowledge no such project has been setup yet, probably due to the technical difficulties inherent to REs. Because of its performances and exhaustiveness, HERV-V3 could benefit such project.

Additional files

Additional file 1: Supplementary notes. (DOCX 22 kb)

Additional file 2: Table S1. Chromosome locations of the prototypes used in HERVgDB4 generation. For each of the 70 prototypes associated with 42 HERV families, the family name, the sub-region annotation (full length provirus, int = gag + pol + env, LTRs, U3, R, U5 subdomains, and gag, pol, dUTPase, env genes), chromosome location (chromosome, start, end) and strand are provided. The 42 HERV families split into, 28 class I, 11 class II and 3 class III sub-families. (XLS 76 kb)

Additional file 3: Figure S1. Models performance illustrated on gene CD59. (PDF 223 kb)

Additional file 4: Table S2. List of the 1559 genes used for the PEHM hybridization model evaluation. For each gene, abbreviated name, full name, alias and accession number are provided. As indicated in the paper, each of these genes is targeted by three probesets, two derived from Affymetrix arrays U133 (GeneChip Human Genome U133 Plus 2.0 Array), HTA (GeneChip Human Transcriptome Array 2.0) and one designed using our probes and probesets selection procedure. (XLS 119 kb)

Additional file 5: Figure S2. Correlation between gU133 probesets on HG-U133 Plus 2.0 and HERV-V3 microarrays. (PDF 281 kb)

Additional file 6: Figure S3. RT-PCR validation on loci specific of placenta, colon and prostate tumor tissues, and, embryonic stem cells. (PDF 357 kb)

Additional file 7: Figure S4. HERV-V3 specificity evaluation. (PDF 315 kb)

Abbreviations

DEG: Differentially expressed gene; DEL: Differentially expressed locus; dgPEHM: downsized gPEHM; EST: Expressed sequence tag; gU133, gHTA, gPEHM: Probesets that originate from commercial Affymetrix arrays (U133 and HTA v2) and our design. They target 1559 genes involved in eight pathways (immunity, inflammation, cancer, central nervous system affections, differentiation, telomere maintenance, chromatin structure and gag-like genes, see Additional file 4: Table S2); HERV: Human endogenous retrovirus; hESCs: human stem cells; lincRNA: long intergenic non-coding RNA; lncRNA: long non-coding RNA; LTR: Long terminal repeat; MaLR: Mammalian apparent LTR-retrotransposon; MM: Mismatch probe; PEHM: Pentamer rEgion-dependent Hybridization Model; PM: Perfect match probe; RE: Repetitive elements; TSS: Transcription start site

Acknowledgements

We are grateful to Pierre Savatier and Pierre-Yves Bourillot for generously providing RNAs from undifferentiated OSCAR cells and embryoid bodies. We further thank Myriam Decaussin-Petrucci for providing us with prostatectomy samples. We also wish to thank Emmanuelle Lerat for her kind advices.

Funding

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

Availability of data and materials

Microarray expression data has been deposited on NCBI Gene Expression Ominibus and are accessible through GEO accession number GSE87134.

Authors' contributions

JB developed PEHM, implemented HERVgDB4, designed the array, performed the analyses and wrote the paper. FM defined the array content. FM and PP designed the prototypes, contributed to the design of the array (database creation, probes/probeset design) and the data interpretation. JBV and NM provided advice on the bioinformatics and statistical aspects of the project. VM and GO performed the microarray experiments. MM, OT and JT designed and carried out the comparison between U133 and HERV-V3 presented Additional file 5: Figure S2. All authors read and approved the final manuscript.

Competing interests

The authors declare they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

MAQC, colon and placenta samples were purchased from Ambion and Clinisciences.

RNA extracted from the OSCAR and EBJ14 primary human cell lines were provided by the Brain Research Institute (INSERM U846, Université Lyon 1, Lyon, France).

The prostate samples were isolated from post-surgery (radical prostatectomy) prostate cancer and prostate normal tissue, then treated by micro-dissection. Post-surgery prostate sample were provided by the Tumorothèque du Centre Hospitalier Lyon-Sud (Pierre Benite, France). The tissue samples conservation after prostate surgery in Centre Hospitalier Lyon-Sud was performed with the local ethics committee approval (Comité de Protection des Personnes de Lyon). All patients were informed through an individual notice during the hospital admission and then gave their verbal consent, as required by the French Loi de Bioéthique (2004), for the sample conservation and research use.

Publisher's Note

Author details

[1]Joint research unit, Hospice Civils de Lyon, bioMerieux, Centre Hospitalier Lyon Sud, 165 Chemin du Grand Revoyet, 69310 Pierre-Benite, France. [2]Bioinformatics Research Department, bioMerieux, 376 Chemin de l'Orme, 69280 Marcy l'Etoile, France. [3]EA 7426 Pathophysiology of Injury-induced Immunosuppression, University of Lyon1-Hospices Civils de Lyon-bioMérieux, Hôpital Edouard Herriot, 5 Place d'Arsonval, 69437 Lyon Cedex 3, France.

Received: 23 September 2016 Accepted: 28 March 2017
Published online: 08 April 2017

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Evans GA, Athanasiou M, Schultz R, Patrinos A, Morgan MJ. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.
2. Sperber GO, Airola T, Jern P, Blomberg J. Automated recognition of retroviral sequences in genomic data–RetroTector. Nucleic Acids Res. 2007;35(15):4964–76.
3. Mager DL, Medstrand P. Retroviral repeat sequences. Chichester: eLS. Wiley; 2005.
4. Gifford R, Tristem M. The evolution, distribution and diversity of endogenous retroviruses. Virus Genes. 2003;26(3):291–315.
5. Bannert N, Kurth R. Retroelements and the human genome: new perspectives on an old relation. Proc Natl Acad Sci U S A. 2004;101:14572–9.
6. Perot P, Mugnier N, Montgiraud C, Gimenez J, Jaillard M, Bonnaud B, Mallet F. Microarray-based sketches of the HERV transcriptome landscape. PLoS One. 2012;7(6):e40194.
7. Cohen CJ, Lock WM, Mager DL. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. Gene. 2009;448(2):105–14.
8. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, Waki K, Hornig N, Arakawa T, Takahashi H, Kawai J, Forrest ARR, Suzuki H, Hayashizaki Y, Hume DA, Orlando V, Grimmond SM, Carninci P. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet. 2009;41(5):563–71.
9. Mager DL, Hunter DG, Schertzer M, Freeman JD. Endogenous retroviruses provide the primary polyadenylation signal for two new human genes (HHLA2 and HHLA3). Genomics. 1999;59(3):255–63.
10. Medstrand P, Van de Lagemaat L, Dunn CA, Landry J-R, Svenback D, Mager DL. Impact of transposable elements on the evolution of mammalian gene regulation. Cytogenet Genome Res. 2005;110(1–4):342–52.
11. Li L, Feng T, Lian Y, Zhang G, Garen A, Song X. Role of human noncoding RNAs in the control of tumorigenesis. Proc Natl Acad Sci. 2009;106(31):12956–61.
12. Perot P, Bolze P-A, Mallet F. From viruses to genes: syncytins. In: Witzany G, editor. Viruses, Essential Agents of Life. Netherlands: Springer; 2012. p. 325–61.

13. Wang J, Xie G, Singh M, Ghanbarian AT, Rasko T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, Schumann GG, Chen W, Lorincz MC, Ivics Z, Hurst LD, Izsvak Z. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature. 2014;516(7531):405–9.
14. Antony JM, van Marle G, Opii W, Butterfield DA, Mallet F, Yong VW, Wallace JL, Deacon RM, Warren K, Power C. Human endogenous retrovirus glycoprotein-mediated induction of redox reactants causes oligodendrocyte death and demyelination. Nat Neurosci. 2004;7(10):1088–95.
15. Balada E, Vilardell-Tarres M, Ordi-Ros J. Implication of human endogenous retroviruses in the development of autoimmune diseases. Int Rev Immunol. 2010;29(4):351–70.
16. Christensen T. HERVs in neuropathogenesis. J Neuroimmune Pharmacol. 2010;5(3):326–35.
17. Yu HL, Zhao ZK, Zhu F. The role of human endogenous retroviral long terminal repeat sequences in human cancer. Int J Mol Med. 2013;32(4):755–762.
18. Pichon J-P, Bonnaud B, Cleuziat P, Mallet F. Multiplex degenerate PCR coupled with an oligo sorbent array for human endogenous retrovirus expression profiling. Nucleic Acids Res. 2006;34(6):e46.
19. Contreras-Galindo R, Kaplan MH, Leissner P, Verjat T, Ferlenghi I, Bagnoli F, Giusti F, Dosik MH, Hayes DF, Gitlin SD, Markovitz DM. Human endogenous retrovirus K (HML-2) elements in the plasma of people with lymphoma and breast cancer. J Virol. 2008;82(19):9329–36.
20. Forsman A, Yun Z, Hu L, Uzhameckis D, Jern P, Blomberg J. Development of broadly targeted human endogenous gammaretroviral pol-based real time PCRs Quantitation of RNA expression in human tissues. J Virol Methods. 2005;129(1):16–30.
21. Seifarth W, Frank O, Zeilfelder U, Spiess B, Greenwood AD, Hehlmann R, Leib-Mösch C. Comprehensive Analysis of Human Endogenous Retrovirus Transcriptional Activity in Human Tissues with a Retrovirus-Specific Microarray. J Virol. 2005;79(1):341–52.
22. Oja M, Peltonen J, Blomberg J, Kaski S. Methods for estimating human endogenous retrovirus activities from EST databases. BMC Bioinformatics. 2007;8(Suppl 2):S11.
23. Young GR, Mavrommatis B, Kassiotis G. Microarray analysis reveals global modulation of endogenous retroelement transcription by microbes. Retrovirology. 2014;11(1):59.
24. Fort A, Hashimoto K, Yamada D, Salimullah M, Keya CA, Saxena A, Bonetti A, Voineagu I, Bertin N, Kratz A, Noro Y, Wong C-H, de Hoon M, Andersson R, Sandelin A, Suzuki H, Wei C-L, Koseki H, The FANTOM Consortium, Hasegawa Y, Forrest ARR, Carninci P. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nat Genet. 2014;46(6):558–66.
25. Bhardwaj N, Montesion M, Roy F, Coffin JM. Differential expression of herv-k (hml-2) proviruses in cells and virions of the teratocarcinoma cell line tera-1. Viruses. 2015;7(3):939–68.
26. Sokol M, Jessen KM, Pedersen FS. Utility of next-generation rna-sequencing in identifying chimeric transcription involving human endogenous retroviruses. APMIS. 2016;124(1–2):127–39.
27. Wang J, Huda A, Lunyak VV, Jordan IK. A gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. Bioinformatics. 2010;26(20):2501–8.
28. Rowe HM, Trono D. Dynamic control of endogenous retroviruses during development. Virology. 2011;411(2):273–87.
29. Gimenez J, Montgiraud C, Pichon J-P, Bonnaud B, Arsac M, Ruel K, Bouton O, Mallet F. Custom human endogenous retroviruses dedicated microarray identifies self-induced HERV-W family elements reactivated in testicular cancer upon methylation control. Nucleic Acids Res. 2010;38(7):2229–46.
30. Perot P, Cheynet V, Decaussin-Petrucci M, Oriol G, Mugnier N, Rodriguez-Lafrasse C, Ruffion A, Mallet F. Microarray-based identification of individual HERV loci expression. application to biomarker discovery in prostate cancer. J Vis Exp. 2013;81:e50713.
31. Perot P, Mullins CS, Naville M, Bressan C, Huhns M, Gock M, Kuhn F, Volff JN, Trillet-Lenoir V, Linnebacher M, Mallet F. Expression of young HERV-H loci in the course of colorectal carcinoma and correlation with molecular subtypes. Oncotarget. 2015;6(37):40095–111.
32. Blomberg J, Benachenhou F, Blikstad V, Sperber G, Mayer J. Classification and nomenclature of endogenous retroviral sequences (ERVs): problems and recommendations. Gene. 2009;448(2):115–23.
33. De Parseval N, Lazar V, Casella J-F, Benit L, Heidmann T. Survey of human genes of retroviral origin: identification and transcriptome of the genes with coding capacity for complete envelope proteins. J Virol. 2003;77(19):10414–22.
34. Strissel PL, Ruebner M, Thiel F, Wachter D, Ekici AB, Wolf F, Thieme F, Ruprecht K, Beckmann MW, Strick R. Reactivation of codogenic endogenous retroviral (ERV) envelope genes in human endometrial carcinoma and prestages: Emergence of new molecular targets. Oncotarget. 2012;3(10):1204–19.
35. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110(1-4):462–7.
36. Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, Smit AFA, Finn RD. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res. 2013;41:D70–82.
37. Mei R, Hubbell E, Bekiranov S, Mittmann M, Christians FC, Shen M-M, Lu G, Fang J, Liu W-M, Ryder T, Kaplan P, Kulp D, Webster TA. Probe selection for high-density oligonucleotide arrays. Proc Natl Acad Sci. 2003;100(20):11237–42.
38. Ono N, Suzuki S, Furusawa C, Agata T, Kashiwagi A, Shimizu H, Yomo T. An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays. Bioinformatics. 2008;24(10):1278–85.
39. Li S, Pozhitkov A, Brouwer M. A competitive hybridization model predicts probe signal intensity on high density DNA microarrays. Nucleic Acids Res. 2008;36(20):6585–91.
40. Zhang L, Miles MF, Aldape KD. A model of molecular interactions on short oligonucleotide microarrays. Nat Biotechnol. 2003;21(7):818–21.
41. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proc Natl Acad Sci. 2001;98(1):31–6.
42. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249–64.
43. Hadiwikarta WW, Walter J-C, Hooyberghs J, Carlon E. Probing hybridization parameters from microarray experiments: nearest-neighbor model and beyond. Nucleic Acids Res. 2012;40:e138.
44. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B. 1996;58(1):267–288.
45. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589–95.
46. Chen H, Aksoy I, Gonnot F, Osteil P, Aubry M, Hamela C, et al. Reinforcement of STAT3 activity reprogrammes human embryonic stem cells to naive-like pluripotency. Nat Commun. 2015;6:7095.
47. McCall MN, Murakami PN, Lukk M, Huber W, Irizarry RA. Assessing Affymetrix GeneChip microarray quality. BMC Bioinformatics. 2011;12(1):137.
48. Affymetrix: Quality assessment of exon and gene 10 st arrays Affymetrix White Paper; 2009. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-449.
49. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3(1):1–25.
50. Contreras-Galindo R, Kaplan MH, He S, Contreras-Galindo AC, Gonzalez-Hernandez MJ, Kappes F, Dube D, Chan SM, Robinson D, Meng F, Dai M, Gitlin SD, Chinnaiyan AM, Omenn GS, Markovitz DM. HIV infection reveals widespread expansion of novel centromeric human endogenous retroviruses. Genome Res. 2013;23(9):1505–13.
51. Penzkofer T, Dandekar T, Zemojtel T. L1base: from functional annotation to prediction of active LINE-1 elements. Nucleic Acids Res. 2005;33:D498–500.
52. Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P. dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum Mutat. 2006;27(4):323–9.
53. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Morales DR, Thomas K, Presser A, Bernstein BE, Oudenaarden A, Regev A, Lander ES, Rinn JL. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci. 2009;106(28):11667–72.
54. Laurent GS, Shtokalo D, Dong B, Tackett MR, Fan X, Lazorthes S, Nicolas E, Sang N, Triche TJ, McCaffrey TA, Xiao W, Kapranov P. VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. Genome Biol. 2013;14(7):R73.
55. Dunn CA, Romanish MT, Gutierrez LE, van de Lagemaat LN, Mager DL. Transcription of two human genes from a bidirectional endogenous retrovirus promoter. Gene. 2006;366(2):335–42.
56. Mestdagh P, Hartmann N, Baeriswyl L, Andreasen D, Bernard N, Chen C, et al. Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. Nat Methods. 2014;11(8):809–15.
57. Reimers M. Making informed choices about microarray data analysis. PLoS Comput Biol. 2010;6(5):e1000786.

58. Lu J, Lee JC, Salit ML, Cam MC. Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: High-resolution annotation for microarrays. BMC Bioinformatics. 2007;8(1):108.
59. Mallet F, Bouton O, Prudhomme S, Cheynet V, Oriol G, Bonnaud B, Lucotte G, Duret L, Mandrand B. The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. Proc Natl Acad Sci U S A. 2004;101(6):1731–6.
60. Blaise S, de Parseval N, Benit L, Heidmann T. Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. Proc Natl Acad Sci. 2003;100(22):13013–8.
61. Goering W, Ribarska T, Schulz WA. Selective changes of retroelement expression in human prostate cancer. Carcinogenesis. 2011;32(10):1484–92.
62. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res. 2002;12(4):656–64.
63. Chu W, Ghahramani Z, Falciani F, Wild DL. Biomarker discovery in microarray gene expression data with Gaussian processes. Bioinformatics. 2005;21(16):3385–93.
64. Anjum S, Doucet A, Holmes CC. A boosting approach to structure learning of graphs with and without prior knowledge. Bioinformatics. 2009;25(22):2929–36.
65. Fernandes TG, Diogo MM, Clark DS, Dordick JS, Cabral JMS. High-throughput cellular microarray platforms: applications in drug discovery, toxicology and stem cell research. Trends Biotechnol. 2009;27(6):342–9.
66. Flockerzi A, Ruggieri A, Frank O, Sauter M, Maldener E, Kopper B, Wullich B, Seifarth W, Müller-Lantzsch N, Leib-Mösch C, Meese E, Mayer J. 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. 2008;9(1):354.

3.2 EXPRESSION DES HERV EN SITUATIONS NORMALES ET D'AGRESSIONS INFLAMMATOIRES

Suite à une agression inflammatoire, telle que le choc septique, des brulures graves ou un traumatisme important, le système immunitaire de l'hôte répond par deux phases concomitantes : une hyper-inflammation et une immunosuppression. La première est initialement dominante, et si le patient survit, la seconde phase devient plus importante. Durant ces réponses, l'exome de l'hôte est largement modulé, comme décrit dans l'introduction (section 1.1.2). Si le lien avec la mortalité et la survenue d'infections secondaires est assez clairement illustré, il est cependant difficile de déterminer l'état immunitaire du patient à partir de l'expression des gènes. D'autre part, dans des contextes similaires d'un point de vue immunitaire, tels que le cancer ou des maladies auto-immunes, il est connu que certains HERV sont exprimés, dont certains peuvent jouer un rôle sur la réponse immunitaire.

A l'aide de la puce HERV-V3, mais aussi de puces commerciales U133plus2 ou encore du RNAseq , on veut décrire l'expression des HERV dans le sang, suite à une agression inflammatoire, spécialement après un choc septique. On cherche également à évaluer la modulation des HERV entre patients, selon leur statut immunitaire (dont un marqueur imparfait est la mesure du mHLA-DR). Enfin, on introduit la notion de co-expression des HERV avec des gènes situés proches dans le génome, suggérant un possible rôle de ces éléments (par leur LTR) sur l'expression des gènes.

3.2.1 MODULATION APRES UNE AGRESSION INFLAMMATOIRE GRAVE

3.2.1.1 RESUME DE L'ETUDE

La première étude sur l'expression de quelques centaines de HERV dans des contextes de réponses inflammatoires graves permet de montrer qu'il existe une expression et une modulation HERV dans ces contextes comparé à l'état sain. Le but de cette étude est de montrer qu'il existe une réponse transcriptomique HERV sur trois modèles d'agression inflammatoire aiguë : des patients brulés graves, des traumatisés et des chocs septiques.

Pour cela, nous avons exploité 4 jeux de données de puces commerciales Affymetrix U133plus2, déjà existants. Bien que censés cibler des gènes conventionnels, certains probesets présents sur cette puce ciblent en réalité des HERV situés proches de ces gènes. Il y a ainsi 337 probesets qui mappent sur des éléments HERV sur cette puce. Dans cette analyse, nous montrons qu'en moyenne, seuls 26% des HERV sont exprimés dans chacun des jeux de données. Pour chaque cohorte, une vingtaine de probesets sont modulés entre patients et volontaires sains (6%). De manière très intéressante, 5 probesets sont différentiellement exprimés dans les 4 cohortes. 1 probeset est de mauvaise qualité (multi-mapping), les 4 autres sont plus fortement exprimés chez les patients que chez les volontaires sains. Nous avons ensuite réalisé des validations en RT-qPCR de ces HERV et de leur gènes situés proches, tous liés à la réponse immunitaire. Nous mettons en évidence ainsi que 2 HERV sont co-exprimés avec leur gène voisin, suggérant une régulation commune de leur expression ; et les 2 autres ont une expression autonome par rapport au gène voisin. Cette étude permet de conclure à une réponse transcriptionnelle HERV spécifique de l'inflammation, de suggérer une co-régulation de certains HERV avec leur gène voisin, impliqué dans la réponse de l'hôte, et pour aller plus loin de poser l'hypothèse d'un rôle de ces HERV sur l'expression de ces gènes.

3.2.1.2 ARTICLE

ENDOGENOUS RETROVIRUSES TRANSCRIPTIONAL MODULATION AFTER SEVERE INFECTION, TRAUMA AND BURN

Olivier Tabone, Marine Mommert, Camille Jourdan, Elisabeth Cerrato, Matthieu Legrand, Alain Lepape, Bernard Allaouchiche, Thomas Rimmelé, Alexandre Pachot, Guillaume Monneret, Fabienne Venet, François Mallet, Julien Textoris

Année 2018

En cours de révision dans *Frontiers in immunology*

Endogenous retroviruses transcriptional modulation after severe inflammatory injuries.

Olivier Tabone[1], Marine Mommert[1,2], Camille Jourdan[1], Elisabeth Cerrato[1], Matthieu Legrand[3], Alain Lepape[4], Bernard Allaouchiche[4,5], Thomas Rimmelé[1], Alexandre Pachot[1], Guillaume Monneret[1,6], Fabienne Venet[1,6], François Mallet[1,2], Julien Textoris[1,7,*]

[1] EA7426 Hospices Civils de Lyon – bioMérieux – UCBL1 "Pathophysiology of Injury Induced Immunosuppression", Groupement Hospitalier Edouard Herriot, Lyon, France.
[2] Joint research unit, Hospice Civils de Lyon, bioMerieux, Centre Hospitalier Lyon Sud, Pierre-Benite, France
[3] Department of Anesthesiology and Critical Care and Burn Unit, Groupe Hospitalier St-Louis-Lariboisière, Assistance publique - Hôpitaux de Paris, Paris, France.
[4] Hospices Civils de Lyon, Intensive Care Unit, Centre Hospitalier Lyon Sud, Pierre Bénite, France.
[5] Agressions Pulmonaires et Circulatoires dans le Sepsis APCSe VetAgro Sup UPSP 2016.A101, Centre Hospitalier Lyon-Sud, Pierre Bénite, France.
[6] Hospices Civils de Lyon, Immunology Laboratory, Groupement Hospitalier Edouard Herriot, Lyon, France.
[7] Hospices Civils de Lyon, Department of Anaesthesiology and Critical Care Medicine, Groupement Hospitalier Edouard Herriot, Université Claude Bernard Lyon 1, Lyon, France.

Correspondence**:** Julien Textoris
Laboratoire Commun de Recherche bioMérieux / HCL / UCBL1
Hôpital E. Herriot, Pavillon P
5 place d'Arsonval, 69437 LYON Cedex 03
Tel: +33 426 038747; Fax: +33 472 119 547
Email: julien.textoris@biomerieux.com
ORCID: orcid.org/0000-0002-3821-9337

number of words: 5504
number of figures: 10

Keywords

Endogenous retroviruses – severe inflammatory injuries – septic shock – burn – trauma – transcriptome – host response.

Abstract

Although human endogenous retroviruses (HERVs) expression is a growing subject of interest, no study focused before on specific endogenous retroviruses loci activation in severely injured patients. Yet, HERV reactivation is observed in immunity compromised settings like some cancers and auto-immune diseases. Our objective was to assess the transcriptional modulation of HERVs in burn, trauma and septic shock patients. We analyzed HERV transcriptome with microarray data from whole blood samples of a burn cohort (n=30), a trauma cohort (n=105) and 2 septic shock cohorts (n=28, n=51), and healthy volunteers (HV, n=60). We described expression of the 337 probesets targeting HERV from U133 plus 2.0 microarray in each dataset and then we compared HERVs transcriptional modulation of patients compared to healthy volunteers. Although all 4 cohorts contained critically ill patients, the majority of the 337 HERVs was not expressed (around 74% in mean). Each cohort had differentially expressed probesets in patients compared to HV (from 19 to 46). Strikingly, 5 HERVs were in common in all types of severely injured patients, with 4 being up-modulated in patients. We highlighted co-expressed profiles between HERV and nearby CD55 and CD300LF genes as well as autonomous HERV expression. We suggest an inflammatory-specific HERV transcriptional response, and importantly, we introduce that the HERVs close to immunity-related genes might have a role on its expression.

Introduction

Human Endogenous Retroviruses (HERVs) are former exogenous retroviruses which have infected germinal cells and became integrated in our genome million years ago (Young, Stoye, and Kassiotis 2013). These rare events happened several times in evolution. As retrotransposons, they are able to duplicate across the genome and they represent today more than 8% of our genome. Each insertion therefore led to distinct groups or families, each including multiple copies. Current classification annotates around 100 such groups.

HERV loci initially shared a common structure with exogenous retroviruses: internal protein coding regions (*gag, pro, pol, env*) flanked by two identical Long Terminal Repeats (LTRs). The accumulation of mutations and recombination events during evolution made most of these elements incomplete and defective for replication. Indeed, most of HERVs in our genome are now solo LTRs (Young, Stoye, and Kassiotis 2013) resulting from recombination between 5‘ and 3’ proviral LTRs. LTRs are critical elements that control viral gene expression either as promoters, enhancers or as polyadenylation signals. When inserted upstream, within or downstream of a “conventional” protein coding gene, LTRs can modulate its expression pattern (Cohen, Lock, and Mager 2009; Isbel and Whitelaw 2012). For example, the presence of intronic LTR can result in novel transcripts, by providing alternative promoters, enhancers or polyadenylation signals, or by altering RNA splicing (Jern and Coffin 2008; Mager et al. 1999; Dunn and Mager 2005). Very few is known about of the transcriptional modulation of such elements in pathological contexts but in cancers (like testicular cancer (J. Gimenez et al. 2010) or colorectal cancer (Pérot et al. 2015)) and auto-immune diseases (like multiple sclerosis (Laska et al. 2012; Balada, Vilardell-Tarrés, and Ordi-Ros 2010; Madeira et al. 2016)).

Few studies focused on HERVs reactivation in acute inflammatory contexts. In mice, modulation of HERVs expression has been shown to be quite specific, with signatures related to pathogen-associated molecular pattern (PAMPs) (Young et al. 2012). In human, LPS or PMA stimulations of myeloid cells revealed an increase expression of four HERVs families (Johnston et al. 2001). In vivo, HERVs expression has been detected in the plasma and whole blood samples of burn patients (Y.-J. Lee et al. 2013; K.-H. Lee et al. 2014) although the studies focused on whole HERVs families, not on specific loci. Studying HERV transcriptome modulation after severe inflammatory injuries could help to better understand pathological states of patients.

After severe injuries like septic shock, burn or trauma, leading to an important inflammatory response, we and others have shown that the blood transcriptome is highly modulated, with early and profound changes in adaptive and innate immune responses (Plassais et al. 2017; Xiao et al. 2011). Moreover in these contexts, viral reactivation is often observed, especially for Herpes Viruses (Ong et al. 2017; Textoris and Mallet 2017). This reactivation is associated with an immunosuppressive state (Walton et al. 2014). We therefore hypothesize that HERV, like latent viruses, may reactivate and be transcribed in vivo after inflammatory injuries. Given that several groups showed that some probes of commercial whole genome microarray do target HERV loci (Young, Mavrommatis, and Kassiotis 2014; Reichmann et al. 2012) (such as Affymetrix U133 plus 2), we retrospectively explored microarray datasets obtained in our lab to study the HERV transcriptome modulation in various contexts of injuries *in vivo*.

Material and Methods

Patients and sample collection

Microarray analyzed cohort:

Burns cohort: 30 severe burn patients admitted at Hospices Civils de Lyon, France (HCL) were included in a placebo-controlled, randomized, double-blind study assessing the efficacy of hydrocortisone administration on burn shock duration. Inclusion / exclusion criteria, clinical description and ethical considerations of the cohort have been previously published elsewhere (Venet et al. 2015; Plassais et al. 2017). Thirteen healthy volunteers were also recruited within Hospices Civils de Lyon to serve as controls for the transcriptional study. Whole blood samples were collected at inclusion (severe shock, before any treatment, Day 1) and in the following days (around day 2 (D2), day 5 (D5) and day 7 (D7) after inclusion).

Traumas cohort: 105 patients with severe trauma were admitted at HCL. Briefly, patients were included when they were under mechanic ventilation, with an Injury Severity Score (ISS) over 25 and were at least 18 years old. Inclusion / exclusion criteria and ethical considerations of the cohort have been previously published elsewhere (Gouel-Chéron et al. 2015). The main clinical variables are summarized on Table S1. Samples were collected at day 1 (D1) or day 2 (D2) after trauma. Data from 22 healthy volunteers were also used to make comparisons with patients (identical with septic shock cohort 2).

Septic shock cohort 1 (SS1): 28 septic shock patients and 25 HV admitted into 2 ICUs of HCL were included in this study to explore the early transcriptome modulation after septic shock. Inclusion / exclusion criteria, clinical description and ethical considerations of the cohort have been previously published elsewhere (Cazalis et al. 2014). The first blood sample was collected at the onset of shock (i.e., within 30 min after the beginning of vasoactive treatment, D0) and at day 1 (D1) and day 2 (D2) after shock.

Septic shock cohort 2 (SS2): 51 septic shock patients admitted to two Intensive Care Units (ICU) of HCL and 22 HV were included in a prognostic biomarker study. Inclusion / exclusion criteria, clinical description and ethical considerations of the cohort have been previously published elsewhere (Venet et al. 2017). Samples were collected at day 1 (D1), day 2 (D2) and day 3 (D3) after shock.

RT-qPCR validation cohorts:

Patients: Subset of cohorts used for microarray analysis were used for validation cohort: 10 burn samples at D1, 10 traumas samples at D1, 10 SS1 samples at D1, 10 SS2 samples at D1. Each subset was matched with its corresponding cohort on: Age, sex and Total Burn Surface Area (TBSA) for burns - Sex, Sepsis at D7 and Death at D28 for traumas - Age, sex and SAPS II for SS1 - Age, Sex and Death at D28 for SS2.

Healthy Volunteers: Whole blood samples were purchased from the Etablissement Français du Sang (n=12). The mean age of HV is 56, with a standard error of 9. According to the standardized procedure for blood donation, written informed consent was obtained from healthy volunteers (HVs) and personal data for blood donors were anonymized at time of blood donation and before blood transfer to a research lab.

Flow cytometry validation cohort:

Burns: Whole blood samples (EDTA tubes) from 13 burn patients sampled at D1 and D7 and admitted in Edouard Herriot hospital at Lyon, France were recruited as part of the EARLYBURN study (NCT02940171). Patients were aged from 21 to 84 (mean = 53), 12 men. The mean TBSA was 33% (from 20% to 52%). All samples from these patients were used for CD300LF protein analysis, and 7 of these 13 patients were used for *CD55* protein analysis.

Septic shocks: Whole blood samples (EDTA tubes) from 22 septic shock patients sampled at D1/D2, D3/D4/D5 and D6/D7/D8 after shock and admitted in Edouard Herriot hospital at Lyon, France were recruited as part of IMMUNOSEPSIS study (NCT02803346). Patients were aged from 23 to 81 (mean = 68), 16 men. Eleven samples were used for CD300LF protein analysis and 11 other samples for *CD55* protein analysis.

Healthy volunteers: Whole blood samples (EDTA tubes) were purchased from the Etablissement Français du Sang (n=18). Donors were aged from 21 to 63 (mean = 50), 12 men and 6 women. They were age-matched with burn and septic shock cohorts. According to the standardized procedure for blood donation, written informed consent was obtained from healthy volunteers (HVs) and personal data for blood donors were anonymized at time of blood donation and before blood transfer to a research lab.

RNA extraction and microarrays

Total RNA was extracted with PAXgene™ Blood RNA kit (PreAnalytix, Hilden, Germany). Whole blood from PAXGene™ tubes was preferred to either buffy coat or PBMCs to ensure reproducibility and avoid missing samples within the context of a clinical study. RNA integrity was assessed using Agilent 2100 Bioanalyser (Agilent Technologies, Waldbrom, Germany) and Lab-on-chip RNA 6000 Nano Assay (Agilent Technologies). Double-stranded cDNA was prepared from total RNA and an oligo-dT primer using GeneChip One-Cycle cDNA Synthesis Kit (Affymetrix, Santa Clara, United States). Three µg labeled cRNA were hybridized onto Human Genome U133 Plus 2.0 GeneChips (Affymetrix), revealed and washed using FS450 fluidic station. GeneChips were scanned using a 5G scanner (Affymetrix) and images (DAT files) were converted to CEL files using GCOS software (Affymetrix).

Microarray analysis

Microarray data are available on the Gene Expression Omnibus (GEO) website for Burn [GEO:GSE77791], SS1 [GEO:GSE57065] and SS2 [GEO:GSE95233] cohorts. The preprocessing methods were comparable in all datasets. Microarray normalization and statistical analysis were performed using R/Bioconductor (R v3.2.3). Quality assessment was performed through simpleaffy (v2.46.0) (Wilson and Miller 2005). After removing outlier samples the raw data were normalized, adjusted for background noise and summarized using the GCRMA (Guanine Cytosine Robust Multi-Array) algorithm with default parameters (Wu and Irizarry 2005). COMBAT algorithm (Johnson, Li, and Rabinovic 2007) was used to remove batch effect on Burn and Trauma cohorts. The 337 probesets from the U133 Plus2.0 microarray targeting HERVs have been identified and selected as described elsewhere (Young, Mavrommatis, and Kassiotis 2014; Reichmann et al. 2012).

All the analysis were made with R (3.2.3). The differential expression analysis was performed with Limma package (3.26.9) (Ritchie et al. 2015). A probeset was considered significantly statistically differentially expressed between two conditions when absolute log2 Fold Change was higher than 0.5 and adjusted P-values (Benjamini-Hochberg correction (Benjamini and Hochberg 1995)) lower than 0.01.

Reverse transcription and quantitative PCR:

RNA from the cohorts, according to the above criteria, and new RNA from HV were selected. RNA concentration was determined using Quant-iT RNA, BR assay on Qubit (Life Technologies, Chicago, Ilinois, United States). RNA integrity was assessed with the RNA 6000 Nano Kit on a Bioanalyzer (Agilent Technologies, Santa Clara, California, United States). Samples with RNA integrity number ≤ 6 were excluded due to poor quality RNA. Total RNA was reverse transcribed in complementary DNA (200ng in a final volume of 20 µL) using QuantiTect Reverse Transcription kit (Qiagen) as recommended by the manufacturer. The expression levels of genes (*CD55, CD300LF, SLC8A1, NFE4, PTTG1IP* and *HPRT1* as reference gene) and associated HERVs were quantified using quantitative-real time polymerase chain reaction (qPCR). qPCR were performed on a LightCycler instrument using Light Cycler 480 Probes Master for the genes and reference genes and on SYBR Green I master for HERVs. Final volume of 20µL contains 0.5µM of primers. For genes, an initial denaturation step of 10min at 95°C followed by 45 cycles, 10 sec at 95°C, 29 sec annealing at 60°C, and 1 sec extension at

72°C, Taqman) was performed. For HERVs, an initial denaturation step of 5min at 95°C followed by 45 cycles of a PCR protocol (10 sec at 95°C, 15sec at 55°C and 15sec at 72°C, SYBR Green program), melting curve protocol was performed. The Second Derivative Maximum Method was used with the LightCycler software (Release 1.5.1) to automatically determine the crossing point for individual samples. Standard curves were generated by using serial dilutions of cDNA standards prepared from purified PCR amplicons obtained with the corresponding primers (Table S2). Relative standard curves describing the PCR efficiency of selected targets were created and used to perform efficiency-corrected quantification with the LightCycler Relative Quantification Software. Targets expression normalization was performed using a selected housekeeping gene (hypoxanthine phosphoribosyltransferase 1 [HPRT1, (Friggeri et al. 2016)]), and results were expressed as normalized concentration ratio.

Flow cytometry

Sampling and staining: The following antibodies were used: anti CD14-BV510, anti CD3–BV421 and anti CD56–PECy7 from BD Biosciences; anti CD300lf-PE from BD Biosciences or anti *CD55*-APC from Biolegend; anti CD16-APC from BD Biosciences or anti CD16-FITC from Beckman Coulter (Miami, FL) and PE Mouse IgG1, κ Isotype Control from BD Biosciences or APC Mouse IgG1, κ Isotype Control from R&D System. Red blood cell lysis was performed using Versalyse lysing solution (Beckman Coulter). *CD300LF* and *CD55* expression were measured using Navios flow cytometer (Beckman-Coulter). Results were analyzed with Kaluza software (Beckman-Coulter) expressed as Medians of Fluorescence Intensity (MFI).

Statistics

Wilcoxon signed rank tests were done for RT-qPCR and flow cytometry results, by comparison between HV and each cohort of patients, for each target.

Ethics approval and consent to participate

EDTA blood tubes were obtained from EFS (Etablissement Français du Sang) and used immediately. In accordance with EFS standardized procedures for blood donation, written no-objection was obtained from healthy volunteers to use the blood for the research and personal data for blood donors were anonymized before blood transfer to our research lab.
Protocols of the discovery and validation cohorts were approved by local ethics committees. Non-opposition to inclusion in the protocols was systematically recorded from patients or next of kin.

Results

We studied the *in vivo* modulation of the HERV transcriptome in three clinical relevant models of acute inflammatory injury: a burn, a trauma and 2 septic shock cohorts. We analyzed expression from each cohort independently comparing patients with healthy volunteers. All cohorts included severely injured patients (Table 1). The 30 burn patients had a median total burn surface area (TBSA) of 70% and high severity scores (median Baux: 110, median Abbreviated Burn Severity Index (ABSI): 11). The 105 trauma patients had a median Injury Severity Score

(ISS) score of 34 and a median Simplified Acute Physiology Score II (SAPSII) of 44. The 28 septic shocks from SS1 cohort had a median SAPSII of 45 and a median Charlson score of 2. The 51 patients from SS2 cohort had a median SAPSII of 51.

As previously published (Young, Mavrommatis, and Kassiotis 2014; Reichmann et al. 2012), we extracted data from 337 probesets targeting HERVs loci from the whole genome U133 plus 2.0 microarray datasets. Among them, a majority had low expression levels, within background levels (Supplemental Figure 1). Based on hierarchical clustering analysis, 64 probesets (19%) were expressed (i.e. above background) for burns, 60 probesets (18%) for traumas, 164 for septic shock 1 (49%) and 63 for septic shocks 2 (19%). The 25% most variant probesets (n=84) across samples in each dataset revealed that several probesets were even highly expressed (Figure 1). In each dataset, the hierarchical clustering highlighted a clear difference between patients and HV, suggesting a modulation of HERV expression following injury. Interestingly, over these top 25% most variant probesets selected in each dataset (resulting of 127 distinct probesets), 44 (35%) were similarly modulated in the four datasets, and 102 (80%) in at least 2 datasets (Supplemental Figure 2). In order to analyze the HERV transcriptome modulation associated with injury, we performed a supervised analysis comparing HERV expression in injured patients at D1 (admission) and HV, in each dataset separately. The comparison (accounting for multiple testing correction with absolute fold change higher than 1 and corrected p-value lower than 0.01) between burn patients and HV resulted in 19 differentially expressed HERVs (Supplemental Figure 3A). The comparison between trauma patients and HV resulted in 27 differentially expressed HERVs (Supplemental Figure 3B). The comparison between septic shock patients and HV resulted in 19 and 46 differentially expressed HERVs for cohorts 1 and 2 respectively (Supplemental Figure 3C and D). Altogether, 56 distinct probesets targeting HERVs were differentially expressed among all 4 datasets, clearly discriminating HV from patients at ICU admission (Supplemental Figure 4, Table S3). Taking into account the global profile for each probeset, 16 (28.6%) had higher expression in patients compared to HV and 40 (71.4%) were down-modulated in patients. Interestingly, 5 probesets were differentially expressed in all 4 datasets and 16 in at least 3 of them (Figure 2A). All 5 commonly modulated probesets had consistent expression profile across the 4 datasets. Four were over-expressed in patients compared to healthy volunteers (Figure 2B). The 5th probeset, down-modulated in all datasets, maps at multiple locations in the genome and was not considered in further analyses. Among the 4 remaining modulated probesets, 1 HERV from ERV24B_Prim-int family (236982_at), is within 2kb from the *PTTG1IP* gene and 3 are within a gene. A HERV from LTR33 family (230354_at) is within an intron of *SLC8A1* gene. A HERV from MLT1H family (1556107_at) and one from LTR16B2 family (1559777_at) are located in the 3'UTR of *CD55* and *MIR3945HG* genes respectively (Table 2).

Moreover, we selected 2 other probesets of interest (1553043_a_at and 1560527_at, Figure 2C). The first one targets a MLT1D HERV located in the 3'UTR of *CD300LF*. It was up-modulated in burn and SS2 cohorts. It had a strong up-modulation at D1 in burn patients compared to HV, decreasing over the first week towards HV expression level at D7 (Supplemental figure 5). The second one targets a LTR101_Mam HERV located in a 3'UTR of a processed transcripts of *NFE4* gene. It was differentially expressed in the 2 septic shock cohorts. This probeset had the highest log2FC among the 5 septic shock-specific modulated probesets.

To validate these transcriptional HERV modulations, we designed primers on the 6 described HERV loci above, and on nearby genes by RT-qPCR (Table S2). For each targeted

region, we made multiple RT-qPCR designs. We identified several distinct patterns of expression comparing HERVs and nearby genes (Tables 2): (i) for *PTTG1IP* and *MIR3945HG* regions, we observed no or low signal from the HERV loci (data not shown), (ii) for SLC8A1 (Figure 3) and *NFE4* (Figure 4) regions, we observed a high signal from HERVs elements, but no or lower signal on the genes, (iii) for *CD55* (Figure 5) and *CD300LF* (Figure 6) regions, we observed a middle or high signal from both HERV loci and genes.

To better interpret the results, we extracted from Ensembl the genome annotation and showed in genomic context, the microarray and the RT-qPCR results of *SLC8A1* (Figure 3), *NFE4* (Figure 4), *CD55* (Figure 5) and *CD300LF* (Figure 6) regions. *SLC8A1* has 11 known transcripts. All but one are located in 3' of the LTR33 HERV element targeted by the 230354_at probeset, which is located in the first intron of SLC8A1-204 transcript (Figure 3A). The up-modulation of the LTR33 element in septic shock patients observed on microarray was confirmed by RT-qPCR (Figure 2B, Figure 3B). The up-modulation observed for other cohorts was not confirmed by RT-qPCR. The gene *SLC8A1* was not expressed in patients or HV, as seen on various microarray probesets and confirmed by RT-qPCR (SLC8A1_gene, var210, var211_212).

NFE4 gene has 2 transcripts (Figure 4A) and only one is coding for a protein (NFE4-202). The LTR101_Mam HERV element, targeted by the 2560527_at probeset, is located in 3'UTR of NFE4-201, the non-protein-coding transcript. Although the same trends are observed between microarray and RT-qPCR , the up-modulation of the LTR101_Mam element observed in septic shock patients with microarray was not statistically significant in RT-qPCR , (Figure 2C, Figure 4B). There was low or no signal on designs targeting gene transcripts (NFE4_gene and NFE4_gene_var201).

CD55 gene has 11 transcripts. The MLT1H HERV element, targeted by the 1556107_at probeset, is located in the 3'UTR of CD55-211 transcript (Figure 5A). The HERV element overlaps the 3'UTR of transcript CD55-211 and a long intergenic noncoding RNA (lincRNA, a class of long transcribed RNA molecules longer than 200 nucleotides and not coding for proteins) (Figure 5B). The up-modulation of MLT1H seen with microarray in the 4 cohorts was partially confirmed by RT-qPCR on trauma and septic shock cohorts (Figure 2B, Figure 5C). The designs targeting MLT1H or close neighborhood (PCR3, 4 and 5) presented the same profile, with a significant difference in septic shock and trauma cohorts compared to HV (PCR4). The design targeting the gene showed also up-modulation of *CD55* and a very high absolute normalized expression in patients compared to HV (Figure 5C). (Of note 1555950_a_at probeset, targeting most of *CD55* transcripts, was also up-modulated in patients, and with a high expression level (data not shown)). We also confirmed by flow cytometry on monocytes and neutrophils that *CD55* expression was higher in patients than in HV, confirming an up-modulation at the protein level in patients (Figure 5D).

The MLT1D HERV element, targeted by the 1553043_a_at probeset is located in 3'UTR of CD300LF-201, 202, 203, 204 and 207 protein-coding transcripts (Figure 6A). We made several RT-qPCR designs, targeting either the HERV locus only (PCR1) or both HERV and 3'UTR of *CD300LF* (PCR2, Figure 6B). We confirmed the expression of HERV locus, but the up-modulation seen in burn and septic shock 2 cohorts compared to HV on microarray was not confirmed by RT-qPCR, neither for gene nor for HERV designs (Figure 2C, Figure 6C). PCR1 showed no signal at all. PCR2 design showed a slight higher expression level in burn and septic shock cohorts compared to HV. We also confirmed an higher expression at the protein level by

flow cytometry on neutrophils in burn and septic shock patients, compared to HV (Figure 6D). In monocytes, protein level in burn patients at D1 seemed slightly higher than HV.

Discussion

We took advantage of previous microarray analyses on four cohorts of severely injured patients to assess the modulation of HERV transcriptome in acute inflammation. We showed that several loci were expressed and modulated after acute injury. Surprisingly, a large majority among the modulated HERVs were down-modulated in patients compared to HV, whereas a global and massive gene up-modulation has been observed after severe injuries (Xiao et al. 2011).

Five HERVs were modulated in patients compared to HV in all four datasets and 16 HERVs in at least 3 datasets, suggesting a similar inflammatory triggered modulation in all models. We validated expression profiles by RT-qPCR on 6 regions, allowing us to explore more precisely the modulation pattern of the HERVs and the neighbor genes. Interestingly, all these 6 HERVs have detected signals in RNAseq experiments from lymphoid cells and whole blood datasets (Ensembl Rnaseq tracks, (Aken et al. 2017)). Some authors already focused on HERV detection in blood of burn patients using pan-family RT-PCRs (Y.-J. Lee et al. 2013; K.-H. Lee et al. 2014).

Moreover, very few data are available in human diseases for specific loci. No study had yet evaluated the expression of HERVs in acute inflammatory contexts by using multiple cohorts with different types of inflammatory injuries.

Several groups showed that huge epigenetic modifications occur after acute inflammation, regulating transcriptional profiles in the immune system, especially in sepsis (J. L. G. Gimenez et al. 2016; Saeed et al. 2014). These epigenetic modifications may explain the polarization profiles such as tolerance or trained immunity, observed after various stimulations of innate cells (Saeed et al. 2014). We hypothesized and confirmed *in vivo* that other elements than genes, especially HERVs which are known to be tightly controlled by epigenetic modifications (Daskalakis et al. 2018), might be modulated in acute inflammatory situations. This has also been demonstrated in other pathophysiological contexts such as cancer (J. Gimenez et al. 2010; Pérot et al. 2015; Lamprecht et al. 2010; Beyer et al. 2016), where global epigenetic modifications are also observed (Chiappinelli et al. 2015; Groh and Schotta 2017).

Interestingly in cancer, epigenetic modifications that gave access to HERV cis sequences through open chromatin, have also revealed a very role in pathophysiology (Lamprecht et al. 2010; Cohen, Lock, and Mager 2009; Mager et al. 1999). Indeed, by providing alternative promoter sequences to classical protein coding genes, these epigenetic modifications explain part of the ectopic expression of myeloid-growth factor receptors in lymphoid cells (Lamprecht et al. 2010). Recently, it has been suggested that HERV could provide transcription factor binding sites, modulating immune-related gene expression, as illustrated by MER41 elements which bring STAT1 motifs to IFNγ inducible genes (Chuong, Elde, and Feschotte 2016). An exhaustive study on HERV expression with a different tool, like CHIPseq technology, would bring valuable data to find potential TFBS on HERV sequences. This underlines how HERV elements, in particular their LTRs, could modulate gene expression and the host immune response to injury. In our

study, the four commonly modulated HERVs were LTRs located nearby genes related to the immune response. In several cases (*NFE4*, *CD300LF*), we found a polyadenylation signal (AAUAAA) provided by the HERV LTR in 3' of some of the alternative transcripts of the genes. The case of *CD300LF* is interesting as this protein acts as an inhibitory receptor for myeloid cells (Alvarez-Errico et al. 2004). The LTR might stabilize specific transcripts and enhance expression of CD300LF protein, which we confirmed by flow cytometry in severe burn patients early after admission. This up regulation might participate in the compensatory anti-inflammatory response. The precise understanding of the mechanisms through which specific HERV LTRs might impact immune gene expression is not possible in such translational research setting with patient samples. This will require in the future *in vitro* experimental models to validate and understand our observations.

Our RT-qPCR validation assays also showed inter-individual variability and underlined that exploring such repertoire of our genome, repetitive sequences, may face specificity issues, and will require specific tools. Indeed, as a first attempt, we used commercial microarray where probesets were not initially designed to target HERV elements. Moreover, as the probesets targeting HERVs were initially supposed to target conventional genes, the majority of explored HERVs are close to or within a gene. To better understand HERV expression in these settings, targeting HERVs localized far from genes seems important. Until now, the lack of tool made difficult the exploration of HERV expression. It would be interesting to reproduce these analyses, with a more exhaustive technology designed to specifically target HERVs, like the HERV-V3 Affymetrix microarray we recently published (Becker et al. 2017), or even RNAseq. It will allow us to better describe the whole HERV transcriptome modulation and understand the putative global role of HERV in the host response.

Finally, it would be of importance to consider HERV expression in further blood transcriptome analyses, especially in such acute inflammatory contexts, to better understand HERV expression during host response. Such studies, based on well-defined cohorts including criteria for patient stratification and taking into account drug treatment, should allow to estimate whether HERVs could be good markers of the different immune phases that follow acute injury. More, whether HERV could become potential therapeutic targets would basically require to decipher which circulating cell type produces which HERV. It will thus be appropriate to develop dedicated cellular models to , in one hand better understand the contribution of each blood-cell type to HERV expression and in another hand how HERV expression may contribute to the cell response to stimuli.To conclude, we showed for the first time that specific HERV loci are transcribed in whole blood of ICU patients. Our design allowed us to identify specific transcriptional signatures of HERVs elements, *in vivo*, linked to the acute inflammatory response. Moreover, the similarities observed in three models of acute injuries suggest common regulatory mechanisms and a specificity of the observed modulation. We also unravel the potential regulatory role of these elements within the host immune response. Further studies are needed to better understand such mechanisms and how HERVs may contribute to the pathophysiology of the host immune response, a key part of the pathophysiology of sepsis.

List of abbreviations

HERV: Human endogenous retrovirus; LTR: Long Terminal Repeats; PAMP: pathogen-associated molecular pattern; LPS: Lipopolysaccharide; PMA: phorbol-12-myristate-13-acetate;

HV: healthy volunteers; ICU: Intensive Care Unit; TBSA: Total Burn Surface Area; ABSI: Abbreviated Burn Severity Index; ISS: Injury Severity Score; SAPSII: Simplified Acute Physiology Score II; MFI: Medians of Fluorescence Intensity;

Acknowledgements

The authors would like to gratefully thank Maria-Paola Pisano, Marie-Angélique Cazalis, Boris Meunier, Julie Mouillaux and Estelle Peronnet for their kind advices. They also thank all clinical research assistant for the collection of blood samples, especially Hélène Vallin and Valérie Cerro. Finally they gratefully thank Anne Portier and Marie-Angélique Cazalis for all experiments made on samples.

Author Contributions Statement

OT and JT designed the project, performed the analyses and wrote the paper. CJ and FV performed cytometry experiments. EC performed RT-qPCR validations. ML, AC, BA, TR recruited patients in the various cohorts. OT, MM, FV , FM, JT read and discussed the manuscript. All authors drafted or revised critically the manuscript for important intellectual contents. All authors read and approved the final manuscript.

Conflict of Interest Statement

OT, MM, CJ, EC, AP, FM and JT are employees of an *in-vitro* diagnostic company. The other authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This work was supported by bioMerieux SA and HCL. MM and OT were supported by doctoral grants from bioMerieux. In addition, OT was supported by the Association Nationale de la Recherche et de la Technologie (ANRT), convention N° 2015/1227.

Data availability statement

Microarray expression data has been deposited on NCBI Gene Expression Omnibus and are accessible through GEO accession numbers GEO:GSE77791, GEO:GSE57065 and GEO:GSE95233. Data from microarray experiment for trauma cohort are available at Hospices Civils de Lyon – bioMérieux – UCBL1 "Pathophysiology of Injury Induced Immunosuppression", Groupement Hospitalier Edouard Herriot, France.

References

Aken, Bronwen L., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Friederike Bernsdorff, Jyothish Bhai, Konstantinos Billis, et al. 2017. "Ensembl 2017." *Nucleic Acids Research* 45 (D1): D635–42. https://doi.org/10.1093/nar/gkw1104.

Alvarez-Errico, Damiana, Helena Aguilar, Friederike Kitzig, Tamara Brckalo, Joan Sayós, and Miguel López-Botet. 2004. "IREM-1 Is a Novel Inhibitory Receptor Expressed by Myeloid Cells." *European Journal of Immunology* 34 (12): 3690–3701. https://doi.org/10.1002/eji.200425433.

Balada, Eva, Miquel Vilardell-Tarrés, and Josep Ordi-Ros. 2010. "Implication of Human Endogenous Retroviruses in the Development of Autoimmune Diseases." *International Reviews of Immunology* 29 (4): 351–70. https://doi.org/10.3109/08830185.2010.485333.

Becker, Jérémie, Philippe Pérot, Valérie Cheynet, Guy Oriol, Nathalie Mugnier, Marine Mommert, Olivier Tabone, Julien Textoris, Jean-Baptiste Veyrieras, and François Mallet. 2017. "A Comprehensive Hybridization Model Allows Whole HERV Transcriptome Profiling Using High Density Microarray." *BMC Genomics* 18 (1): 286. https://doi.org/10.1186/s12864-017-3669-7.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300.

Beyer, U., S. K. Krönung, A. Leha, L. Walter, and M. Dobbelstein. 2016. "Comprehensive Identification of Genes Driven by ERV9-LTRs Reveals TNFRSF10B as a Re-Activatable Mediator of Testicular Cancer Cell Death." *Cell Death & Differentiation* 23 (1): 64–75. https://doi.org/10.1038/cdd.2015.68.

Cazalis, Marie-Angélique, Alain Lepape, Fabienne Venet, Florence Frager, Bruno Mougin, Hélène Vallin, Malick Paye, Alexandre Pachot, and Guillaume Monneret. 2014. "Early and Dynamic Changes in Gene Expression in Septic Shock Patients: A Genome-Wide Approach." *Intensive Care Medicine Experimental* 2 (1): 20. https://doi.org/10.1186/s40635-014-0020-3.

Chiappinelli, Katherine B., Pamela L. Strissel, Alexis Desrichard, Huili Li, Christine Henke, Benjamin Akman, Alexander Hein, et al. 2015. "Inhibiting DNA Methylation Causes an Interferon Response in Cancer via DsRNA Including Endogenous Retroviruses." *Cell* 162 (5): 974–86. https://doi.org/10.1016/j.cell.2015.07.011.

Chuong, Edward B., Nels C. Elde, and Cédric Feschotte. 2016. "Regulatory Evolution of Innate Immunity through Co-Option of Endogenous Retroviruses." *Science* 351 (6277): 1083–87. https://doi.org/10.1126/science.aad5497.

Cohen, Carla J., Wynne M. Lock, and Dixie L. Mager. 2009. "Endogenous Retroviral LTRs as Promoters for Human Genes: A Critical Assessment." *Gene*, Genomic Impact of Eukaryotic Transposable Elements, 448 (2): 105–14. https://doi.org/10.1016/j.gene.2009.06.020.

Daskalakis, Michael, David Brocks, Yi-Hua Sheng, Md Saiful Islam, Alzbeta Ressnerova, Yassen Assenov, Till Milde, et al. 2018. "Reactivation of Endogenous Retroviral Elements via Treatment with DNMT- and HDAC-Inhibitors." *Cell Cycle* 0 (0): 1–12. https://doi.org/10.1080/15384101.2018.1442623.

487 Dunn, Catherine A, and Dixie L Mager. 2005. “Transcription of the Human and Rodent
488 SPAM1 / PH-20 Genes Initiates within an Ancient Endogenous Retrovirus.” *BMC*
489 *Genomics* 6 (April): 47. https://doi.org/10.1186/1471-2164-6-47.
490 Friggeri, Arnaud, Marie-Angélique Cazalis, Alexandre Pachot, Martin Cour, Laurent Argaud,
491 Bernard Allaouchiche, Bernard Floccard, et al. 2016. “Decreased CX3CR1 Messenger
492 RNA Expression Is an Independent Molecular Biomarker of Early and Late Mortality
493 in Critically Ill Patients.” *Critical Care* 20: 204. https://doi.org/10.1186/s13054-016-
494 1362-x.
495 Gimenez, Jose Luis Garcia, Nieves Edurne Carbonell, Carlos Roma Mateo, Eva Garcia López,
496 Lorena Palacios, Lorena Peiro Chova, Ester Berenguer, Carla Giménez Garzo,
497 Federico V. Pallardó, and Jose Blanquer. 2016. “Epigenetics As The Driving Force In
498 Long-Term Immunosuppression.” *Journal of Clinical Epigenetics* 2 (2).
499 https://doi.org/10.21767/2472-1158.100017.
500 Gimenez, Juliette, Cécile Montgiraud, Jean-Philippe Pichon, Bertrand Bonnaud, Maud Arsac,
501 Karine Ruel, Olivier Bouton, and François Mallet. 2010. “Custom Human Endogenous
502 Retroviruses Dedicated Microarray Identifies Self-Induced HERV-W Family Elements
503 Reactivated in Testicular Cancer upon Methylation Control.” *Nucleic Acids Research*
504 38 (7): 2229–46. https://doi.org/10.1093/nar/gkp1214.
505 Gouel-Chéron, Aurélie, Bernard Allaouchiche, Bernard Floccard, Thomas Rimmelé, and
506 Guillaume Monneret. 2015. “Early Daily MHLA-DR Monitoring Predicts Forthcoming
507 Sepsis in Severe Trauma Patients.” *Intensive Care Medicine* 41 (12): 2229–30.
508 https://doi.org/10.1007/s00134-015-4045-1.
509 Groh, Sophia, and Gunnar Schotta. 2017. “Silencing of Endogenous Retroviruses by
510 Heterochromatin.” *Cellular and Molecular Life Sciences*, February, 1–11.
511 https://doi.org/10.1007/s00018-017-2454-8.
512 Isbel, Luke, and Emma Whitelaw. 2012. “Endogenous Retroviruses in Mammals: An
513 Emerging Picture of How ERVs Modify Expression of Adjacent Genes.” *BioEssays:*
514 *News and Reviews in Molecular, Cellular and Developmental Biology* 34 (9): 734–38.
515 https://doi.org/10.1002/bies.201200056.
516 Jern, Patric, and John M. Coffin. 2008. “Effects of Retroviruses on Host Genome Function.”
517 *Annual Review of Genetics* 42 (1): 709–32.
518 https://doi.org/10.1146/annurev.genet.42.110807.091501.
519 Johnson, W. Evan, Cheng Li, and Ariel Rabinovic. 2007. “Adjusting Batch Effects in
520 Microarray Expression Data Using Empirical Bayes Methods.” *Biostatistics (Oxford,*
521 *England)* 8 (1): 118–27. https://doi.org/10.1093/biostatistics/kxj037.
522 Johnston, J. B., C. Silva, J. Holden, K. G. Warren, A. W. Clark, and C. Power. 2001. “Monocyte
523 Activation and Differentiation Augment Human Endogenous Retrovirus Expression:
524 Implications for Inflammatory Brain Diseases.” *Annals of Neurology* 50 (4): 434–42.
525 Lamprecht, Björn, Korden Walter, Stephan Kreher, Raman Kumar, Michael Hummel, Dido
526 Lenze, Karl Köchert, et al. 2010. “Derepression of an Endogenous Long Terminal
527 Repeat Activates the CSF1R Proto-Oncogene in Human Lymphoma.” *Nature Medicine*
528 16 (5): 571–79. https://doi.org/10.1038/nm.2129.
529 Laska, Magdalena Janina, Tomasz Brudek, Kari Konstantin Nissen, Tove Christensen, Anné
530 Møller-Larsen, Thor Petersen, and Bjørn Andersen Nexø. 2012. “Expression of HERV-
531 Fc1, a Human Endogenous Retrovirus, Is Increased in Patients with Active Multiple

Sclerosis." *Journal of Virology* 86 (7): 3713–22. https://doi.org/10.1128/JVI.06723-11.

Lee, Kang-Hoon, HyungChul Rah, Tajia Green, Young-Kwan Lee, Debora Lim, Jean Nemzek, Wendy Wahl, David Greenhalgh, and Kiho Cho. 2014. "Divergent and Dynamic Activity of Endogenous Retroviruses in Burn Patients and Their Inflammatory Potential." *Experimental and Molecular Pathology* 96 (2): 178–87. https://doi.org/10.1016/j.yexmp.2014.02.001.

Lee, Yun-Jung, Byung-Hoon Jeong, Jae-Bong Park, Hyung-Joo Kwon, Yong-Sun Kim, and In-Suk Kwak. 2013. "The Prevalence of Human Endogenous Retroviruses in the Plasma of Major Burn Patients." *Burns* 39 (6): 1200–1205. https://doi.org/10.1016/j.burns.2012.12.013.

Madeira, Alexandra, Ingrid Burgelin, Hervé Perron, Francois Curtin, Alois B. Lang, and Raphael Faucard. 2016. "MSRV Envelope Protein Is a Potent, Endogenous and Pathogenic Agonist of Human Toll-like Receptor 4: Relevance of GNbAC1 in Multiple Sclerosis Treatment." *Journal of Neuroimmunology* 291 (February): 29–38. https://doi.org/10.1016/j.jneuroim.2015.12.006.

Mager, D. L., D. G. Hunter, M. Schertzer, and J. D. Freeman. 1999. "Endogenous Retroviruses Provide the Primary Polyadenylation Signal for Two New Human Genes (HHLA2 and HHLA3)." *Genomics* 59 (3): 255–63. https://doi.org/10.1006/geno.1999.5877.

Ong, David S. Y., Marc J. M. Bonten, Cristian Spitoni, Verduyn Lunel, Frans M, Jos F. Frencken, Janneke Horn, et al. 2017. "Epidemiology of Multiple Herpes Viremia in Previously Immunocompetent Patients With Septic Shock." *Clinical Infectious Diseases* 64 (9): 1204–10. https://doi.org/10.1093/cid/cix120.

Osler, Turner, Laurent G. Glance, and David W. Hosmer. 2010. "Simplified Estimates of the Probability of Death after Burn Injuries: Extending and Updating the Baux Score." *The Journal of Trauma* 68 (3): 690–97. https://doi.org/10.1097/TA.0b013e3181c453b3.

Pérot, Philippe, Christina Susanne Mullins, Magali Naville, Cédric Bressan, Maja Hühns, Michael Gock, Florian Kühn, et al. 2015. "Expression of Young HERV-H Loci in the Course of Colorectal Carcinoma and Correlation with Molecular Subtypes." *Oncotarget* 6 (37): 40095–111.

Plassais, Jonathan, Fabienne Venet, Marie-Angélique Cazalis, Diane Le Quang, Alexandre Pachot, Guillaume Monneret, Sylvie Tissot, and Julien Textoris. 2017. "Transcriptome Modulation by Hydrocortisone in Severe Burn Shock: Ancillary Analysis of a Prospective Randomized Trial." *Critical Care* 21 (1): 158. https://doi.org/10.1186/s13054-017-1743-9.

Reichmann, Judith, James H. Crichton, Monika J. Madej, Mary Taggart, Philippe Gautier, Jose Luis Garcia-Perez, Richard R. Meehan, and Ian R. Adams. 2012. "Microarray Analysis of LTR Retrotransposon Silencing Identifies Hdac1 as a Regulator of Retrotransposon Expression in Mouse Embryonic Stem Cells." *PLOS Comput Biol* 8 (4): e1002486. https://doi.org/10.1371/journal.pcbi.1002486.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. "Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47. https://doi.org/10.1093/nar/gkv007.

Saeed, Sadia, Jessica Quintin, Hindrik H. D. Kerstens, Nagesha A. Rao, Ali Aghajanirefah, Filomena Matarese, Shih-Chin Cheng, et al. 2014. “Epigenetic Programming of Monocyte-to-Macrophage Differentiation and Trained Innate Immunity.” *Science* 345 (6204): 1251086. https://doi.org/10.1126/science.1251086.

Textoris, Julien, and François Mallet. 2017. “Immunosuppression and Herpes Viral Reactivation in Intensive Care Unit Patients: One Size Does Not Fit All.” *Critical Care* 21 (August): 230. https://doi.org/10.1186/s13054-017-1803-1.

Venet, Fabienne, Jonathan Plassais, Julien Textoris, Marie-Angélique Cazalis, Alexandre Pachot, Marc Bertin-Maghit, Christophe Magnin, Thomas Rimmelé, Guillaume Monneret, and Sylvie Tissot. 2015. “Low-Dose Hydrocortisone Reduces Norepinephrine Duration in Severe Burn Patients: A Randomized Clinical Trial.” *Critical Care (London, England)* 19: 21. https://doi.org/10.1186/s13054-015-0740-0.

Venet, Fabienne, Jeremy Schilling, Marie-Angélique Cazalis, Julie Demaret, Fanny Poujol, Thibaut Girardot, Christelle Rouget, et al. 2017. “Modulation of LILRB2 Protein and MRNA Expressions in Septic Shock Patients and after Ex Vivo Lipopolysaccharide Stimulation.” *Human Immunology*, March. https://doi.org/10.1016/j.humimm.2017.03.010.

Walton, Andrew H., Jared T. Muenzer, David Rasche, Jonathan S. Boomer, Bryan Sato, Bernard H. Brownstein, Alexandre Pachot, et al. 2014. “Reactivation of Multiple Viruses in Patients with Sepsis.” *PLoS ONE* 9 (6): e98819. https://doi.org/10.1371/journal.pone.0098819.

Wilson, Claire L., and Crispin J. Miller. 2005. “Simpleaffy: A BioConductor Package for Affymetrix Quality Control and Data Analysis.” *Bioinformatics (Oxford, England)* 21 (18): 3683–85. https://doi.org/10.1093/bioinformatics/bti605.

Wu, Zhijin, and Rafael A. Irizarry. 2005. “Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Arrays.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 12 (6): 882–93. https://doi.org/10.1089/cmb.2005.12.882.

Xiao, Wenzhong, Michael N. Mindrinos, Junhee Seok, Joseph Cuschieri, Alex G. Cuenca, Hong Gao, Douglas L. Hayden, et al. 2011. “A Genomic Storm in Critically Injured Humans.” *Journal of Experimental Medicine* 208 (13): 2581–90. https://doi.org/10.1084/jem.20111354.

Young, George R., Urszula Eksmond, Rosalba Salcedo, Lena Alexopoulou, Jonathan P. Stoye, and George Kassiotis. 2012. “Resurrection of Endogenous Retroviruses in Antibody-Deficient Mice.” *Nature* 491 (7426): 774–78. https://doi.org/10.1038/nature11599.

Young, George R, Bettina Mavrommatis, and George Kassiotis. 2014. “Microarray Analysis Reveals Global Modulation of Endogenous Retroelement Transcription by Microbes.” *Retrovirology* 11 (1): 59. https://doi.org/10.1186/1742-4690-11-59.

Young, George R, Jonathan P Stoye, and George Kassiotis. 2013. “Are Human Endogenous Retroviruses Pathogenic? An Approach to Testing the Hypothesis.” *Bioessays* 35 (9): 794–803. https://doi.org/10.1002/bies.201300049.

Figure Legends

Supplemental figure 1: Heatmap representation of HERVs in three models of injury. Heatmap of the 337 probesets targeting HERVs in the four datasets: burn, trauma and 2 septic shock cohorts. Probesets are in rows and samples in columns. Samples are annotated (colored bars on the top) by type of samples (HV in pink, patients in cyan) and day after inclusion (blue scaled). Expression levels are color-coded from blue (low expression) to red (high expression). Similar patterns of expression are highlighted through hierarchical clustering of probesets (rows) with Euclidean distance and complete clustering method. **(A)** Expression levels in burns. **(B)** Expression levels in traumas. **(C)** Expression levels in septic shock 1. **(D)** Expression levels in septic shock 2. On each heatmap, the percentage of probesets with low intensity is shown.

Figure 1: Heatmap representation of HERVs in three models of injury. Heatmap of the 25% most variant probesets targeting HERVs in the four datasets: burn, trauma and 2 septic shock cohorts. Probesets are in rows and samples in columns. Samples are annotated (colored bars on the top) by type of samples (HV in pink, patients in cyan) and day after inclusion (blue scaled). Expression levels are color-coded from blue (low expression) to red (high expression). Similar patterns of expression are highlighted through hierarchical clustering of probesets (rows) and samples (columns) with Euclidean distance and complete clustering method. **(A)** Expression levels in burn patients. **(B)** Expression levels in trauma patients. **(C)** Expression levels in septic shock 1 patients. **(D)** Expression levels in septic shock 2 patients.

Supplemental figure 2: Most variant HERVs in severely injured patients. Venn diagram of the 84 most variant HERV probesets (25%) selected in each of the four datasets.

Supplemental figure 3: Volcano plots of differentially expressed HERVs. (A) in burn cohort. **(B)** in trauma cohort. **(C)** in septic shock cohort 1 and **(D)** in septic shock cohort 2. The x-axis represents the log2 fold change between patient and HV, the y-axis the –log10 of adjusted p-values. Each point represents a probeset targeting HERV, in red the statistically differentially expressed between patients at D1 and HV. On each volcano plot, the number indicates the number of differentially expressed probesets.

Supplemental figure 4: Heatmap representation of the modulated HERVs in severely injured patients at D1. Heatmap of the 56 differentially expressed probesets in at least 1 dataset. On the top bar, samples are color-coded in blue for HV and in red for Patients. On the bar below, samples are in green for Burn study, in yellow for Trauma study, in purple for Septic Shock 1 (SS1) study and in light red for Septic Shock 2 (SS2). Probesets are in rows and samples in columns. Expression levels from each cohort have been normalized (centered and reduced). Normalized expression levels are color-coded from blue (low expression) to red (high expression). Similar patterns of expression are highlighted through hierarchical clustering of probesets (rows) and samples (columns) with Euclidean distance and complete clustering method.

Figure 2: Differentially expressed HERVs in severely injured patients. (A) Venn diagram of differentially expressed HERVs for each dataset. **(B)** Expression profiles of commonly modulated probesets targeting HERVs in the 4 datasets, at D1. **(C)** Expression profiles of 2 selected probesets targeting HERVs. Boxes are color-coded by cohort. For each graphic, from top to bottom, title contains: probeset name, HERV name and closest gene.

Supplemental figure 5: Differentially expressed HERVs in severely injured patients. (A) Venn diagram of differentially expressed HERVs for each dataset. **(B)** Expression profiles of

commonly modulated probesets targeting HERVs in the 4 datasets. Boxes are color-coded by day after inclusion. **(C)** Expression profiles of 2 selected probesets targeting HERVs. For each graphic, from top to bottom, title contains: probeset name, HERV name and closest gene.

Figure 3: LTR33 HERV and *SLC8A1* gene expression. (A) *SLC8A1* genomic region, with the position of HERV in green, probeset in dark blue and PCR designs in purple. **(B)** Expression levels of specific transcripts by RT-qPCR, as described in A, in HV and patients at D1. Expression levels (copy number / µl) were normalized with reference gene (*HPRT1*). Boxes are color-coded by cohort. Statistically significant difference with HV is marked by * (Wilcoxon signed rank test, p-value <0.05).

Figure 4 : LTR101_Mam HERV and *NFE4* gene expression. (A) *NFE4* genomic region, with the position of HERV in green, of probeset in dark blue, of PCR designs in purple. **(B)** Expression levels of specific transcripts by RT-qPCR, as described in A, in HV and patients at D1. Expression levels (copy number / µl) were normalized with reference gene (*HPRT1*). Boxes are color-coded by cohort. Statistically significant difference with HV is marked by * (Wilcoxon signed rank test, p-value <0.05).

Figure 5: *CD55* associated HERV. (A) *CD55* genomic region, with the positions of HERV in green, of probeset in dark blue, of PCR designs in purple. **(B)** Zoom in genomic region of HERV showing PCR designs in detail. **(C)** Expression levels of specific transcripts by RT-qPCR , as described in A and B, in HV and patients at D1. Expression levels (copy number / µl) were normalized with reference gene (*HPRT1*). Boxes are color-coded by cohort. **(D)** Protein expression levels (MFI), on monocytes (left) and neutrophils (right) from 8 burn patients (red), 11 septic shock patients (blue) and 9 HV (purple). Columns ISO B, ISO SS and ISO HV correspond to isotypes for burn, septic shock and HV respectively. Statistically significant difference with HV is marked by * (Wilcoxon signed rank test, p-value <0.05).

Figure 6: *CD300LF* associated HERV. (A) *CD300LF* genomic region, with the positions of HERV in green, of probeset in dark blue, of PCR designs in purple. **(B)** Zoom in genomic region of HERV showing PCR designs in detail. **(C)** Expression levels of specific transcripts by RT-qPCR , as described in A and B, in HV and patients at D1. Expression levels (copy number / µl) were normalized with reference gene (*HPRT1*). Boxes are color-coded by cohort. **(D)** Protein expression levels (MFI), on monocytes (left) and neutrophils (right) from 14 burn patients (red), 11 septic shock patients (blue) and 10 HV (purple). Columns ISO B, ISO SS and ISO HV correspond to isotypes for burn, septic shock and HV respectively. Statistically significant difference with HV is marked by * (Wilcoxon signed rank test, p-value <0.05)

Tables

Table 1: Patients characteristics of burn, trauma and septic shock cohorts included in microarray analyses. TBSA: Total Burn Surface Area (severe patient > 30%) ; Baux score: Predictor of mortality due to severe burns (severe patient > 100); ABSI: Abbreviated Burn Severity Index (severe patient > 9) ; ISS: Injury Severity Score (severe patient > 15) ; SOFA: Sequential organ failure assessment score (severe patient ≥ 3) ; SAPSII: Simplified Acute Physiology Score II (severe patient > 30). In each cohort, mortality has been assessed at the 28th day after ICU admission.

Variable	Burn (n=30)	Trauma (n=105)	Septic shock 1 (n=28)	Septic shock 2 (n=51)
Age, years	48 [39-55]	38 [25-54]	62 [54-76]	65 [53-74]
Gender, women, n (%)	8 (27%)	34 (32%)	9 (32%)	18 (35%)
Weight, kg	94 [77-104]	78 [67-92]	-	-
TBSA (%)	70 [48-84]	-	-	-
Baux score	110 [102-125]	-	-	-
ABSI score	11 [10-12]	-	-	-
ISS score	-	34 [29-41]	-	-
SOFA score	-	5 [1-7]	10 [9-13]	10 [8-12]
SAPSII score	-	44 [29-56]	45 [34-56]	51 [43-62]
Secondary septic shock	12 (40%)	29 (28%)	-	-
ICU length of stay, days	66 [22-89]	9 [5-17]	10 [5-14]	-
Mortality, n (%)	8 (27%)	4 (4%)	5 (18%)	17 (33%)

Table2: Genomic and transcriptomic features of the 6 probesets of interest.

Transcriptomic & genomic features	HERV Probesets [a]					
	1556107_at	**230354_at**	**1553043_a_at**	**1560527_at**	**1559777_at**	**236982_at**
Patients vs. HVs	UP	UP	UP (for Burn & SS2)	UP (for SS1 & SS2)	UP	UP
log2FC [b] in						
Burn	1.13	1.73	1.48	-0.55	1.05	0.77
Trauma	1.31	1.47	-0.33	1.50	0.79	1.57
SS1	0.57	2.03	0.26	0.72	1.07	0.90
SS2	1.45	2.97	1.08	2.00	1.26	1.12
Confirmed [c]	Yes*	Yes*	Yes*	Yes*	No	No
HERV family	MLT1H	LTR33	MLT1D	LTR101_Mam	LTR16B2	ERV24B_Prim-int
HERV coordinates [d]	chr1 207372720-207272854	chr2 40545338-40545778	chr17 74694268-74694744	chr7 102988743-102988923	chr4 184844993-184845324	chr21 44875454-44876122
Closest gene	CD55	SLC8A1	CD300LF	NFE4	MIR3945HG	PTTG1IP
Localization [e]	3' UTR	intron 1	3'UTR	3'end	3' UTR	promoter region

a HGU133plus2 Affymetrix probesets mapping on a HERV locus

b A positive log2 FC means that the probeset is more expressed in patients than in HV.

c Confirmed by RT-qPCR.

**Expression confirmed. Modulation between patients and HV not always statistically confirmed, mainly due to high inter-individual variability.*

d Grch38 genomic coordinates of HERV locus.

e Localization of HERV element according to the nearby gene.

Figure 1

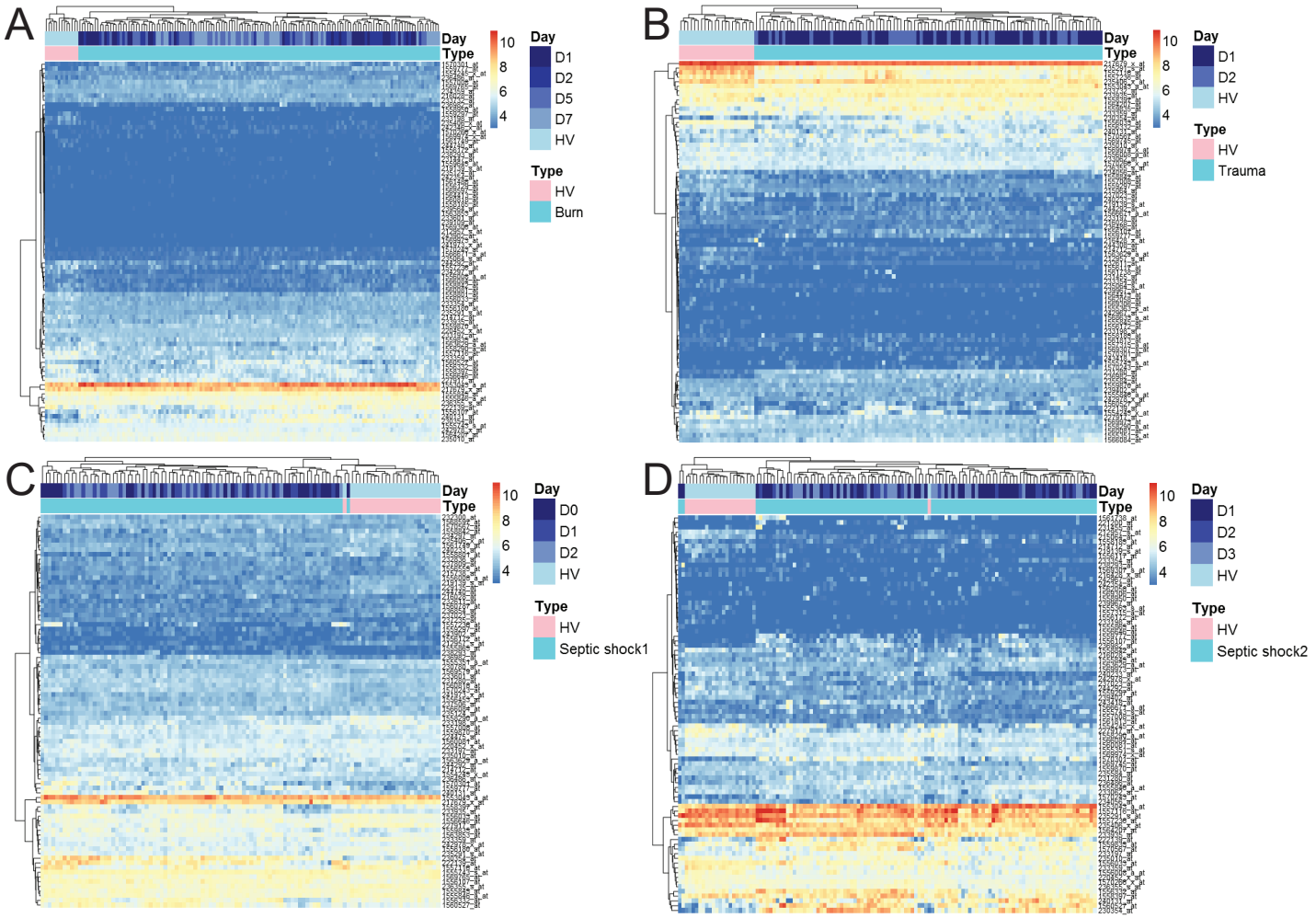

A
Day
Type
10
8
6
4
Day
D1
D2
D5
D7
HV
Type
HV
Burn
B
Day
Type
10
8
6
4
Day
D1
D2
HV
Type
HV
Trauma
C
Day
Type
10
8
6
4
Day
D0
D1
D2
HV
Type
HV
Septic shock1
D
Day
Type
10
8
6
4
Day
D1
D2
D3
HV
Type
HV
Septic shock2

Figure 2

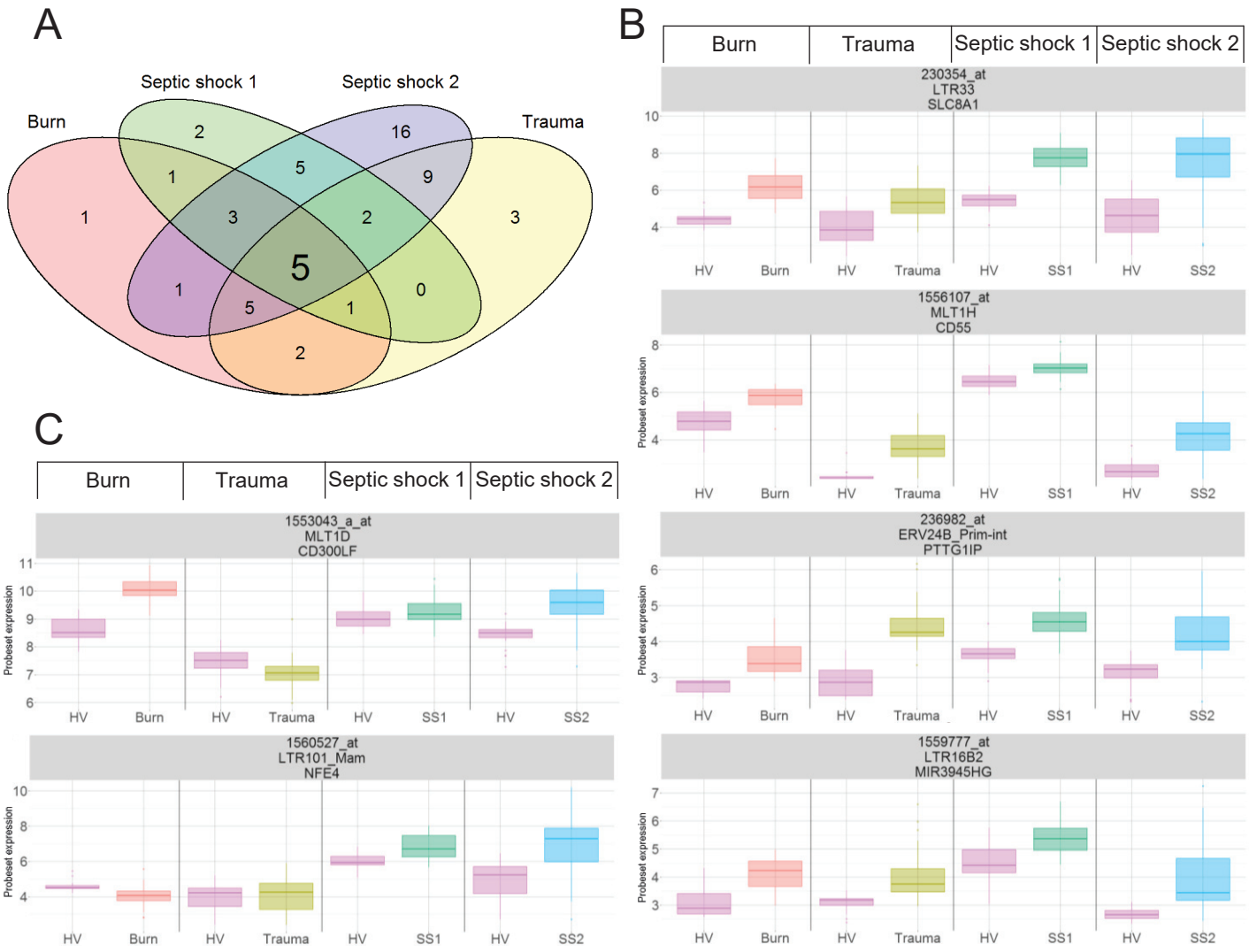
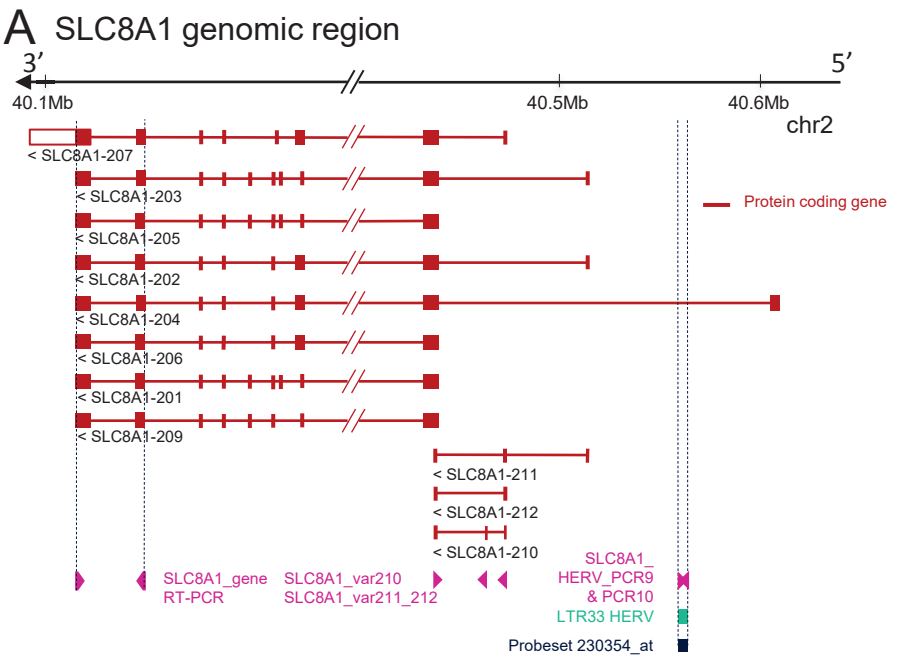

A
Septic shock 1
Septic shock 2
Burn
Trauma
2
16
1
5
9
1
3
2
3
5
1
0
5
1
2
B
Burn
Trauma
Septic shock 1
Septic shock 2
230354_at
LTR33
SLC8A1
1556107_at
MLT1H
CD55
236982_at
ERV24B_Prim-int
PTTG1IP
1559777_at
LTR16B2
MIR3945HG
Probeset expression
HV
Burn
Trauma
SS1
SS2
C
1553043_a_at
MLT1D
CD300LF
1560527_at
LTR101_Mam
NFE4

Figure 3


A SLC8A1 genomic region
3'
5'
40.1Mb
40.5Mb
40.6Mb
chr2
< SLC8A1-207
< SLC8A1-203
< SLC8A1-205
< SLC8A1-202
< SLC8A1-204
< SLC8A1-206
< SLC8A1-201
< SLC8A1-209
< SLC8A1-211
< SLC8A1-212
< SLC8A1-210
Protein coding gene
SLC8A1_gene RT-PCR
SLC8A1_var210 SLC8A1_var211_212
SLC8A1_ HERV_PCR9 & PCR10
LTR33 HERV
Probeset 230354_at

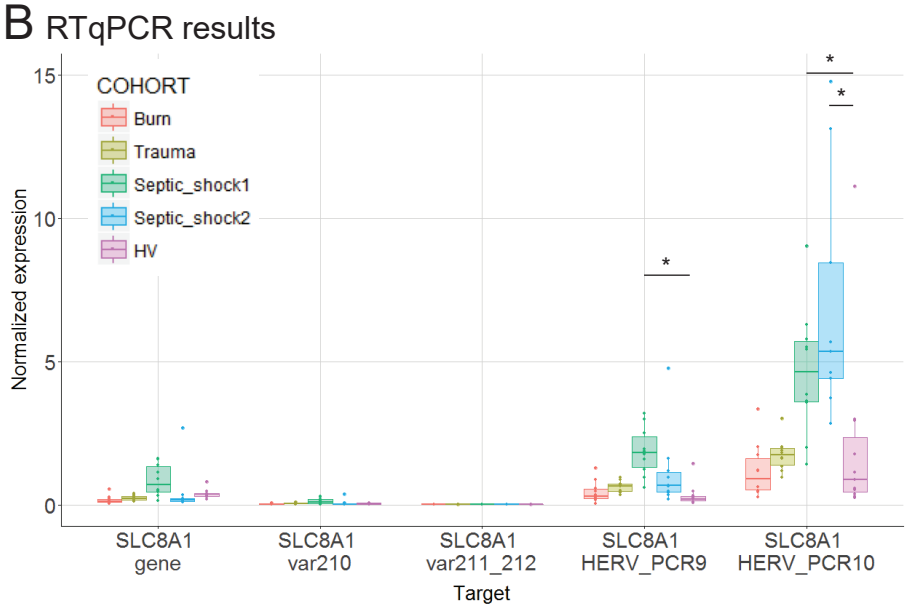

B RTqPCR results
COHORT
Burn
Trauma
Septic_shock1
Septic_shock2
HV
Normalized expression
15
10
5
0
*
*
*
SLC8A1 gene
SLC8A1 var210
SLC8A1 var211_212
SLC8A1 HERV_PCR9
SLC8A1 HERV_PCR10
Target

Figure 4

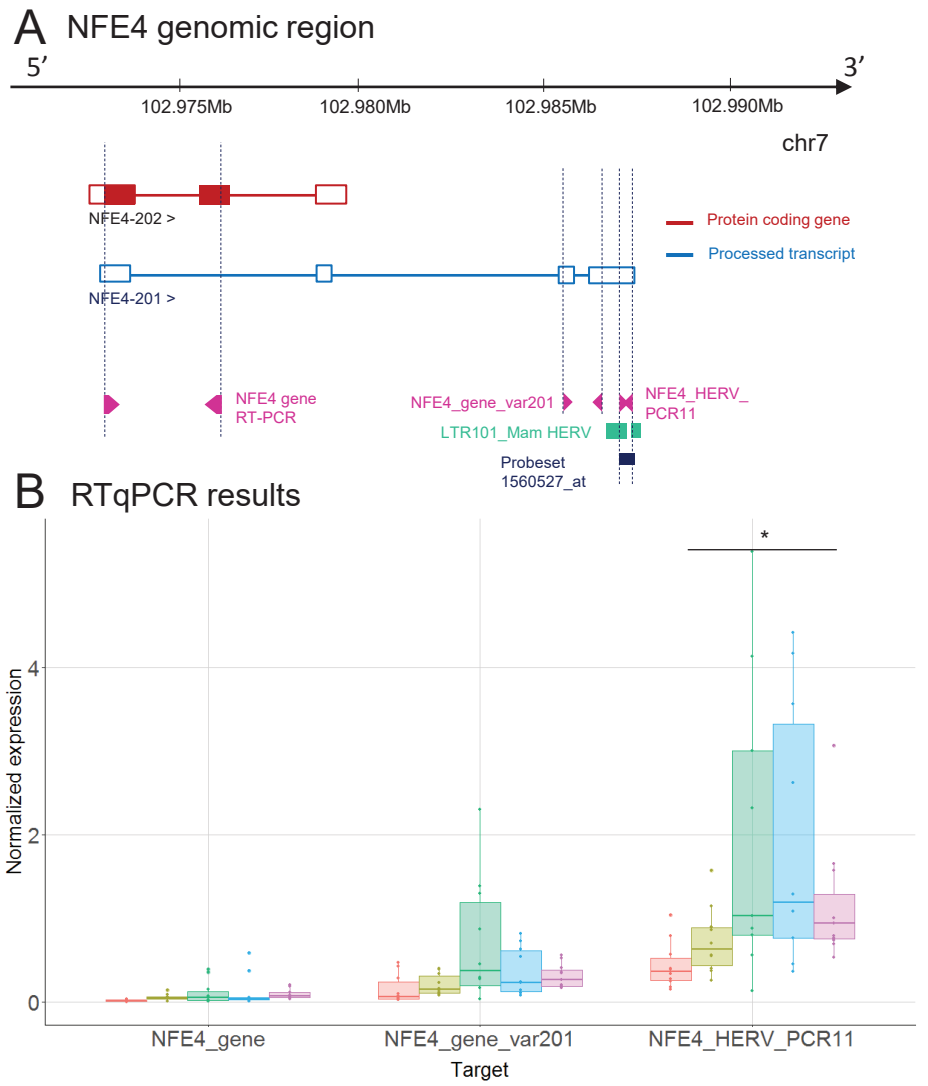
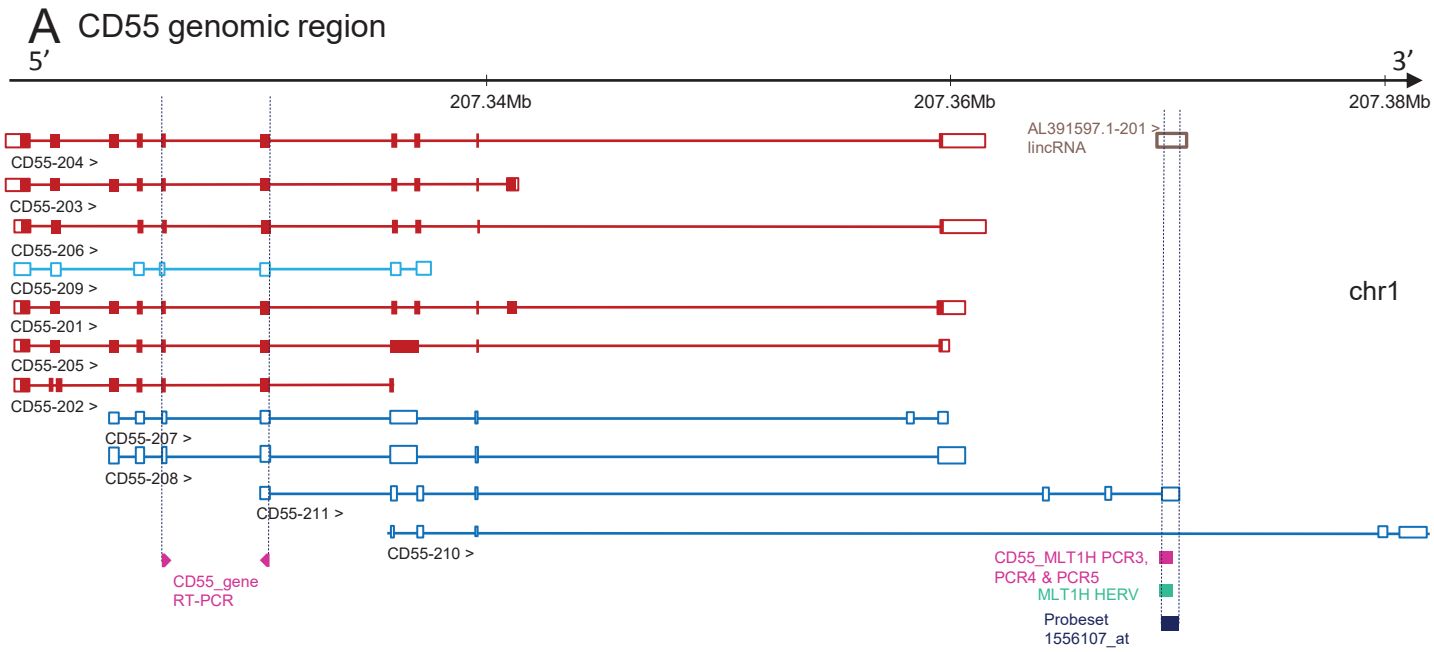
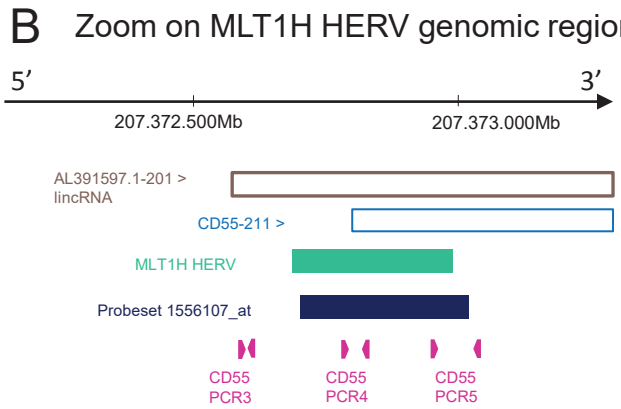
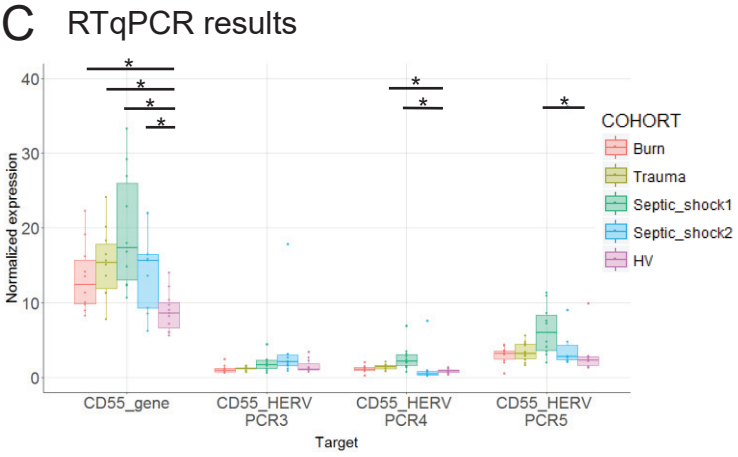
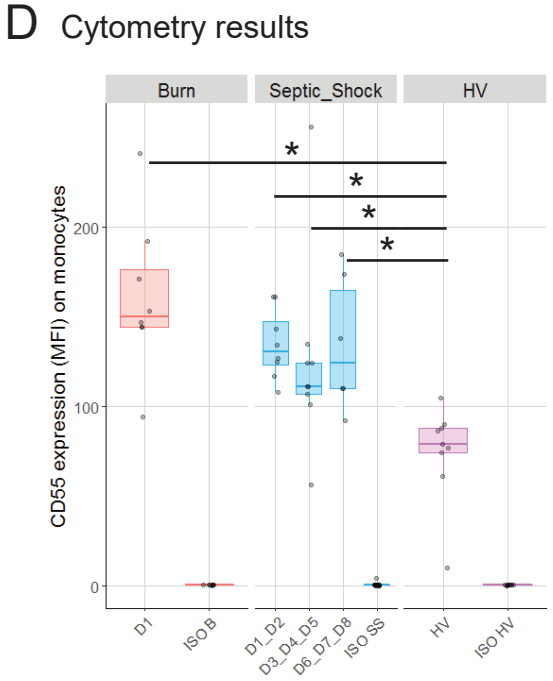

A NFE4 genomic region
5'
3'
102.975Mb
102.980Mb
102.985Mb
102.990Mb
chr7
NFE4-202 >
NFE4-201 >
Protein coding gene
Processed transcript
NFE4 gene
RT-PCR
NFE4_gene_var201
NFE4_HERV_
PCR11
LTR101_Mam HERV
Probeset
1560527_at
B RTqPCR results
*
Normalized expression
0
2
4
NFE4_gene
NFE4_gene_var201
NFE4_HERV_PCR11
Target

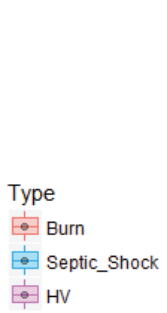
Figure 5


A CD55 genomic region
5'
3'
207.34Mb
207.36Mb
207.38Mb
AL391597.1-201 >
lincRNA
CD55-204 >
CD55-203 >
CD55-206 >
CD55-209 >
CD55-201 >
CD55-205 >
CD55-202 >
CD55-207 >
CD55-208 >
CD55-211 >
CD55-210 >
chr1
CD55_gene
RT-PCR
CD55_MLT1H PCR3,
PCR4 & PCR5
MLT1H HERV
Probeset
1556107_at


B Zoom on MLT1H HERV genomic region
5'
3'
207.372.500Mb
207.373.000Mb
AL391597.1-201 >
lincRNA
CD55-211 >
MLT1H HERV
Probeset 1556107_at
CD55
PCR3
CD55
PCR4
CD55
PCR5


C RTqPCR results
Normalized expression
40
30
20
10
0
CD55_gene
CD55_HERV
PCR3
CD55_HERV
PCR4
CD55_HERV
PCR5
Target
COHORT
Burn
Trauma
Septic_shock1
Septic_shock2
HV


D Cytometry results
Burn
Septic_Shock
HV
CD55 expression (MFI) on monocytes
200
100
0
D1
ISO B
D1_D2
D3_D4_D5
D6_D7_D8
ISO SS
HV
ISO HV


Type
Burn
Septic_Shock
HV

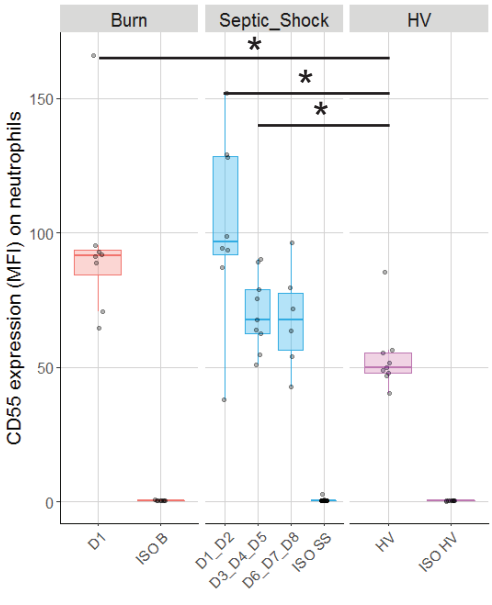
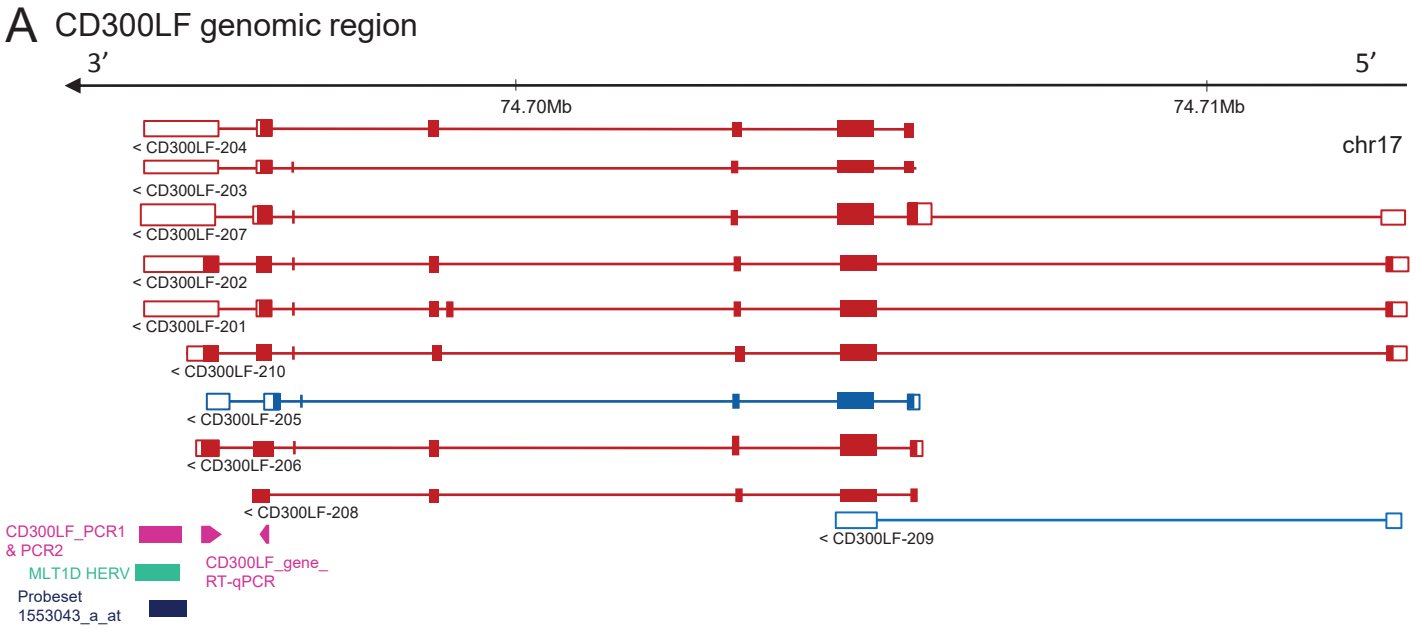
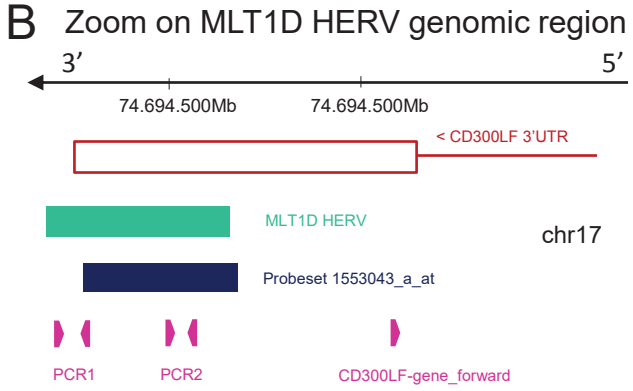
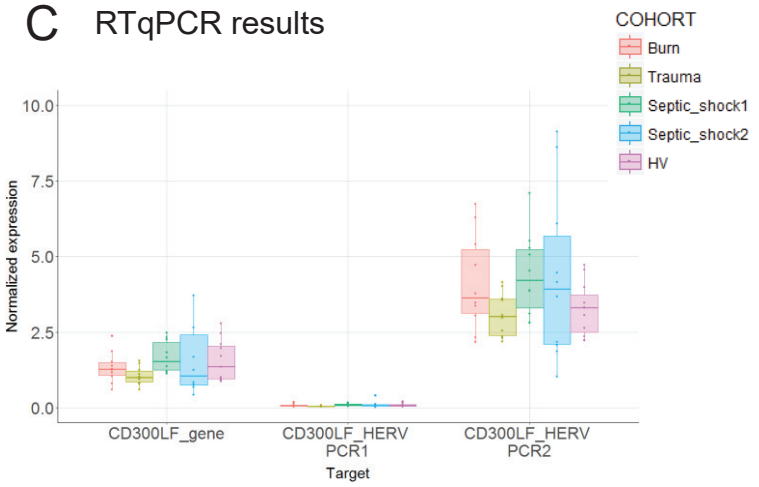
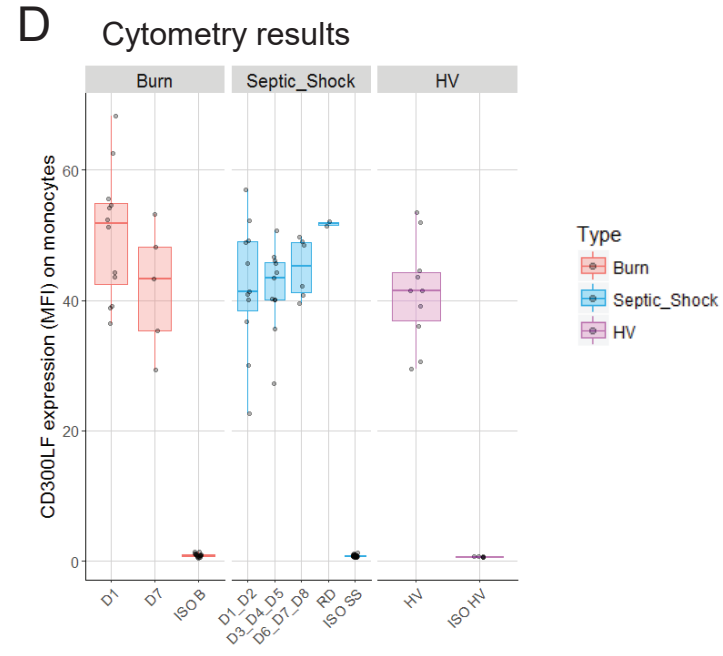

Burn
Septic_Shock
HV
CD55 expression (MFI) on neutrophils
150
100
50
0
D1
ISO B
D1_D2
D3_D4_D5
D6_D7_D8
ISO SS
HV
ISO HV

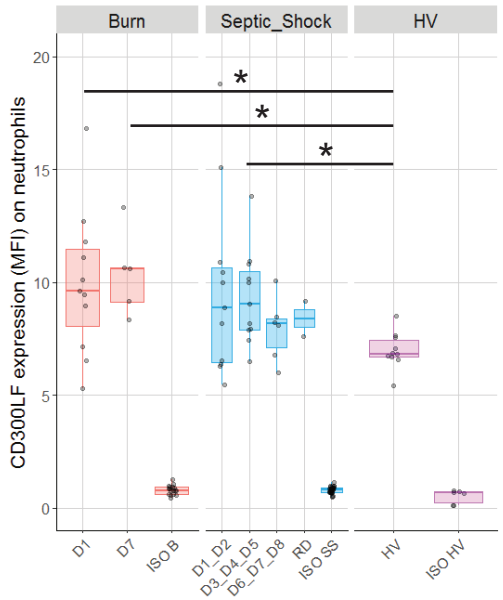
Figure 6


A CD300LF genomic region
3'
5'
74.70Mb
74.71Mb
chr17
< CD300LF-204
< CD300LF-203
< CD300LF-207
< CD300LF-202
< CD300LF-201
< CD300LF-210
< CD300LF-205
< CD300LF-206
< CD300LF-208
< CD300LF-209
CD300LF_PCR1 & PCR2
MLT1D HERV
Probeset 1553043_a_at
CD300LF_gene_ RT-qPCR

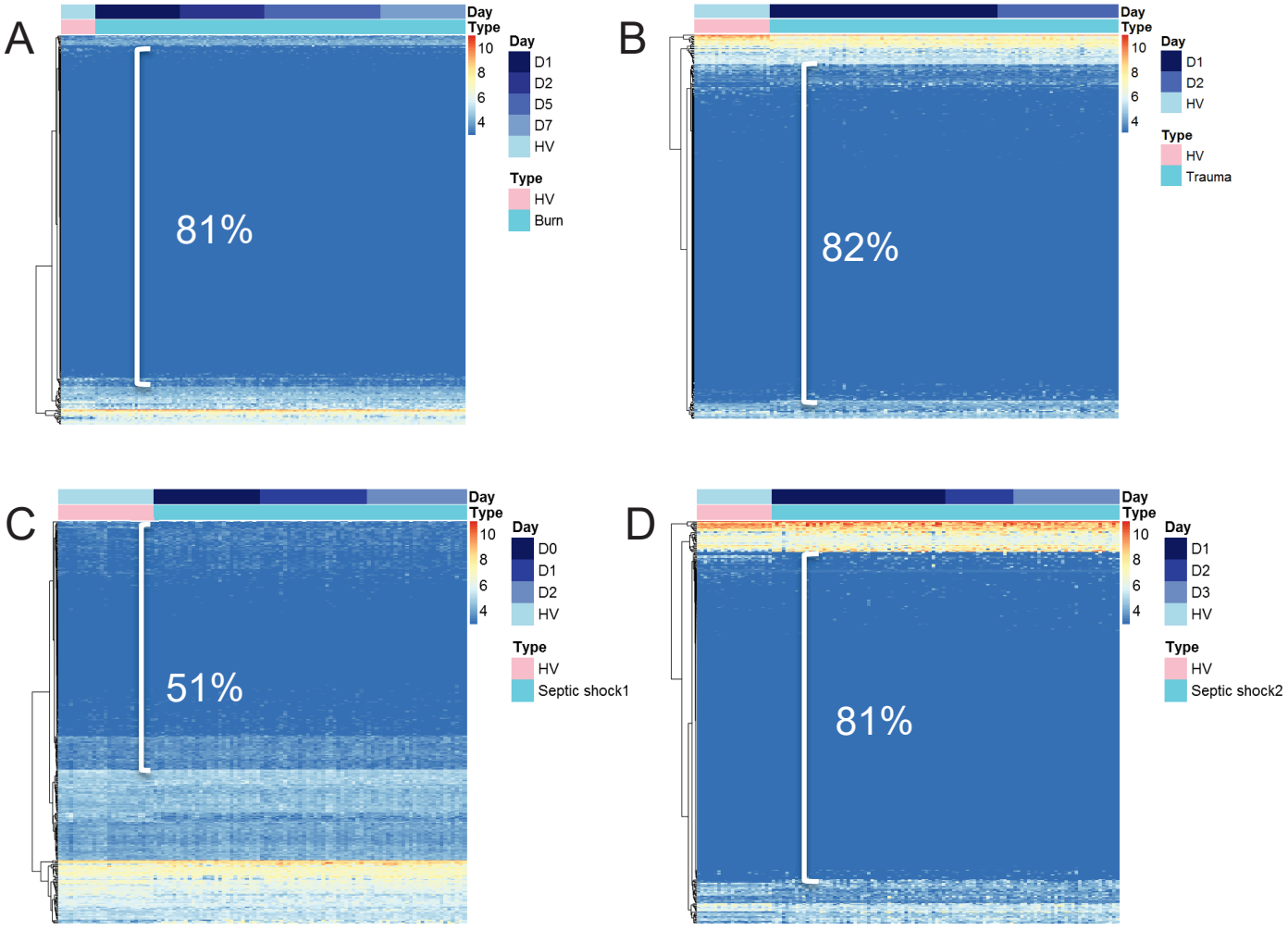

B Zoom on MLT1D HERV genomic region
3'
5'
74.694.500Mb
74.694.500Mb
< CD300LF 3'UTR
MLT1D HERV
chr17
Probeset 1553043_a_at
PCR1
PCR2
CD300LF-gene_forward


C RTqPCR results
COHORT
Burn
Trauma
Septic_shock1
Septic_shock2
HV
Normalized expression
10.0
7.5
5.0
2.5
0.0
CD300LF_gene
CD300LF_HERV PCR1
CD300LF_HERV PCR2
Target

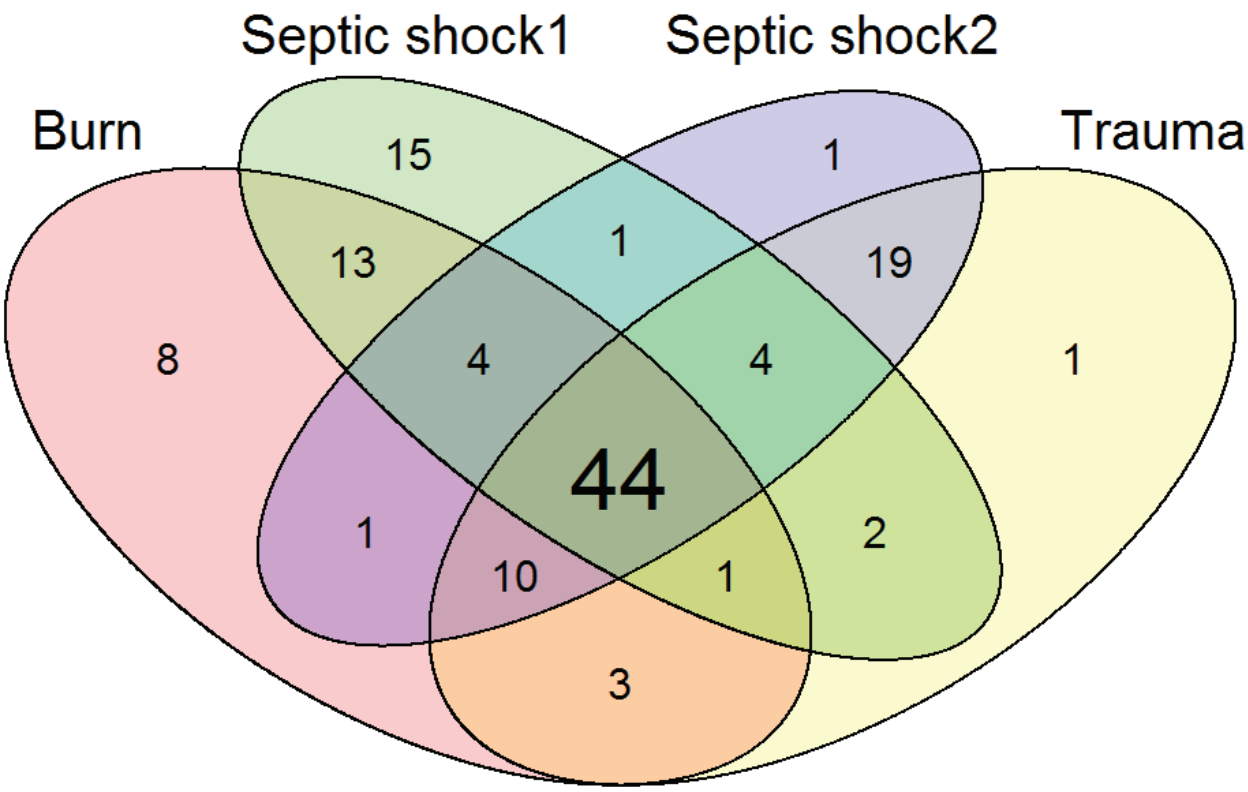

D Cytometry results
Burn
Septic_Shock
HV
CD300LF expression (MFI) on monocytes
60
40
20
0
D1
D7
ISO B
D1_D2
D3_D4_D5
D6_D7_D8
RD
ISO SS
HV
ISO HV
Type
Burn
Septic_Shock
HV


Burn
Septic_Shock
HV
CD300LF expression (MFI) on neutrophils
20
15
10
5
0
*
*
*
D1
D7
ISO B
D1_D2
D3_D4_D5
D6_D7_D8
RD
ISO SS
HV
ISO HV

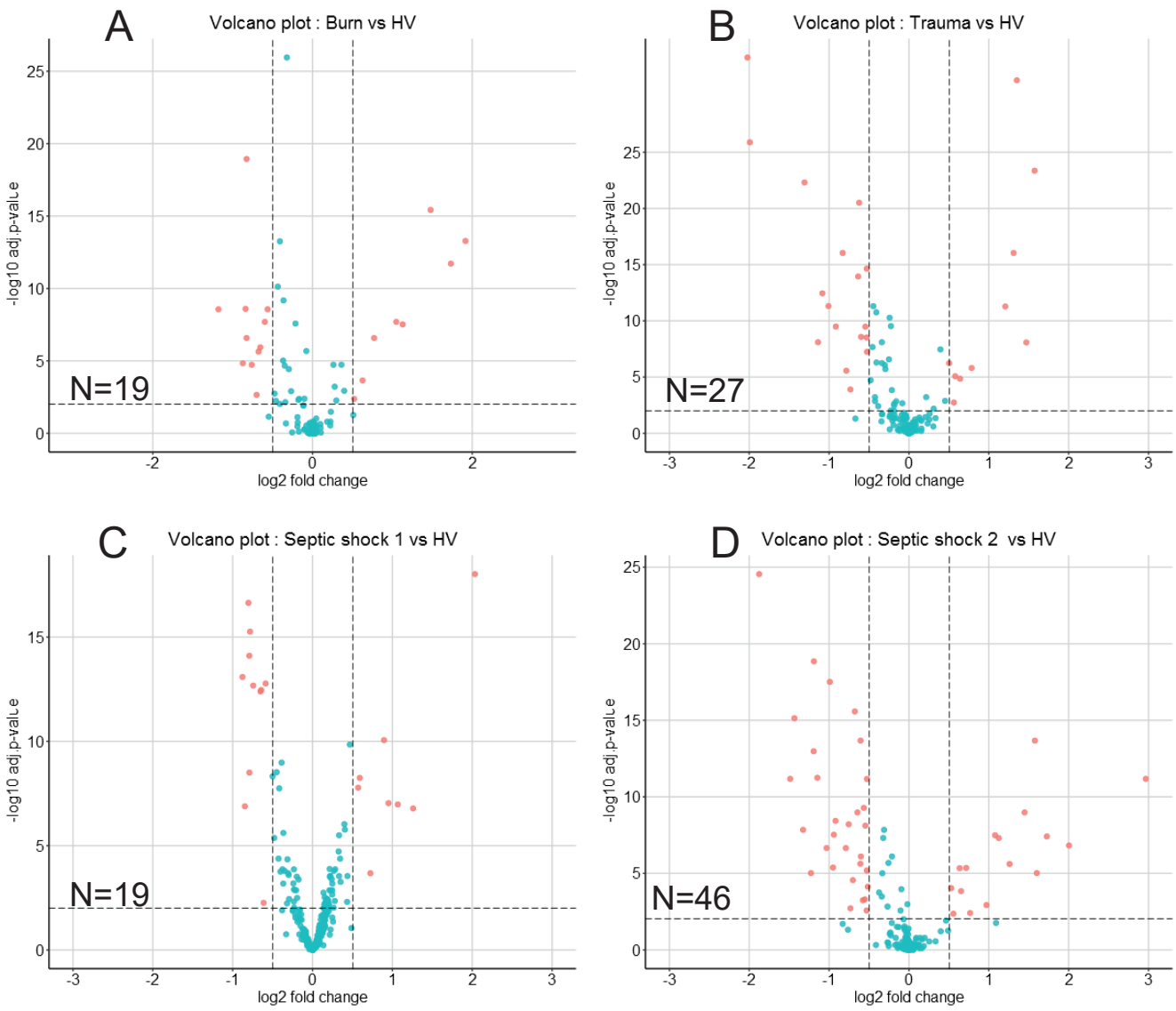
Supplemental figure 1



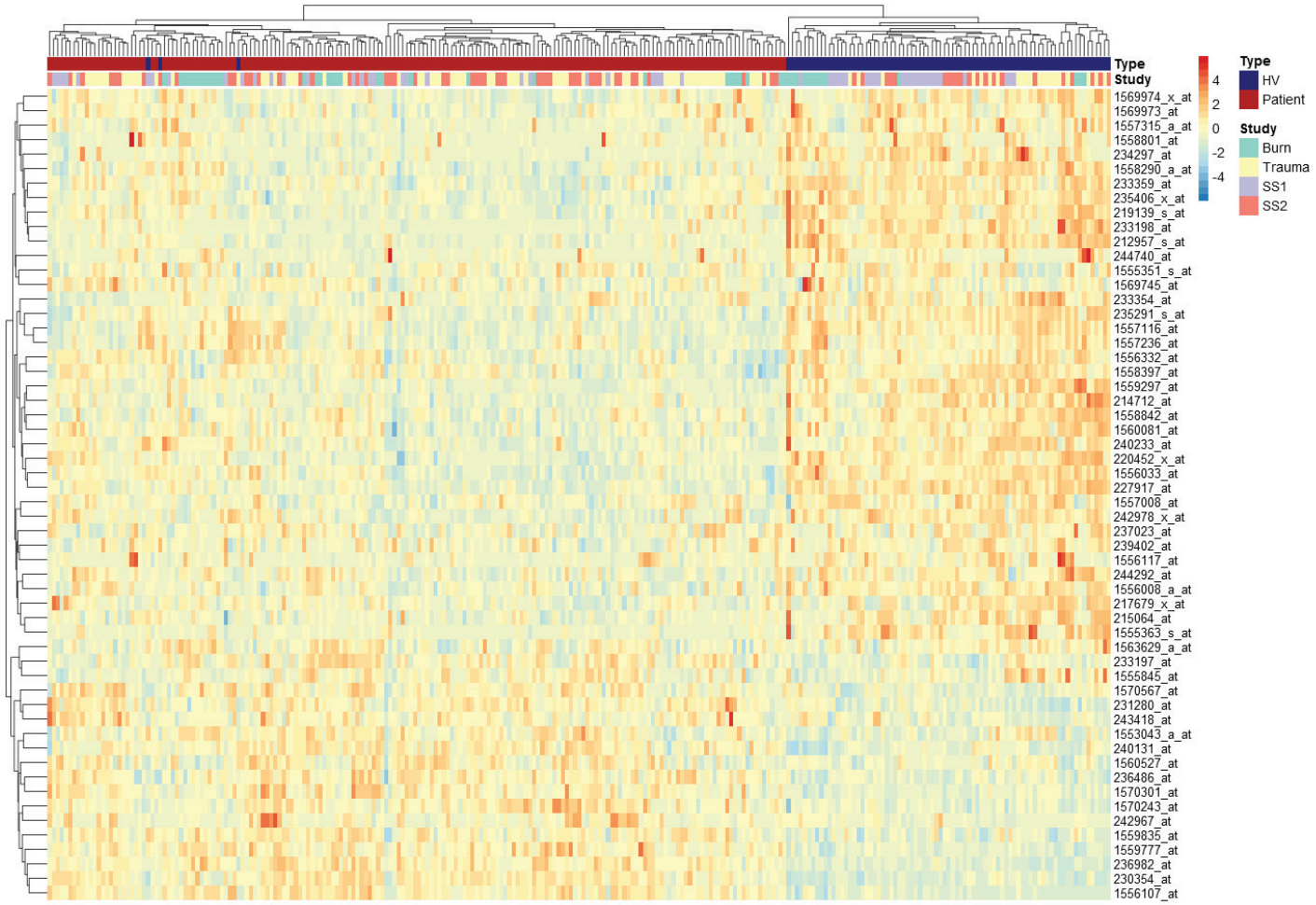
Supplemental figure 2



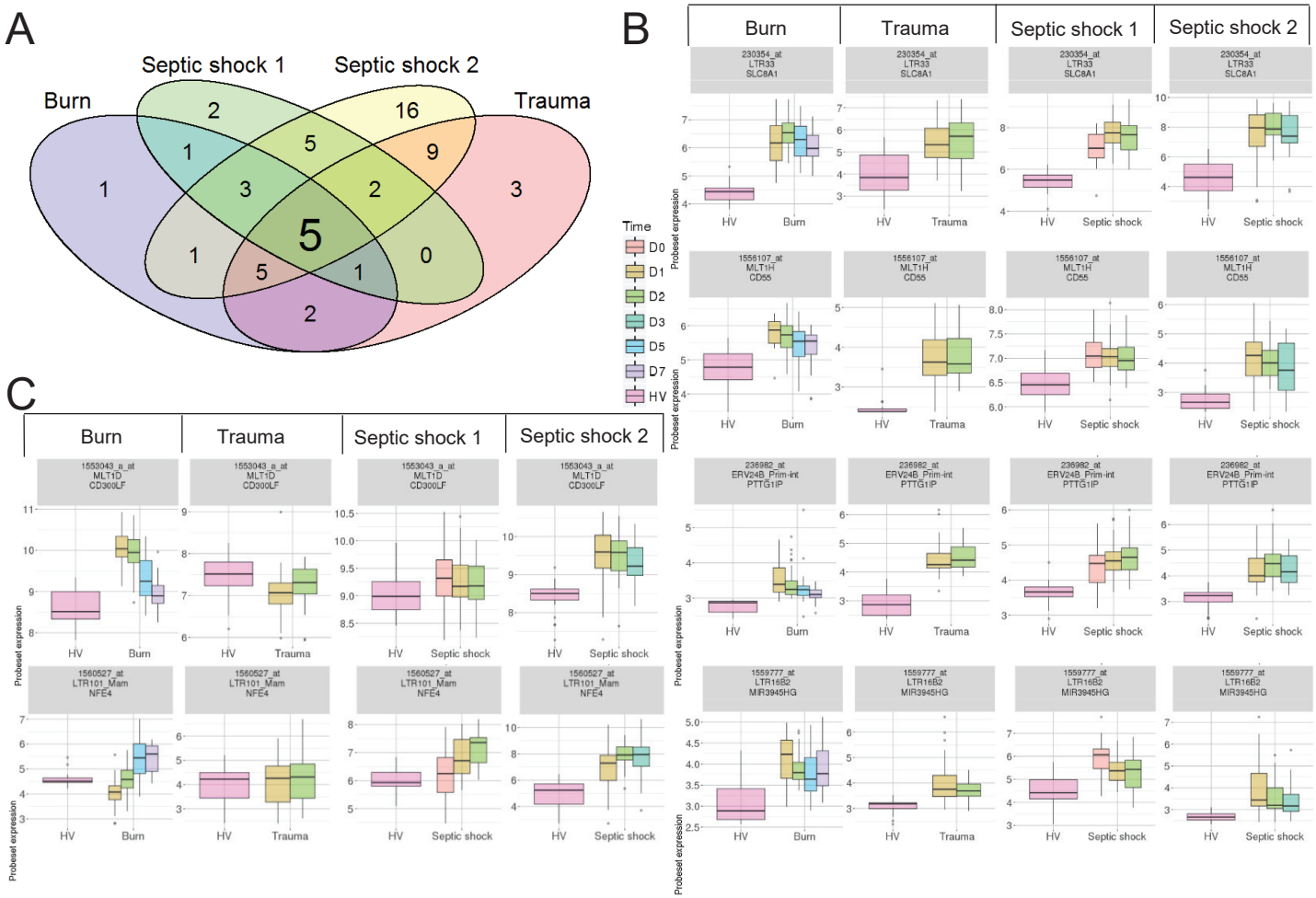
Supplemental figure 3



Supplemental figure 4



Supplemental figure 5



3.2.2 EXPRESSION DANS UN MODELE DE TOLERANCE A L'ENDOTOXINE

3.2.2.1 RESUME DE L'ETUDE

L'étude précédente a montré qu'une modulation HERV existe dans les cellules du sang, dans ces contextes inflammatoires aigus. Comme présenté dans l'introduction, le modèle de tolérance à l'endotoxine (modèle ET) reproduit certaines caractéristiques de la réponse de l'hôte suite à un choc septique. Sur des PBMCs de volontaires sains, une seule stimulation forte de LPS va mimer la réponse inflammatoire des monocytes ou état d'inflammation (condition LPS). Une faible dose de LPS précédant une dose forte de LPS va mimer l'anergie monocytaire retrouvée chez les patients en choc septique, ou état d'immunodépression (condition ET , (Allantaz-Frager et al., 2013; Cavaillon and Adib-Conquy, 2006)). Au total, le jeu de données groupe des PBMCs de 5 volontaires sains, 3 différentes conditions et des triplicats pour chaque condition, ce qui donne 45 échantillons. Avec la puce HERV-V3, permettant de cibler quasi exhaustivement les HERV du génome, nous avons regardé leur expression et leur modulation dans ce modèle *ex vivo*.

Cette étude a pour but de décrire pour la première fois le transcriptome HERV dans les PBMCs de volontaires sains ainsi que leur modulation entre les conditions du modèle. Elle permet également de suggérer un rôle des HERV sur la réponse de l'hôte, d'une part en intégrant les HERV dans les réseaux fonctionnels de gènes, et d'autre part en assignant des rôles putatifs aux LTR, sur lesquels nous reviendrons par la suite.

3.2.2.2 DESCRIPTION PRELIMINAIRE

Cette section présente un exemple d'analyses préliminaires, réalisées pour chacun des jeux de données obtenus avec la puce HERV-V3. L'exemple est réalisé sur le modèle ET.

La distribution des données d'expression par répertoire montre 2 types d'expression bien distincts (Figure 3-7). D'un côté, quel que soit le répertoire ciblant les 1500 gènes (U133, HTA ou Opti), les distributions sont bimodales, avec un pic d'expression à environ 2 d'intensité, et un autre à environ 5 d'intensité. D'un autre côté, les répertoires HERV et MaLR montrent une distribution plutôt uni modale, avec un pic d'intensité à 2. Cependant on observe aussi pour les répertoire HERV de l'expression à une intensité de 5 ou plus. Ces

différentes formes entre gènes et HERV s'explique par le fait que la majorité des probesets ciblant les HERV ne sont pas exprimés, alors que ce n'est pas le cas pour les gènes. Ce Grand nombre d'éléments à faible niveau d'expression (< 5) valide le fait de filtrer le jeu de données afin d'enlever le bruit de fond et gagner en puissance statistique. Après filtre sur intensité des sondes (Figure 3-5, Figure 3-6 et méthodes du papier section 3.2.2.4). Ce type d'analyses préliminaires a été réalisé pour chaque jeu de données analysé par la puce HERV-V3, et les mêmes tendance ont à chaque fois été observées.

Comparé au jeu de données complet, le jeu de données filtré contient nettement plus de probesets ciblant des gènes (35,7 % vs. 5,7 %), et moins de probeset ciblant les répertoires Dfam (54,7 % vs. 86,2%, Figure 3-8). Ceci est expliqué par le grand nombre de probesets HERV et MaLR non exprimés, visible sur la Figure 3-7. On voit également que proportionnellement, les probeset du répertoire prototype ont été moins filtrés que les probesets des répertoires Dfam. L'analyse en composante principale sur les données filtrées révèle que la variabilité expliquée par la composante 2 est majoritairement due à la distinction entre condition NS et conditions LPS/ET (Figure 3-9). La composante 1, dans une moindre mesure, distingue la condition LPS des conditions NS/ET. Cela suggère globalement une forte variabilité entre les conditions stimulées et non stimulées, mais également qu'il existe une modulation entre les conditions LPS et ET. L'écart entre les différents points signe également une hétérogénéité de réponse selon les volontaires.

La suite de l'étude se trouve dans l'article de la section 3.2.2.4.

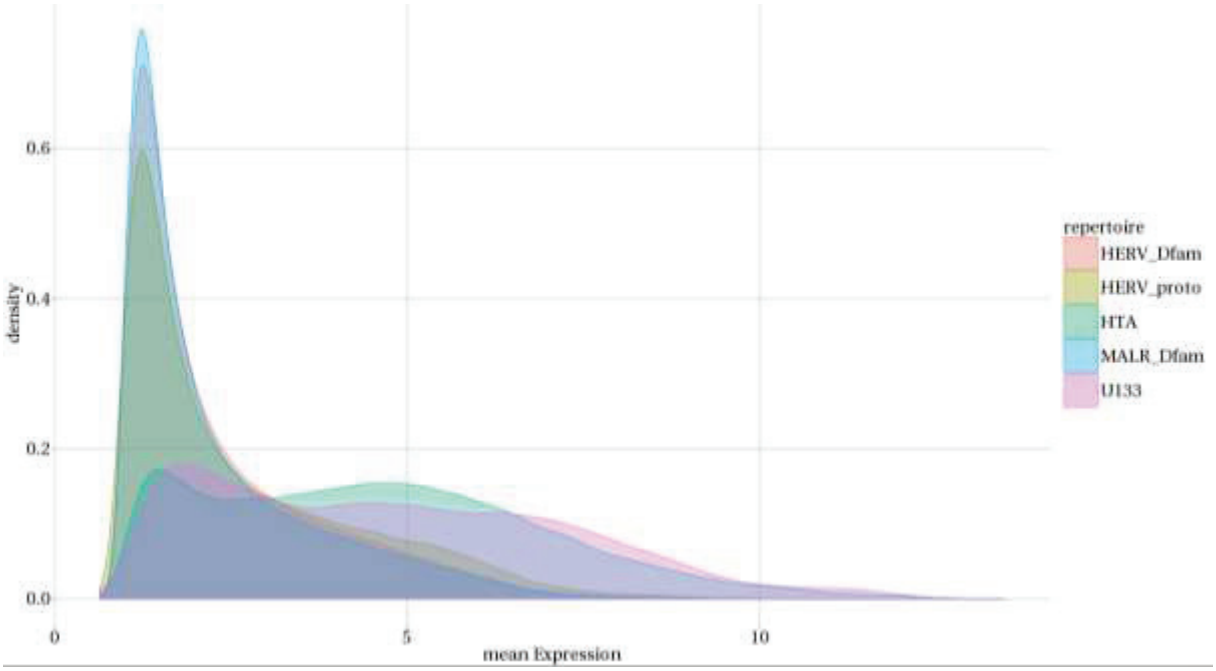


Figure 3-7 : Distribution des intensités par répertoire. L'axe des x représente l'expression moyenne pour chaque probeset. Chaque courbe de densité est colorée en fonction du répertoire.

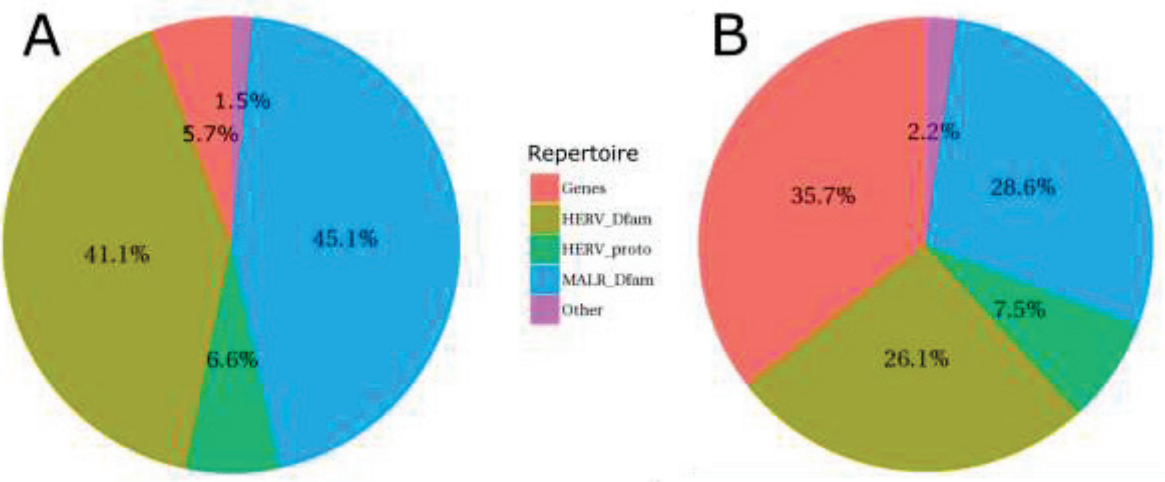


Figure 3-8 : Proportions relatives des principaux répertoires dans les jeux de données. A. Jeu de données complet. B. Jeu de données filtré.

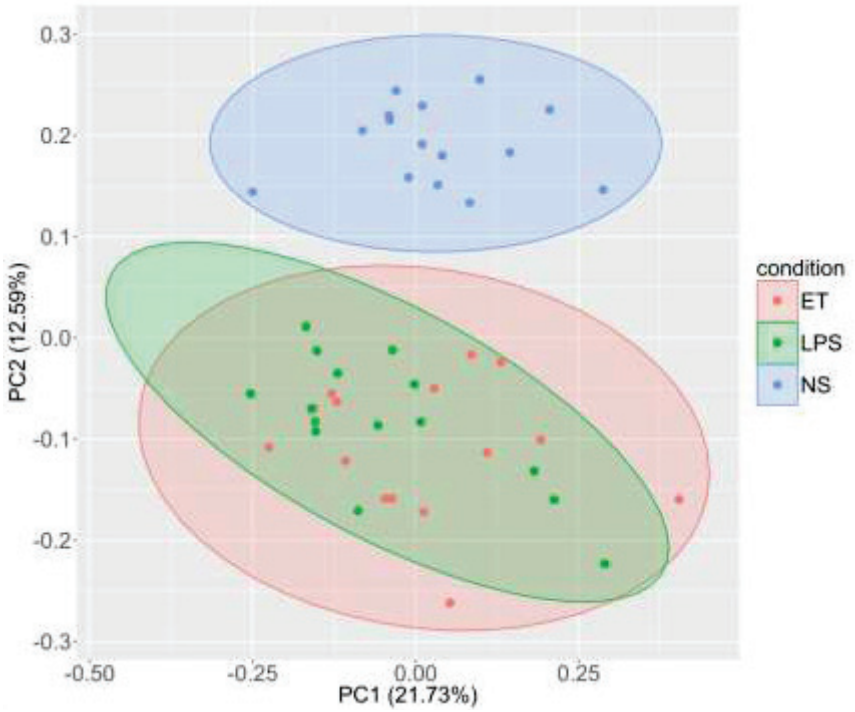


Figure 3-9 : Analyse en composante principale. Chaque point représente un échantillon, et est coloré en fonction de la condition NS (bleu), LPS (vert) ou ET (rouge). La composante 1 explique 21,7% de la variabilité totale, la composante 2 en explique 12,6%.

3.2.2.3 CONCLUSIONS

Nous montrons dans ce travail qu'environ 5% des HERV et MaLR sont exprimés dans les PBMCs. Parmi ces éléments, nous mettons en évidence que le groupe HERV-H est parmi le plus abondant, comportant le plus d'éléments exprimés (17% des HERV-H). Le récent groupe HML-2, considéré comme le plus actif, a quant à lui, relativement peu d'éléments exprimés (9% du groupe). Nous décrivons également la modulation des HERV entre les conditions non stimulée (NS), d'inflammation (LPS) et d'immunodépression (ET). De manière intéressante, les proportions de HERV modulés entre les conditions NS et LPS sont similaires, alors que 86% des HERV modulés entre ET et LPS sont down-modulés (42 vs 5). Un pourcentage similaire de gènes down-modulés est observé (86%). De plus, nous mettons en évidence un grand nombre de HERV qui ont un profil d'expression similaire avec les gènes ciblés par la puce HERV-V3, que ce soit un profil tolérisable (plus fortement exprimé dans la condition LPS par rapport aux autres conditions), ou non tolérisable (plus fortement exprimé dans les conditions LPS et ET par rapport à NS). A partir de cette co-expression, nous concluons à une régulation commune de l'expression entre des gènes de l'immunité et 64 loci HERV, nous

permettant d'intégrer les HERV dans les réseaux de régulation de gènes de l'immunité. Nous identifions, entre autre, des HERV, co-exprimés avec et proches de gènes de la réponse interféron, OAS3 et IFI44L, importants dans la reconnaissance de pathogènes et plus particulièrement des virus. De manière intéressante, un groupe de HERV (MER41) a déjà été identifié comme pouvant avoir un rôle de site de fixation de facteur de transcription des gènes de la voie interféron (gène AIM2 notamment, introduction, section 1.2.2). Nous introduisons également la notion de fonctions possibles des LTR sur l'expression de gènes. Par une décomposition du signal observé sur les LTR, nous sommes en mesure d'attribuer une fonction putative aux LTR. Nous montrons ainsi qu'un dixième des LTR peuvent passer d'un état inactif à un état potentiellement actif dans le modèle ET. Je reviendrai plus en détail sur les fonctions des LTR dans la section 3.4.3.

Cette étude nous permet de décrire pour la première fois, avec la puce HERV-V3, le transcriptome HERV, sa modulation dans les PBMCs, stimulés ou non et la co-expression de nombreux HERV avec des gènes de la réponse immunitaire. Cependant, elle comporte certaines limitations. L'utilisation d'un modèle ex-vivo ne peut reproduire à l'identique ce qui peut se passer dans un organisme entier. De plus, la stimulation au LPS ne mime pas l'ensemble des pathogènes induisant une réponse immunitaire lors d'un sepsis.

3.2.2.4 ARTICLE

LTR-RETROTRANSPOSON TRANSCRIPTOME MODULATION IN RESPONSE TO ENDOTOXIN-INDUCED STRESS IN PBMCS

Marine Mommert*, **Olivier Tabone***, Guy Oriol, Elisabeth Cerrato, Audrey Guichard, Magali Naville, Paola Fournier, Jean-Nicolas Volff, Alexandre Pachot, Guillaume Monneret, Fabienne Venet, Karen Brengel-Pesce, Julien Textoris and François Mallet

* Contribution à parts égales

Année 2018

Publié dans *BMC Genomics (2018) 19 :522*

Mommert et al. BMC Genomics (2018) 19:522
https://doi.org/10.1186/s12864-018-4901-9

BMC Genomics

RESEARCH ARTICLE Open Access

LTR-retrotransposon transcriptome modulation in response to endotoxin-induced stress in PBMCs

Marine Mommert[1,2*†], Olivier Tabone[2†], Guy Oriol[1], Elisabeth Cerrato[2], Audrey Guichard[1,2], Magali Naville[3], Paola Fournier[1], Jean-Nicolas Volff[3], Alexandre Pachot[2], Guillaume Monneret[2,4], Fabienne Venet[2,4], Karen Brengel-Pesce[1], Julien Textoris[2,5] and François Mallet[1,2*†]

Abstract

Background: Human Endogenous Retroviruses (HERVs) and Mammalian apparent LTR-retrotransposons (MaLRs) represent the 8% of our genome and are distributed among our 46 chromosomes. These LTR-retrotransposons are thought to be essentially silent except in cancer, autoimmunity and placental development. Their Long Terminal Repeats (LTRs) constitute putative promoter or polyA regulatory sequences. In this study, we used a recently described high-density microarray which can be used to study HERV/MaLR transcriptome including 353,994 HERV/MaLR loci and 1559 immunity-related genes.

Results: We described, for the first time, the HERV transcriptome in peripheral blood mononuclear cells (PBMCs) using a cellular model mimicking inflammatory response and monocyte anergy observed after septic shock. About 5.6% of the HERV/MaLR repertoire is transcribed in PBMCs. Roughly one-tenth [5.7–13.1%] of LTRs exhibit a putative constitutive promoter or polyA function while one-quarter [19.5–27.6%] may shift from silent to active. Evidence was given that some HERVs/MaLRs and genes may share similar regulation control under lipopolysaccharide (LPS) stimulation conditions. Stimulus-dependent response confirms that HERV expression is tightly regulated in PBMCs. Altogether, these observations make it possible to integrate 62 HERVs/MaLRs and 26 genes in 11 canonical pathways and suggest a link between HERV expression and immune response. The transcriptional modulation of HERVs located close to genes such as OAS2/3 and IFI44/IFI44L or at a great distance from genes was discussed.

Conclusion: This microarray-based approach revealed the expression of about 47,466 distinct HERV loci and identified 951 putative promoter LTRs and 744 putative polyA LTRs in PBMCs. HERV/MaLR expression was shown to be tightly modulated under several stimuli including high-dose and low-dose LPS and Interferon-γ (IFN-γ). HERV incorporation at the crossroads of immune response pathways paves the way for further functional studies and analyses of the HERV transcriptome in altered immune responses in vivo such as in sepsis.

Keywords: HERV transcriptome, PBMCs, LPS, Endotoxin tolerance, Signalling pathways, Sepsis

* Correspondence: marine.mommert@biomerieux.com; francois.mallet@biomerieux.com
†Marine Mommert, Olivier Tabone and François Mallet contributed equally to this work.
[1]Joint research unit, Hospice Civils de Lyon, bioMerieux, Centre Hospitalier Lyon Sud, 165 Chemin du Grand Revoyet, 69310 Pierre-Benite, France
Full list of author information is available at the end of the article

Background

Retrovirus-like sequences represent the 8% of the human genome [1]. They consist of some 200,000 Human Endogenous Retroviruses (HERVs) and 240,000 Mammalian Apparent LTR retrotransposons (MaLRs). HERVs are remnants of ancestral and independent retroviral infections within the germ line. The parental integrated retroviral DNA or provirus is flanked by two Long Terminal Repeats (LTR). The 5'LTR contains the promoter and enhancer signals initiating transcription, while the 3'LTR contains the polyadenylation signal terminating transcription. Between the two LTRs lie at least three genes coding for the structural proteins (*gag*), the enzymatic proteins (*pro-pol*), and the envelope glycoprotein (*env*). MaLR structure is similar except for the absence of an *env* gene. Since endogenization, proviruses have propagated all along the genome by reinfection and retrotransposition events. Due to the general absence of selection pressure, most of the elements contain substitutions, insertions and deletions. However, a few hundred large open reading frames (ORF) remain [2], including *env* ORFs, of which Syncytins support essential functions in placental development [3]. In addition, numerous HERV elements consist of solitary LTRs, resulting from the loss of coding genes by recombination between the two flanking LTRs. All these mechanisms lead to complex multicopy groups, (reviewed in [4–6]). Each group consists of heterogeneous elements, all defective for replication, and thus engaged in a vertical mode of transmission exclusively. While bioinformatics approaches have identified 103 HERV groups and 1 MaLR group [1], only 40 HERV groups have been characterised in wet-lab studies [7–9]. Based on the homology of *pol* sequences between HERVs and exogenous retroviruses, these well-defined groups can be classified as gamma-, beta-, spuma- and epsilon- retroviruses [10].

As repeated elements and due to their organisation into groups, (H)ERVs may be involved in genomic plasticity during evolution, being preferential recombination sites within or between chromosomes [11]. Under physiological conditions, HERV elements are subject to strong epigenetic controls either in terms of methylation or histone code [12]. Nevertheless, in various diverse situations, HERV elements have been shown to be transcribed. HERV transcription has been observed in organ-specific (e.g. brain for multiple sclerosis) and systemic (e.g. lupus erythematosus) autoimmune diseases, and HERV-driven mechanisms involving molecular mimicry and immune dysregulation have been proposed [13–15]. HERV expression has also been researched in cancer with regard to the oncogenic properties of infectious retroviruses and epigenetic changes observed in cancer [12, 16]. The latter highlighted LTR-driven transactivation of cellular proto-oncogenes or the expression of HERV-encoded *Env,* NP9 or Rec candidate oncogenes. HERVs also contribute to the physiopathology of their host at multiple levels. In brief, (i) solo or proviral LTR can modulate the expression of adjacent cellular genes, in addition to their autonomous function in controlling retroviral expression [17–20], (ii) the expression of HERV proteins with conventional retroviral functions can influence the host's physiological or pathological states, like fusion for Syncytin-1 [21], immunomodulation for Env HERV-H and Syncytin-2 [22, 23], RNA nuclear export for Rec [24], and even viral-like particle formation derived from HML-2 [25], and finally (iii) non-coding HERV-expressed sequences may also be biologically active, e.g. HERV-H loci involved in the maintenance of pluripotency in human cells [26]. In all cases, deciphering an HERV biological function starts with the analysis of its expression, which is far from being simple due to both complex biological mechanisms and technical challenges. Such biological complexity is illustrated by one extensively described HERV-W element, the ERVWE1 locus, encoding placental fusogenic Syncytin-1 Env. ERVWE1 expression is driven by its own 5'LTR U3 promoter and adjacent MaLR LTR enhancer. Restricted expression outside the trophoblast is controlled at the LTRslevel by CpG methylation and/or repressive histone mark H3K9me3, and at the splicing level, at least in part, by H3K36me3 along the intron–exon boundary (reviewed in [3]). Experimentally, the challenge of the individual identification of transcriptionally active HERV loci was recently addressed using NGS [27] and high-density microarray [28–30] technologies. Indeed, first [28], second [29] and third [30] generations of custom HERV-dedicated microarrays aimed to solve the antagonism between the specificity of individual locus recognition and exhaustiveness of the HERVome. Although addressing a limited number of groups, the first two generations of HERV-dedicated microarrays confirmed that reproductive organs and solid tumours are major sites of HERV expression, and highlighted the tissue specificity/tropism of expressed HERV elements [28, 29].

Most of the literature concerning HERVs focuses on autoimmune diseases, cancers and placental physiopathology, all these contexts being associated with local or systemic modulation of the immune response. Indeed, there is a growing line of evidence that HERVs may directly shape and regulate our immune system [31–33]. The first evidence of the expression of HERVs beta-retroviruses in PBMCs was reported in healthy volunteers 20 years ago, using *pol*-based pan retroviral PCR [34], following Northern-blot-based seminal scrutiny [35]. This observation was extended to the HERV-H and HERV-W gamma-retrovirus groups using similar PCR-based technology 20 years later [36]. It was demonstrated that the level of HERV expression in the PBMC compartment is modulated in solid organ cancers, autoimmune

diseases, infectious diseases, as well as immunocompromised states, e.g. HML-2 in prostate cancer [37], HERV-W/MSRV in multiple sclerosis [38], HERV-W in EBV-infected multiple sclerosis patients [39], and HML-2 in HIV infected patients [40]. In line with this, HERV expression in PBMCs is modulated by microbial components, the differentiation state of the cell and cytokines. Bacteria-derived components such as LPS increase HERV-W, HERV-K, HERV-H, and decrease HERV-E group expression in monocyte-derived macrophage (MDM) cell lines [41]. Recently, taking advantage of probes targeting HERVs in commercial microarrays, the immune cell activation by microbial signals in vitro induces global modulation of endogenous retroelements [42]. HERV-W, HERV-K and HERV-H RNA levels are increased during monocyte differentiation [41]. HERV expression is modulated by cytokines, as observed for ERV3 in the U-937 monocytic cell line [43], or also for MSRV released from PBMCs stimulated by TNF-α and IFN-γ and inhibited by IFN-α [44]. Moreover, Env of HERV-W and HERV-H groups are expressed on the surface of B cells and monocytes in patients with active multiple sclerosis [45]. Syncytin-1 has also been observed in other pro-inflammatory states in skin-homing non-recirculating mycosis fungoides T cells [46]. Conversely, immunosuppressive properties of Syncytin-1 [47, 48] and Env HERV-H [22] have been described in different contexts [49]. These data support the hypothesis that HERVs are expressed in inflammatory and immunosuppressive contexts and have direct or indirect interactions with the immune host response.

Thus, the considerable improvement inthe molecular and cellular biology tools helps demonstrate that many HERVs are not silent in many contexts. The use ofunique locus candidate approaches helps understand their role in pathophysiological phenomena. We recently introduced a third generation of HERV-dedicated microarray, in high-density Affymetrix format, which makes it possible to measure HERVs at the individual locus level. It targets almost complete coverage of HERVs and MaLRs LTR-retrotransposons of the GRCh38 version of the human genome and Dfam 1.1 database of repetitive DNA element sequence alignments. The chip also targets more than 1500 human genes mainly coding for immunity-related proteins [30]. We investigated HERVs and MaLRs expression in PBMCs by modelling monocyte endotoxin tolerance, consisting in low-dose LPS priming of PBMCs and mimicking monocyte anergy observed in sepsis patients [50–52]. Symmetrically, PBMCs were stimulated by single high-dose LPS mimicking gram-negative bacterial infection and illustrating an inflammatory context. This article i) provides an overview of the HERV transcriptome in PBMCs associated with functional LTR characterisation; ii) shows the stimulus-dependent co-modulation of HERVs/MaLRs and genes, including tolerisable reversible phenotypes; and iii) demonstrates an integrative view of HERVs/MaLRs and genes in canonical immunity pathways, illustrating the multifaceted nature of interactions between LTR retrotransposons and genes.

Results

Detection of the HERV transcriptome in PBMCs

In order to present an overview of the HERV/MaLR transcriptome in PBMCs, we used various LPS challenges to mimic healthy, inflammatory and immunocompromised states using the previously described endotoxin tolerance model [50]. The transcriptome was scrutinised using a custom Affymetrix HERV-V3 microarray [30] which can discriminate 174,852 HERV elements, 179,142 MaLR elements and a set of 1559 genes. In addition to the gene probesets designed in U133plus2 and HTA Affymetrix formats, the chip contains a set of highly informative probesets corresponding to accurately annotated HERV loci hereinafter referred to as “HERV_prototypes”, and two sets of probesets corresponding to roughly annotated HERVs/MaLRs elements called hereafter “HERV_Dfam” and “MaLR_Dfam” (summarised in Additional file 2: Table S1). In brief, the “HERV prototypes” repertoire was retrieved from selected prototype loci. These are sets of loci which maintain the largest open reading frames for *gag pol env* genes within a proviral structure flanked by two complete LTR sequences. From these elements, a repeatmasker-based alignment procedure retrieved 29,271 loci divided into 42 groups. The “HERV_Dfam” and “MaLR_Dfam” repertoires were retrieved from Dfam, a database of repetitive elements detected by RepBase consensus and based on Hidden Markov Models (HMM) [53] (“Proto versus Dfam” tab, Additional file 2: Table S1). Redundancy between repertoires was removed from the HERV/MaLR repertoire. A summary of the absolute counts and the relative abundance of transcriptionally active elements of the HERV/MaLR transcriptome is given in Fig. 1, at the probeset level. Notably, although the hybridisation assessment quality was equivalent across repeated elements and gene repertoires [30], a higher proportion of gene probesets were transcriptionally active, i.e. 52% (42,560 probesets). Overall, 5.6% of targeted HERVs/MaLRs (71,063 probesets) were transcriptionally active in PBMCs. More precisely, 9.4% of the well-described “HERV_prototypes” repertoire, and 5.5% of HERV_Dfam and MaLR_Dfam were expressed. Among the 9.4% expressed prototype elements, 6.1, 1.7 and 1.6% belonged to gamma-, beta- and spuma/epsilon-like retrovirus classes, respectively. On moving from classes towards groups, all well-defined HERVs groups had expressed loci in PBMCs. Notably, within gamma-retroviruses, the largest HERV-H group is that with the highest proportion of active probesets (940 expressed probesets, 17% of the whole group). The lesser

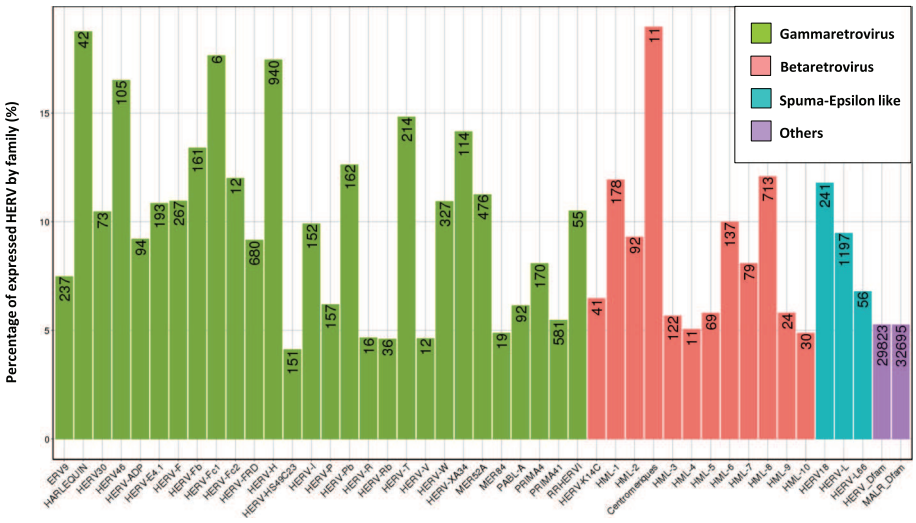


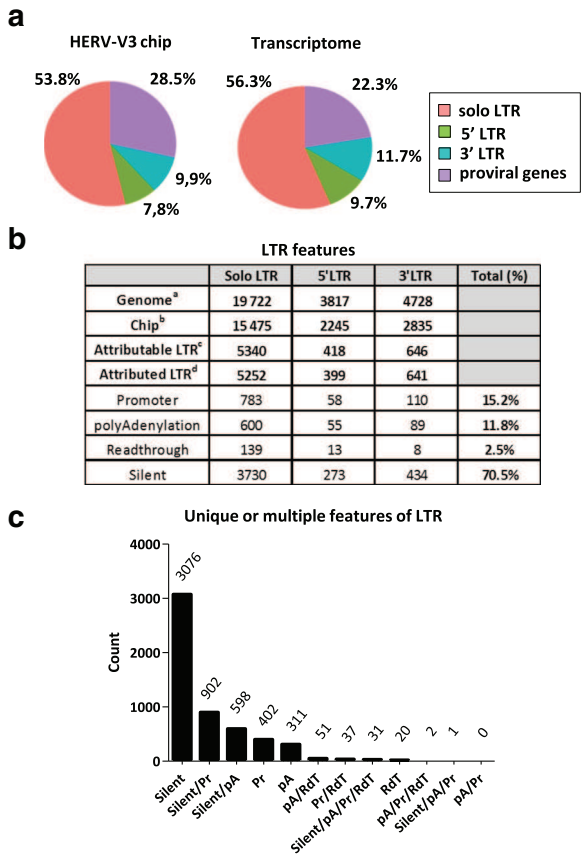
Fig. 1 The HERV transcriptome in PBMCs. Percentages and absolute counts of positive signal-associated probesets within individual groups of the "HERV_prototypes" repertoire and HERV_Dfam and MaLR_Dfam repertoires. A probeset was included as reflecting a significant transcriptional activity if its normalised intensity was over an intensity threshold of $2^{5.5}$ in at least 14 out of the 45 samples (for all conditions). This conservatory threshold was defined as the minimal intensity level shared by all repertoires that exhibited an acceptable variability, i.e. the 75th percentile of the distribution of the variation coefficient as a function of intensity should be lower than 10% (illustrated in Additional file 1: Figure S1). HERV prototype groups were grouped by retrovirus classes, namely gammaretrovirus (green), betaretrovirus (red) and spuma-epsilon like retrovirus (blue). The HERV and MaLR Dfam repertoires were each depicted as a global homogeneous entity (purple)

known PRIMA-41 group is the third most represented among gamma-retroviridae (581 probesets, 5% of the whole group). HML-8 (713 probesets) and HML-1 (178 probesets) are the groups with the highest proportion of expressed probesets (both 12%, of the whole corresponding group) among beta-retroviridae. The HML2 group, which is considered to be highly active, only showed 92 expressed probesets (9% of the whole group). Interestingly, although represented by a much smaller subgroup, centromeric HML2 elements seemed to be more expressed than the non-centromeric HML-2 ones. Finally, in spuma-retroviridae, HERV-L group provides the largest amount of active probesets (1197 probesets, 9% of the whole group). It should be noted that we did not observe significant enrichment or depletion of a specific repertoire, class or HERV group, according to healthy, inflammatory or immunocompromised-like experimental conditions (data not shown).

Functional characterisation of the HERV transcriptome in PBMCs

Beyond the basic observation of transcriptional expression, it would be informative to know whether HERVs/MaLRs exhibit any transcription bias due to their structure. The HERV/MaLR transcriptomic activity observed collectively in stimulated and unstimulated PBMCs is depicted at the LTR/*gag*/*pol*/*env* region levels for each group/repertoire (Additional file 2: Table S1). Figure 2a provides a simplified comparative view of the "HERV_prototype" repertoire addressed by the microarray and its related transcriptome. We observed some differences between chip and transcriptome. Active LTRs represent 77.7% of the HERV transcriptome, whereas they represent 71.5% of the chip. This means that LTRs are more represented in active elements than internal HERV/MaLR regions. More specifically, solo LTRs are more abundant than proviral LTRs (5'LTRs and 3'LTRs). Nevertheless, it seems that proviral LTRs are over-represented in the transcriptome (21.4% versus 17.7%), whereas solo LTRs are not (56.3% versus 53.8%). Notably, internal proviral genes are under-represented in the transcriptome (22.3% versus 28.5%). Surprisingly, if we compare the "HERV_prototypes", "HERV_Dfam" and "MaLR_Dfam" repertoire transcriptional activities, the HERV prototypes (14.3%) had higher numbers of expressed loci than Dfam elements (5.2% for HERVs and 4.4% for MaLRs). Attributable LTRs (solo, 5' or 3') are defined as LTRs bearing U3 and U5 regions. For these LTRs, we are able to attribute promoter or polyA functions. In the prototype repertoire, 6404 (31%) were attributable LTRs. According to cut-offs and fold change criteria (see material and methods), a function could be attributed to these LTRs in each of the 45 samples (Fig. 2b). Promoter (Pr) activity was assigned to 15.2% of LTRs and polyadenylation (pA) signal was observed in 11.8%. Most of the LTRs were silent (70.5%) and a minority (2.5%) were classified as readthrough (RdT).

Finally, we wanted to know whether a single LTR had the potential to change status (Fig. 2c). As expected,



	Solo LTR	5'LTR	3'LTR	Total (%)
Genome[a]	19 722	3817	4728	
Chip[b]	15 475	2245	2835	
Attributable LTR[c]	5340	418	646	
Attributed LTR[d]	5252	399	641	
Promoter	783	58	110	15.2%
polyAdenylation	600	55	89	11.8%
Readthrough	139	13	8	2.5%
Silent	3730	273	434	70.5%

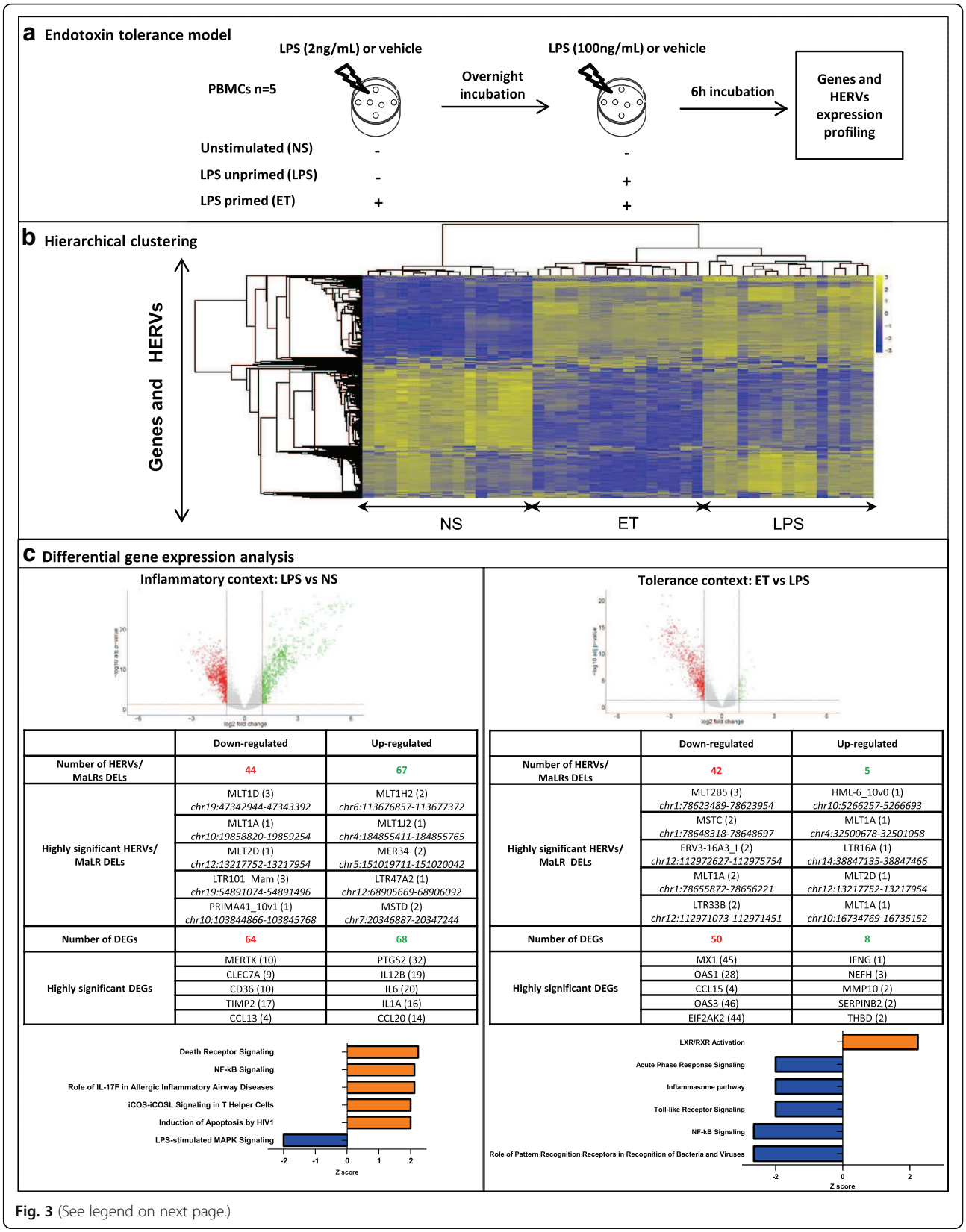
Fig. 2 Genomic, transcriptomic and functional projections of the HERV prototype repertoire. **a** HERV structure distribution. The 42 HERV groups corresponding to the "HERV_prototypes" repertoire are voluntarily depicted as 100% of the HERV-V3 chip. Solo LTR, 5′ LTR, 3′ LTR and proviral genes account for 100% of the HERV chip in the proportion described in Additional file 2: Table S1b. The transcriptome pie-chart is obtained from results detailed in Additional file 2: Table S1c. **b** LTR features. The descriptive table summarises the assignment of features. [a,b,c,d] Loss of information from HERV database to understandable functions. [a] Summary of Additional file 2: Table S1a, [b] Summary of Additional file 2: Table S1b, [c] Enumeration of LTRs whose function is attributable, i.e. defined as LTRs combining U3 and U5 adjacent structures on the genome and existing probesets onto the chip allowing discrimination between U3 and U5 expression signals, [d] Enumeration of LTRs whose function is attributed using both a $2^{4.5}$ positive threshold and a fold change of 3 between U3 and U5 regions. More specifically, to retain sensitivity and robustness with regard to function assignation, we voluntarily selected a lower expression level cut-off of $2^{4.5}$ for positive signal attribution. As such, the LTR was referred to as promoter (Pr), polyadenylation signal (pA), readthrough (RdT) or Silent. All other remaining LTRs were classified as undetermined. **c** Specialisation of LTR features. Number of LTRs from the "HERV_prototypes" repertoire according to all the combinations of functions observed in each of the 45 PBMC samples. Silent LTR, Pr: LTR referred to as a promoter, pA: LTR referred to as a polyadenylation signal, RdT: read-through. The combinations obtained could either contain one (e.g. Pr), two (e.g. Silent/Pr), three (e/g. Silent/pA/Pr), or four functions (Silent/pA/Pr/RdT). 987 LTRs were excluded from the analysis as classified at least once as undetermined

56.6% of LTRs (3076 LTRs) were systematically silent in all samples. Interestingly, 27.6% of LTRs shifted from a silent status to a promoter (902 Silent/Pr LTRs) or polyadenylation (598 Silent/pA LTRs) function. Although poorly represented (2%), the same shift is observed for RdT LTRs. Some LTRs are exclusively RdT (20 LTRs) and some shifted from RdT status to a promoter (37 RdT/Pr LTRs) or polyadenylation (51 RdT/pA LTRs) function. A significant proportion of LTRs (13.1%) were in promoter (402 Pr LTRs) or polyA (311 pA LTRs) status in all samples (candidate loci validation, Additional file 8: Figure S6d). Almost no LTR (0.6%) shifted through at least three different features including both promoter and polyA functions. Consequently this observation confirms that the shift from promoter to polyA function is an extremely rare if significant event. The same analysis was performed using the complete HERV/MaLR dataset (Additional file 3: Figure S2). Although results should be considered with caution due to the imprecise annotation of HERV_Dfam and MaLR_Dfam LTRs (Additional file 2: Table S1), the overall trends were similar to those observed at the "HERV_prototypes" repertoire level.

The genomic environment encompassing the functional and silent LTRs is depicted in Additional file 4: Figure S3. As previously observed [29], the gene density ratio is almost 1.2 times higher for promoter LTRs than for silent LTRs. Meanwhile, the proportion of intragenic LTRs is biased towards the antisense representation, with about two-thirds of LTRs being antisense to the gene in which they are located, regardless of their functional category. About 80% of intragenic LTRs overlap with introns. Although no major bias in the genomic environment of intergenic LTRs (Additional file 4: Figure S3b) could be associated with their function, some trends are observed. A plateau is observed with constitutive promoter LTRs, reflecting a slightly lower occurrence of genes in the sense orientation up to 10 kb upstream of these LTRs. Symmetrically, sense gene occurrence apparently rises faster than for antisense genes in the downstream zone of silent LTRs compared to promoter LTRs.

Gene and HERV/MaLR modulation following LPS stimulation

To gain insight into the modulation of genes and HERV/MaLR expression associated with PBMC stimulation conditions, we performed both a hierarchical clustering analysis and a supervised statistical analysis in pairwise conditions. For the record, PBMCs from 5 healthy volunteers were cultured in triplicate, (i) without any additional stimulation (NS for non-stimulated), (ii) with a high concentration of LPS to mimic inflammation, (iii) and primed with a low-dose LPS and latterly boosted with a high-dose LPS defining the so-called endotoxin tolerance model (ET) that mimics monocyte anergy (Fig. 3a). We checked our model by means of TNF-α pro-inflammatory and



	Down-regulated	Up-regulated
Number of HERVs/ MaLRs DELs	44	67
Highly significant HERVs/ MaLR DELs	MLT1D (3) *chr19:47342944-47343392*	MLT1H2 (2) *chr6:113676857-113677372*
	MLT1A (1) *chr10:19858820-19859254*	MLT1J2 (1) *chr4:184855411-184855765*
	MLT2D (1) *chr12:13217752-13217954*	MER34 (2) *chr5:151019711-151020042*
	LTR101_Mam (3) *chr19:54891074-54891496*	LTR47A2 (1) *chr12:68905669-68906092*
	PRIMA41_10v1 (1) *chr10:103844866-103845768*	MSTD (2) *chr7:20346887-20347244*
Number of DEGs	64	68
Highly significant DEGs	MERTK (10)	PTGS2 (32)
	CLEC7A (9)	IL12B (19)
	CD36 (10)	IL6 (20)
	TIMP2 (17)	IL1A (16)
	CCL13 (4)	CCL20 (14)

	Down-regulated	Up-regulated
Number of HERVs/ MaLRs DELs	42	5
Highly significant HERVs/ MaLR DELs	MLT2B5 (3) *chr1:78623489-78623954*	HML-6_10v0 (1) *chr10:5266257-5266693*
	MSTC (2) *chr1:78648318-78648697*	MLT1A (1) *chr4:32500678-32501058*
	ERV3-16A3_I (2) *chr12:112972627-112975754*	LTR16A (1) *chr14:38847135-38847466*
	MLT1A (2) *chr1:78655872-78656221*	MLT2D (1) *chr12:13217752-13217954*
	LTR33B (2) *chr12:112971073-112971451*	MLT1A (1) *chr10:16734769-16735152*
Number of DEGs	50	8
Highly significant DEGs	MX1 (45)	IFNG (1)
	OAS1 (28)	NEFH (3)
	CCL15 (4)	MMP10 (2)
	OAS3 (46)	SERPINB2 (2)
	EIF2AK2 (44)	THBD (2)

Fig. 3 (See legend on next page.)

(See figure on previous page.)
Fig. 3 Genes and HERV/MaLR modulation following LPS stimulation. **a** Schematic representation of the dose-dependent LPS challenges known as the endotoxin tolerance model. Biological triplicates of PBMCs from 5 healthy volunteers were cultured and stimulated. Efficiency of stimulations was validated with TNF-α and IL10 quantitation by ELISA (Additional file 4: Figure S3) prior to HERV-V3 microarray experiments. **b** Heatmap from hierarchical clustering (correlation distance, complete method) of the 1% most variable probesets (all repertoires included), group samples according to their stimulation condition. Non-stimulated (NS), low-dose LPS primed PBMCs (ET), single high-dose lipopolysaccharide challenge (LPS); high-expression level (yellow), low-expression level (blue). **c** Differential gene and HERV/MaLR expression analysis. The first row shows volcano plots derived from the differential expression analysis; on the left for LPS vs NS (inflammatory context), and on the right for ET vs LPS conditions (immunocompromised/unresponsiveness context). The x-axis represents the log2 fold change values, and the y-axis the log10 adjusted *p*-values. Each point represents these values for a probeset. Coloured points show the significantly modulated probesets (adjusted p-value < 0.05, log2FC < − 1 (red) or log2FC > 1 (green)). The tables in the middle row present the number of statistically significantly differentially expressed elements, at locus level (DELs) for HERVs/MaLRs and differentially expressed genes (DEGs). Down-modulated loci are in red, up-modulated loci are in green. For HERV/MaLR elements, the name, number of differentially expressed probesets (between brackets) and chromosomal locations (in italic) are indicated (GRCh38 version of genome). For genes, the current gene symbol and the number of differentially expressed probesets (between brackets) are indicated. The last row represents canonical pathways identified using Ingenuity Pathways Analysis tool (https://analysis.ingenuity.com) and signals derived from HTA probesets contained on the HERV-V3 chip. Canonical pathways predicted to be significantly activated (orange) or inhibited (blue) between LPS vs NS and ET vs LPS conditions are depicted (z-scores ≥2 and z-scores ≤ − 2; *p*-value cut off of 0.05, Fisher's exact test)

IL-10 anti-inflammatory cytokine quantification both in supernatants (protein level) and cellular extracts (mRNA level). The combined profiles of TNF-α as a "tolerisable gene" which exhibits a lower response, and as IL10, a "non tolerisable gene", whose expression was increased or unaltered, validate the endotoxin tolerance model (Additional file 5: Figure S4). Figure 3b shows a heatmap from hierarchical clustering of the 1% most variant probesets (genes and HERVs/MaLRs) across samples. We observed 3 main expression profiles: (i) non tolerisable probesets with low expression for the NS condition and high expression for the ET and LPS conditions (top, Fig. 3b), (ii) tolerisable probesets with low expression for NS and ET, and high expression for the LPS conditions (bottom, Fig. 3b), and (iii) down-modulated probesets with high expression for the NS condition, and low expression for the LPS and ET conditions (middle, Fig. 3b). Hence, as stimulation conditions appeared to be strong drivers of PBMC transcriptome modulation, we performed a differential expression analysis. Differentially expressed genes are hereinafter referred to as DEG and differentially expressed HERVs/MaLRs loci referred to as DEL. Volcano plots are depicted at the probeset level for both genes and HERVs/MaLRs. They illustrated that similar amounts of elements appeared to be over- or under-expressed in an inflammatory context (LPS), whereas more gene and HERV/MaLR probesets were down-modulated under tolerance condition (ET) (Fig. 3c, top row).

Overall, differential expression analysis of LPS versus NS conditions identified 785 up-regulated probesets and 847 down-regulated probesets (corresponding to 243 distinct HERV/MaLRloci), while analysis of ET versus LPS conditions merely identified 38 up-regulated probesets and 677 down-regulated probesets (corresponding to 105 distinct HERV/MaLR loci) (adjusted *p*-value < 0.05, |log2FC| > 1) (Fig. 3c, middle, and Additional file 6: Table S2). Following LPS stimulation, out of the 111 differentially expressed HERV/MaLR elements, 38 belong to HERV_Dfam, 51 to MaLR_Dfam, and 22 to HERV_prototypes. Although DEL analysis does not indicate any group or chromosomal enrichment, some characteristic points could be observed. ERV9, ERV-E4.1, HERV-FRD, HERV-H, HERV-I, HML8, PABL-A and PRIMA4 groups were exclusively modulated under inflammatory condition (LPS vs NS). HERV-Fb, HERV-L, HERV-T and PRIMA41 groups were both modulated for LPS vs NS and ET vs LPS (see below). Among the DELs, 64% consist of solo LTRs and 36% of complete or partial proviruses. After LPS stimulation, PTGS2, IL1B, IL12B, IL6, IL1A and TNF-α pro-inflammatory cytokine as well as CCL20 and PTX3, were the most up-modulated genes. MERTK, CLEC7A, CD36, TIMP2 and CCL13 were the most down-modulated genes. Moreover, we observed, with a pathway analysis, an enrichment of the Death Receptor Signalling and NF-κB Signalling pathways (Fig. 3c, last row). Notably, these results highlighted significant inactivation of "LPS-stimulated MAPK signalling". This may be due to negative feedback resulting from the high production of TNF-α following LPS stimulation.

In a tolerance context, among the 47 differentially-expressed HERV/MaLR elements, 20 belonged to HERV_Dfam, 20 to MaLR_Dfam and 7 to HERV_prototypes. As observed for LPS vs NS, no group is enriched or depleted among the most DELs. Notably, 11 elements out of 47 (23.4%) are located on chromosome 12. The HERV-HS49C23, HERV-F and HML6 elements were exclusively modulated in the ET vs LPS condition. Some groups had differentially expressed loci in LPS vs NS and other differentially expressed loci in ET vs LPS, such as HERV-Fb, HERV-L and HERV-T groups. Among the 47 DELs, 86% consist of solo LTRs and 14% of complete or partial proviruses. In a tolerance context, IFNγ, NEFH, MMP10, SERPINB2 and THBD were the most up-modulated genes. MX1, OAS1, CCL15, OAS3, EIF2AK2 and TNF-α were

the most down-modulated genes. Consistently, there were 3 inhibited pathways: "role of pattern recognition receptors in recognition of bacteria and viruses", "NF-κB signalling" and "TLR signalling"(Fig. 3c, last row). Conversely, "LXR/RXR signalling" was highly activated, putatively reflecting LXR-induced inactivation of the NF-κB signalling pathway leading to the anti-inflammatory macrophage phenotype in atherosclerosis (Fig. 3c, last row; [54]).

Differential expression analysis allowed us to identify modulated HERVs/MaLRs and genes involved in inflammatory and tolerance contexts. Before considering functional linking of HERVs/MaLRs and genes, we validated the results obtained with microarrays through the use of the RT-qPCR reference method.

RT-qPCR confirmation of condition-related expression of HERV/MaLR elements

To confirm modulation of expression observed with the HERV-V3 microarray, we selected the 44 HERV/MaLR probesets for RT-qPCR validation,. Of these, 31 were the most differentially expressed probesets in Dfam repertoires and 13 the most differentially expressed probesets among the prototype repertoire. HERV/MaLR locus-specific RT-qPCR systems were meticulously designed and validated to secure locus specificity (see Additional file 7: Figure S5), leading to 32 primer pairs out of 44 of selected candidates. We confirmed HERV/MaLR modulation on the samples used for microarray analysis, and then on an independent cohort of 6 healthy volunteers. PCR products could be obtained on 23 out of 32 primer pairs (depicted in Additional file 7: Figure S5). Overall, 87% of the detectable elements had concordant profiles with HERV-V3 microarray data. Twenty loci exhibited similar expression profiles in microarray and RT-qPCR experiments and 3 exhibited conflicting profiles (Additional file 8: Figure S6).

Notably, a majority of HERV/MaLR elements exhibited similar patterns as "tolerisable" or "non tolerisable" genes, i.e. divergently modulated following LPS and ET treatments. Figure 4a illustrates the comparable "tolerisable" behaviour of TNF-α, 121601901-HERV0116uL, and 08114670-MALR1129uL HERV loci. Figure 4b depicts the similar "non tolerisable" behaviour of IL10, 070278702-MALR1045uL, and 043166701-MALR1020uL MaLR loci. These two phenotypes were observed regardless of their distance from genes. Nevertheless, 121601901-HERV0116uL, 070278702-MALR1045uL, and 043166701-MALR1020uL are located within OAS3, ITGB8 and MIR3945HG genes, respectively. Conversely, 08114670-MALR1129uL is at a distance of more than 100 kb from the closest gene. Interestingly, as known for TNF-α [51], the tolerisable HERV phenotype was reversed by IFN-γ (Fig. 4a and b, column c). Hence, as HERVs/MaLRs and genes seemed to share similar control of expression following stimulation, we attempted to integrate HERVs/MaLRs in gene pathways based on their common transcriptional behaviour.

An integrative view of HERVs/MaLRs and genes in immunity pathways

Differential expression analysis identified genes and HERV/MaLR loci for which expression was modulated in inflammatory or tolerance contexts. Some HERVs/MaLRs and genes had strongly correlated expression profiles. We sought to identify which HERV elements may be co-expressed with genes which were integrated in regulatory networks. First, using Ingenuity Pathway Analysis (IPA), we identified 13 activated or repressed canonical pathways in inflammatory or tolerance contexts. In 11 out of the 13 pathways, 26 genes belonged exclusively to one pathway. We then identified 72 probesets corresponding to 62 HERV/MaLR loci strongly correlated with the 26 genes (correlation ≥0.8) (Additional file 9: Figure S7a-c, Additional file 10: Table S3). Among those 62 HERV and MaLR loci, 7, 26 and 29 belonged to "HERV_prototypes", HERV_Dfam and MaLR_Dfam repertoires, respectively. This allowed us to build a global network integrating HERV and MaLR loci within gene pathways (see Additional file 9: Figure S7d). Twenty-six out of the 62 HERV/MaLR elements belonged exclusively to one pathway, and 36 HERV/MaLR loci are associated with 2 to 7 pathways. Eleven HERV or MaLR loci were identified in the vicinity of genes (≤ 40 kb), and they were integrated into 1 to 4 pathways. Conversely, although spread on various chromosomes, a bundle of 8 HERV/MaLR elements contributed to both the LXR/RXR activation pathway activated under ET condition and the NF-κB signalling pathway activated during inflammation, via PTGS2 and FLT1 genes (see Additional file 9: Figure S7e).

A gene centred view of the "Role of pattern recognition receptors in recognition of bacteria and viruses" (PRR) pathway is depicted in Fig. 5a. It includes 32 HERV/MaLR loci and 7 genes exclusive to the pathway. Ten out of the 32 HERV/MaLR loci belong exclusively to this pathway. Most of the genes are co-expressed with several retroviral elements, 18 for IFIH1, 18 for IRF7, 17 for OAS3, 15 for OAS2, 7 for PTX3 and 3 for C5AR1, widespread on distinct chromosomes. Interestingly, 4 tolerisable HERV/MaLR loci are present on chromosome 12 within a 39 kb region which overlaps with OAS2 and OAS3 tolerisable genes. Moreover, 4 HERV/MaLR loci are on chromosome 1 in a 32 kb region overlapping with IFI44L and IFI44 genes. RT-PCR amplification of IFI44L, IFI44 genes and 011052301-HERV0472uL HERV loci showed a marked decrease for the ET condition (data not shown). Notably, most of these HERV/MaLR elements expressed belonged to the 3'UTR of several gene transcripts: 121601802-HERV0492uL and 121601901-HERV0116uL

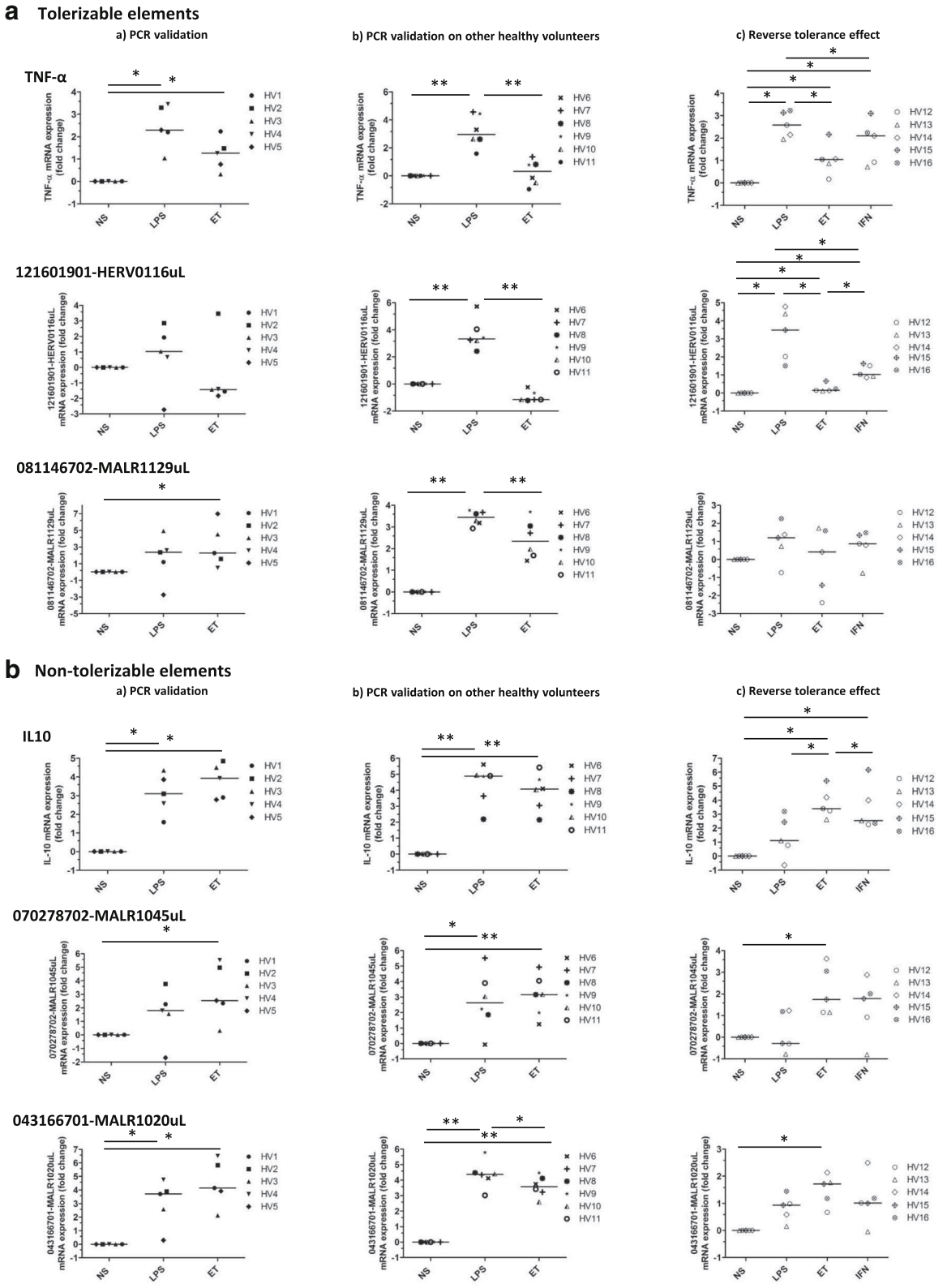


Fig. 4 (See legend on next page.)

(See figure on previous page.)
Fig. 4 RT-qPCR validation of differentially expressed candidate loci following various LPS dose challenges. This figure illustrates the RT-qPCR expression of **a** tolerisable genes (TNF-α) and HERV/MaLR elements (121601901-HERV0116uL and 081146702-MALR1129uL) exhibiting a TNF-α–like pattern of expression and **b** non-tolerisable genes (IL10) and HERV/MaLR elements (070278702-MALR1045uL and 043166701-MALR1020uL) exhibiting an IL10–like pattern of expression. Expression was measured using mRNA derived from the stimulated and unstimulated PBMCs of the healthy volunteers used in the discovery microarray experiment (first column, (Aa and Ba)), mRNA derived from the stimulated and unstimulated PBMCs of 6 additional healthy volunteers managed in the same way (second column, (Ab and Bb)), and finally mRNA derived from the stimulated and unstimulated PBMCs of 5 additional healthy volunteers (third column, (Ac and Bc)). This last column includes additional IFN-γ dependant reversibility of tolerance, consisting of a 2 ng/mL LPS priming step overnight, followed by a 100 ng/mL IFN-γ stimulation step overnight, and finally the 100 ng/mL LPS stimulation step for 6 hours. All PCR reactions were performed in duplicate for each condition. Expression of the housekeeping genes PPIB and RPLP0 was monitored for normalisation. The fold change (FC) was determined using the $2^{-\Delta\Delta Ct}$ method. The final value of the unstimulated condition was arbitrarily set to one and other values scaled-up in order to provide a final relative differential expression (data were represented by a median and using the log2 scale). Statistically significant differences between two conditions are marked (wilkoxon signed rank test. **: p-value < 0.05 and * p-value < 0.1)

for OAS3–201, 121602201-HS49sLRp for OAS2–203, 121602901-MALR1023uL for OAS2–202, and 011052301-HERV0472uL and 011052401-HERV0462uL for IFI44L-201.

Alternatively, an HERV/MaLR-centred view is proposed in Fig. 5b and highlights 18 loci which belong to at least 4 different networks. Eleven out of 15 HERV/MaLR tolerisable loci were shared by 4 pathways (PRR, RIG-1, Interferon and IRF pathways). These 11 elements were co-localised with or close to conventional genes (≤ 10 kb). Non-tolerisable elements 070278702-MaLR1045uL and 043166701-MaLR1020uL and tolerisable element 081146702-MALR1129uL are at the crossroad of several pathways, namely "LXR/RXR activation", "TREM1 signalling", and "Dendritic cell maturation" Altogether, this interplay between HERVs/MaLRs, genes and pathways suggested complex and heterogeneous transcriptional regulation mechanisms.

Discussion

HERV transcriptome in PBMCs

We used the HERV-V3 chip to provide a first overview of the HERV transcriptome in PBMCs. Healthy, inflammatory and immunocompromised/monocyte anergy states were simulated using an endotoxin tolerance model [50]. We observed that about 5.6% of LTR retrotransposons were transcriptionally active in PBMCs, reaching 9.4% in the well-annotated "HERV_prototypes" repertoire. Previous works estimated the extent of the HERV transcriptome, compared to the HERV genome, to be between 7 and 30% [29, 55–57]. These differences could be related to (i) disease conditions [29, 55, 57], (ii) tissue specificity [29, 55], and (iii) technology [29, 55–57]. Although various intensity thresholds (background/expressed) may also account for such differences, we confirmed in our study that a significant amount of HERV and MaLR sequences thought to be silent are actually expressed. HERV or MaLR regions might be either embarked by conventional genes, or self-induced. Indeed, an LTR could possess promoter activity or harbour polyadenylation signals used for proviral gene expression or adjacent non-retroviral genomic sequences. The vast majority of LTRs were silent (70.5%), and the minority were transcribed via readthrough mechanism (2.5%), regardless of the condition. The large amount of silent LTRs is in line with the accumulation of inactivating mutations [58] as well as silencing by epigenetic mechanisms [59, 60]. A relatively balanced state was observed between putative promoter (15.2%) and polyadenylation (11.8%) functions. The same overall trends were previously observed in cancer tissues [29], although the amount of silent LTRs was higher in our study (70.5% versus 47%), probably due to the largely increased number of targeted groups. Altogether, integrating both more precise data derived from the "HERV_prototype" repertoire and results from the complete HERV/MaLR dataset, about one tenth of LTRs exhibit a constitutive promoter or polyA function while roughly one quarter of LTRs may shift between silent and promoter or polyA functions. Only a few thousandths of LTRs can shift between promoter and polyA, as initially described in cancer [29], strengthening what we called "operational determinism", i.e. an LTR is predetermined to act as a promoter or a polyA site. Although all "prototype" groups are active in PBMCs, we observed a higher proportion of gamma-retroviruses, including notably the super spreader HERV-H group and the HERV-W group containing Syncytin 1. This is consistent with group-based PCR approaches in PBMCs [36], and in MDM cell lines [41], as well as with the detection/modulation of expression of MSRV/HERV-W and HERV-H loci in PBMCs of healthy donors and multiple sclerosis patients [15, 45, 61, 62]. Concerning beta-retroviruses, HERV-K/HML elements were expressed in PBMCs as previously observed in healthy subjects [36, 41], in prostate cancer [37], and Henoch-Schönlein purpura [63] patients.Notably, centromeric HML2 elements seemed to be relatively more represented than the other HML-2 elements, as observed in the blood of HIV-infected patients [64]. Nevertheless, this exhaustive HERV-dedicated microarray allowed us to detect expression of poorly characterised groups such as PRIMA41 and MER52A which merit further investigations.

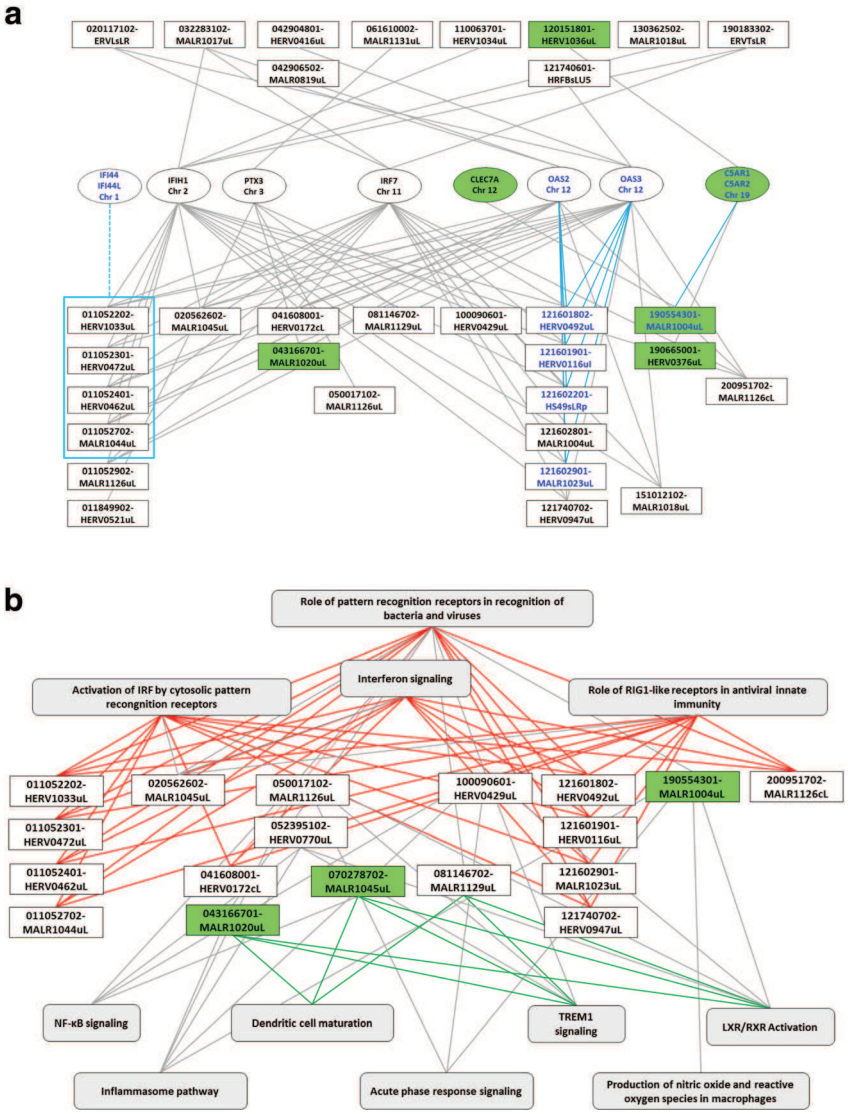


Fig. 5 Expression profile-based integration of HERVs/MaLRs in immunity genes related pathways. **a** The gene-centred network illustrates HERV/MaLR integration in the "role of pattern recognition receptors in recognition of bacteria and viruses" pathway (PRR). Seven genes specific for this pathway are depicted (oval frames) together with their parental chromosome. HERV/MaLR elements associated with these unique genes were identified using both Ingenuity Pathway Analysis and co-expression between HERVs/MaLRs and genes. Tolerisable and non-tolerisable elements are represented by empty or green boxes, respectively. Ten HERV/MaLR elements belong exclusively to the PRR pathway (top rectangular frames), while 22 HERVs/MaLRs loci are shared by several pathways (bottom rectangular frames). Five HERV/MaLR elements are located at a distance of less than 40 kb from the OAS3 and OAS2 or C5AR1/C5AR2 genes (dark blue font and turquoise connectors, C5AR2 absent from the microarray). Four HERV/MaLR elements are located in a 40 kb window (large turquoise rectangle) flanked by IFI44 and IFI44L genes (connected by a dotted line as absent from the microarray). **b** HERVs/MALRs integration in several pathways. The HERV/MaLR-centred network illustrates the integration of HERVs and MaLRs elements within 11 networks (grey boxes). Tolerisable and non-tolerisable elements are represented by empty or green boxes, respectively. Ten out of 15 tolerisable HERV/MaLR elements are exclusively linked to the 4 pathways at the top (red connectors), "activation of IRF by cytosolic pattern recognition receptors", "role of pattern recognition receptors in recognition of bacteria and viruses", "interferon signalling" and "role of RIG1-like receptors in antiviral innate immunity". Three elements are exclusively associated with 3 pathways (2 out of the 3 non-tolerisable elements and 1 tolerisable element), "dendritic cell maturation", "TREM1 signalling", and "LXR/RXR activation" (green connectors)

Modulation of HERVs and genes after inflammation and tolerance induction

After looking at the HERV transcriptome landscape in PBMCs, we analysed whether individual HERVs/MaLRs could be finely regulated upon stimulation. Hierarchical clustering aggregated samples according to their stimulatory status and DE analysis showed that many HERVs/MaLRs and genes were modulated between conditions. We observed a similar amount of up-regulated and down-regulated elements in the LPS

condition compared to NS, as previously shown using group-based RT-PCR systems [36, 41]. Such dichotomic activity of HERVs/MaLRs was found in the highly inflammatory context of burn patients [65, 66]. In mice, LPS stress induced primary lymphoid cell-specific production of MLV-ERV virions [67]. More generally, microbes modulated ERV transcription in mice [42]. In a tolerance context, HERVs/MaLRs and genes tended to be down-regulated. Altogether, these observations suggest various levels of HERV/MaLR control, including a fine-tuned control similar to conventional genes.

The NF-κB signalling pathway was both up-regulated following LPS stimulation and down-regulated following tolerance induction (LPSvsNS, Z score: 2.121; ETvsLPS, Z score: − 2.646). This is highlighted by the pro-inflammatory cytokines, IL12B, IL6, IL1A, TNF-α modulation, as well as IL10 anti-inflammatory variation, as previously described in such contexts [50, 68–70]. As observed for TNF-α and IL10 genes, HERV/MaLR elements exhibit a dichotomy of tolerisable versus non tolerisable phenotypes, although some elements were found to be down-regulated in both conditions. As illustrated by 121601901-HERV0116uL and 081146702-MALR1129uL loci, HERV/MaLR tolerisation was surprisingly reversible upon IFN-γ addition, as previously described for TNF-α [51, 52]. Again, this demonstrates that HERVs/MaLRs and genes share similar regulation control following stimulation inducing inflammation or anergy in PBMCs. These results confirm the existence of tight HERV expression regulation control, as previously suggested by the tropism-related behaviour of HERV elements in solid tissues and in particular in reproductive tissues [3] and cancer [29].

HERV integration within gene immunity pathways

The comparable responses of HERVs/MaLRs and genes to high-dose, low-dose LPS and IFN-γ-dependent reversibility of tolerance may be due to either autonomous HERV expression potentially driving gene expression, or gene expression potentially embedding HERV expression. Pathway and co-expression analyses allowed us to include 62 HERV/MaLR loci and 26 genes which are exclusive to 1 pathway into regulatory networks. Eleven canonical pathways were activated or repressed in inflammatory or tolerance contexts. Among the integrated HERV/MaLR elements, ten loci mapped to the 3'UTR of OAS2, OAS3, IFI44L, IFI44, C5AR1 and C5AR2 genes. These retroviral elements may contribute to the post-transcriptional control of these transcripts, including polyA signalling, nucleocytoplasmic transport, translation efficiency, localisation and stabilisation of mRNA [71]. The complex interplay of HERVs and genes can be illustrated by the IFI44L to IFI44 gene region on chromosome 1. The 2 genes and the 4 HERV/MaLR loci located between them are similarly down-regulated under ET condition. Such similar transcriptional expression of IFI44L and IFI44 was previously described in purified CD14 monocytes of patients with Sjogren's syndrome [72, 73]. Interestingly, dissociated IFI44L versus IFI44 expression was observed in high versus low IFN-γ producers following *Leishmania braziliensis* stimulation in PBMCs [74]. In addition, among the 4 HERV/MaLR elements, the 011052702-MALR1044uL locus corresponded to an annotated CTCF binding region which defines the boundary between active and heterochromatic DNA. The 011052702-MALR1044uL locus may therefore regulates IFI44L and IFI44 expressions in some situations.

The co-expression of genes with a large number of HERV/MaLR elements scattered among different chromosomes, e.g. 18 HERV/MaLR loci located on 10 distinct chromosomes linked with IFIH1, 6 HERVs/MaLRs located on 6 distinct chromosomes linked with PTX3, suggested that they are part of gene networks regulated by shared transcription factors, as previously proposed [33]. However, we did not observe any tissue-specific transcriptional factor binding sites (TFBS), but an enrichment of the AP-1 binding site with promoter LTR was observed as compared to silent LTR (p-value: 1.14 10^{-4}, data not shown). The similar modulation of MALR1126 LTR promoters belonging to LXR/RXR and NF-κB networks and located on chromosome 5 (050017102-MALR1126uL) and 10 (100175702-MALR1126uL) may reflect such shared regulation. The integration of LTRs within several pathways is a first suggestion that HERVs/MaLRs and genes are similarly regulated. The HERVs/MaLRs integrated in PRR, RIG1, IFN and IRF pathways and which are mapped into gene transcripts, are mostly tolerisable and carry H3K36me3, a mark of actively transcribed regions in normal haematopoietic cells. Most HERV/MaLR elements, either tolerisable or non-tolerisable and integrated into LXR/RXR, TREM1 and dendritic cell maturation or LXR/RXR and NF-κB pathways, are at a distance of more than 25 kb from the closest gene and carry mainly H3K27me3, and occasionally H3K9me3, mark of repressed regions in normal haematopoietic cells. Notably, the IFI44-associated 011052702-MALR1044uL and SLC30A4-associated 150312301-HERV0498uL loci appeared to be decorated with H3K9me3 and H3K27me3 repressive histone marks, respectively. As a general trend, LTRs were screened for particular histone modification signals by overlap with Encode peaks for H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3 and H3K9me3 in different immune cells; we found enrichment for the H3K9me3 mark for CD14+ monocytes. The LPS stimulations probably modified the local chromatin configuration leading to global modulation of expression, of the gene and HERV/MaLR [75]. Intriguingly, it appears that CD14-positive

monocytes are at least twice more enriched in H3K9me3 histone mark than CD4- and CD8-positive T cells, B cells and natural killer cells (data not shown). Taken together, the variability of histone marks, the tolerisable/non-tolerisable HERV/MaLR phenotypes and vicinity of genes suggest a common control of expression between HERV/MaLR and genes. In addition to a contribution to cis-regulation, HERV/MaLR presence at the crossroads of different regulatory networks and conservation in the human population (data not shown), may suggest a role in trans-regulation linking enhancer and promoter regions [33, 76]. A better understanding of the causal relationship between HERV/MaLR, genes and regulatory pathways would merit further investigations.

Conclusion

This microarray-based approach revealed the expression of about 47,466 distinct HERV loci and identified 951 putative promoter LTRs and 744 putative polyA LTRs in PBMCs. HERV/MaLR expression was shown to be tightly modulated following several stimuli including high-dose and low-dose LPS as well as IFN-γ. This allowed us to propose an integrative view of HERVs/MaLRs and genes in global functional pathways. Further systematic analyses will be required to gain insight on the modulation of expression of HERV/MaLR loci in different haematopoietic cell types, including monocytes, B and T cells, as well as neutrophils and NK cells. This may help decipher the multiple levels of HERV functions in haematopoietic cells, as locally illustrated by the surface or intracellular envelope on monocytes [45] or glial cells [77], or by non-coding elements involved in the control of cell differentiation [78]. From an in vivo point of view, this approach paves the way for systematic deciphering of modulated retroviral elements associated with autoimmune diseases such as systemic lupus erythematosus, inflammatory diseases such as type 1 diabetes [79] inherited autoimmune and auto-inflammatory disorders such as type 1 interferonopathies (reviewed in [80, 81]), and virus- or drug-induced immunocompromised states [82], as well as resulting from a compensatory response to hyperinflammation such as in sepsis [83]. Notably, it would be of interest to investigate whether the altered histone methylation recently observed for genes in LPS-induced tolerance and in septic patients [75, 84] may affect HERV expression and contribute to sepsis.

Methods

Biological samples and quality control

Citrated pouches or heparinised tubes blood were obtained from EFS (Etablissement Français du Sang) and used immediately. According to EFS standardised procedures for blood donation and to provisions of the articles R.1243–49 and following ones of the French Public Health Code, a written non-opposition to the use of his donation for research purposes was obtained from healthy volunteers. The blood donors' personal data were anonymised before blood transfer to our research lab. We obtained the favourable notice of the Local Ethical Committee (Comité de Protection des Personnes Sud-Est II, Bâtiment Pinel, 59 Boulevard Pinel, 69,500 Bron) and the acceptance of the Ministère de la Recherche (declaration DC-2008-64) for handling and conservation of these samples. Peripheral blood mononuclear cells (PBMCs) were isolated with Unisep tube density gradient centrifugation (Eurobio) and washed with sterile PBS (phosphate buffered saline) (Eurobio). The PBMCs were adjusted to 2×10^6 cell/mL and cultured in X-Vivo 20 Medium (Lonza) at 37 °C and 5% CO_2. All the experiments were perfomed in triplicate. Lipopolysaccharide was purchased from Sigma-Aldrich and was a mix of *Escherichia coli* O111:B4, O55:B5 and O127:B8 (Sigma). In this ex vivo endotoxin tolerance model, the PBMCs were first cultured for 15 h without (control group NS and LPS cells), or with 2 ng/ml LPS (ET cells). After washing steps, the PBMCs were incubated a second time for 6 h without (control group NS), or with 100 ng/ml LPS (LPS and ET cells) (Fig. 3a). In this model, when specified in the text, the effects of recombinant human IFN-γ to reverse tolerance effects were studied. Another incubation phase was performed for 24 h with 100 ng/mL of human IFN-γ1b (Miltenyi Biotec) or vehicle, between the two LPS incubations. At the end of the experiments, the supernatants were retrieved and stored at − 80 °C. Pro-inflammatory cytokine TNF-α and anti-inflammatory cytokine IL10 concentrations in the PBMC culture supernatants were detected using commercially-available ELISA kits from R&D System, in accordance with the supplier's recommendations. The cells were harvested, lysed in RLT buffer supplemented with β mercaptoethanol and stored at − 80 °C until further processing. The total RNA was extracted from PBMCs using RNeasy Mini kit (Qiagen) according to the manufacturer's instructions. For each RNA extraction, the residual genomic DNA was digested using the gDNA Eliminator spin column (Qiagen), and directly on RNeasy spin column using RNase-Free DNase Set (Qiagen). RNA quantity and quality were determined using Nanodrop (Thermo Scientific) Bioanalyser 2100 (Agilent) according to the manufacturer's instructions [51, 52].

Custom Affymetrix HERV-V3 GeneChip microarray

HERV-V3 targets 353,994 loci-elements, represented by 4,410,200 probes. The custom HERV GeneChip can discriminate between distinct HERV elements composed of a set of highly informative probesets (located in U3, R, U5 subdomains of solo, 5' and 3' individual LTRs and

gag/*pol*/ *env* regions), hereinafter referred to as 'HERV prototypes repertoire', and a set of probesets with lower-quality annotations (located in the first third and last third of the complete LTR, and every 2.5 kb in the region in between LTRs), hereinafter referred to as 'HERV/MaLR_Dfam repertoire'. The custom HERV GeneChip also contains probesets targeting LINE1, lncRNA, viruses, and the gene repertoire. The descriptions of the HERVgDB4 database and of the final contents of the HERV-V3 microarray are provided in Additional file 2: Table S1 [30].

RNA amplification, labelling and hybridisation

The cDNA synthesis and amplification steps were performed using 16 ng of RNA with the Ovation Pico WTA System V2 kit (Nugen) according to the manufacturer's instructions. Five micrograms of amplified purified DNA were fragmented into 50–200 bp fragments and were 3-labeled using the Encore Biotin Module kit (Nugen) according to the manufacturer's instructions. The HERV-V3 microarrays were hybridised at 50 °C for 18 h in an oven with constant stirring (60 rpm). Washing and staining were carried out according to the protocol provided by the manufacturer, using the GeneChip fluidics station 450 (Affymetrix). The arrays were finally scanned using the GeneChip scanner 3000 7G (Affymetrix) fluorometric scanner. Images (DAT files) were converted to CEL files using GCOS software (Affymetrix) [30]. The experimental data generated have been filed with the National Center for Biotechnology Information (NCBI) and are available on the GEO DataSets site under access number GSE108239.

Bioinformatics analysis

Microarray analysis pre-processing was detailed in supplementary methods (Additional file 11: Supplementary Methods). We chose a specific threshold to define LTR functions for the HERV_Dfam repertoire. We used the dichotomy of probeset signal targeting. More specifically, to retain sensitivity and robustness with regard tofunction assignation, we voluntarily and arbitrarily selected a relatively low expression level cut-off of $2^{4.5}$ for positive signal attribution coupled with a significant fold change between U3 and U5 signals. Therefore, an LTR was referred to as 'promoter' (Pr) in cases where the signal of the U5-associated probeset was (i) over the threshold, and (ii) at least 3 times higher than its U3 counterpart, and as 'polyadenylation signal' (pA) if the intensity of its U3-associated probeset was (i) over the threshold and (ii) at least 3 times higher than its U5 counterpart. An LTR was assigned as 'readthrough' (RdT) if both U3 and U5 signals (i) were over the threshold, and (ii) without significant fold change between its. Finally, an LTR was assigned as silent if U3 and U5 were both under the threshold; all other remaining LTRs were classified as undetermined (112 LTRs). To visualise HERV and gene co-expression, hierarchical clustering based on correlation distance with the average method was performed on the 1% most variable probesets. Subsequently, comparisons between i) unstimulated PBMCs (NS) and PBMCs stimulated once with LPS (LPS), and ii) tolerant PBMCs re-stimulated with LPS (ET) and LPS were carried out. For all probesets, for differential expression analysis, moderated t-tests were performed (Limma, v3.22.7 [85], and *p*-values adjusted for multiple testing using the Benjamini-Hochberg procedure [86]. A probeset was considered to be statistically significantly differentially expressed when the absolute log2 Fold Change (|log2FC|) was over 1 and the adjusted-p-value under 0.05. Graphs were generated using ggplot2 (v2.2.0) or pheatmap (v1.0.8). Finally, the Ingenuity Pathways Analysis tool (IPA, Ingenuity® Systems, https://analysis.ingenuity.com) was used to assess upstream regulators, canonical pathways, disease, and functions. Details of this analysis were presented in supplementary methods (Additional file 11: Supplementary Methods). HERVs with a correlation coefficient of over 0.8 with an identified gene were selected for integrative pathway view analysis.

Additional files

Additional file 1: Figure S1. Definition of the positive intensity threshold. (PPT 279 kb)

Additional file 2: Table S1. Detection of the HERV transcriptome in PBMCs. (XLSX 145 kb)

Additional file 3: Figure S2. Specialisation of LTR features on the whole dataset. (PPT 212 kb)

Additional file 4: Figure S3. Genomic environment of functional and silent LTRs. (DOC 507 kb)

Additional file 5: Figure S4. TNF-α and IL-10 protein assay and mRNA quantitation in PBMCs following LPS stimulations. (PPT 172 kb)

Additional file 6: Table S2. Differential expression of HERV-V3 chip repertoires induced by stimulation state changes in endotoxin tolerance model. (XLSX 41775 kb)

Additional file 7: Figure S5. Selection, design and quality criteria for the design of locus specific qPCR systems, illustrated with the 121601901-HERV0116uL locus, and PCR systems obtained. (PPT 766 kb)

Additional file 8: Figure S6. RT-qPCR validation of 23 microarray-based identified HERV/MaLR elements differentially expressed following LPS stimulations. (PPT 1309 kb)

Additional file 9: Figure S7. Strategy used to integrate 62 HERVs/MaLRs and 26 genes in 11 canonical immune pathways, global landscape resulting from this analysis, and identification of PTGS2 and FLT1 associated HERVs/MaLRs at the crossroads of "LXR/RXR activation" and "NF-kB signalling" pathways. (PPT 1323 kb)

Additional file 10: Table S3. HERVs or MaLRs with the same expression profiles as genes within pathways. (XLS 53 kb)

Additional file 11: Supplementary Methods. (DOC 55 kb)

Abbreviations

DEG: Differentially expressed gene; DEL: Differentially expressed locus; ET: Endotoxin tolerance; HERV : Human endogenous retrovirus; IFN: Interferon; IPA: Ingenuity pathway analysis; LPS: Lipopolysaccharide; LTR: Long terminal repeat; MaLR: Mammalian apparent LTR-retrotransposons; MSRV: Multiple sclerosis associated retrovirus; NS: Non stimulated; pA: Polyadenylation; PBMC: Peripheral blood mononuclear cell; Pr: Promoter

Acknowledgments
The authors would like to express our thanks to Nadia Gaci and Laurence Ganée for their support in the primer design, and Jérémie Becker and Valérie Cheynet for their kind advice. Many thanks to Sophie Ablott who edited and proofread the English in this paper.

Funding
This work was supported by bioMerieux SA. MM and OT were supported by doctoral grants from bioMerieux. In addition, OT was supported by Association Nationale de la Recherche et de la Technologie (ANRT). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials
The microarray expression data have been filed on the NCBI Gene Expression Omnibus and are accessible via GEO accession number GSE108239.

Authors' contributions
All authors were involved in the analysis and interpretation of data as well as drafting the manuscript or revising it critically for important intellectual content. FM, GM, FV, JT, AP and KBP made substantial contributions to the conception and design of the study, MM, OT and FM designed the experiments, MM performed the endotoxin tolerance model. MM, GO and EC designed and performed the microarray experiments, OT and MM performed the data analyses. MM, OT, JT and FM performed data interpretations, MM, AG and PF performed the primer pairs design, screening and validation, RT-qPCR experiments and statistical analysis, FM, MM, OT, MN and JNV conceived and performed genomic environment and transcription factors enrichments analysis, MM, OT and FM wrote the paper. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Citrated pouches or heparinised tubes blood were obtained from EFS (Etablissement Français du Sang) and used immediately. According to EFS standardised procedures for blood donation and to provisions of the articles R.1243–49 and following ones of the French Public Health Code, a written non-opposition to the use of his donation for research purposes was obtained from healthy volunteers. The blood donors' personal data were anonymised before blood transfer to our research lab. We obtained the favourable notice of the Local Ethical Committee (Comité de Protection des Personnes Sud-Est II, Bâtiment Pinel, 59 Boulevard Pinel, 69,500 Bron) and the acceptance of the Ministère de la Recherche (declaration DC-2008-64) for handling and conservation of these samples.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Joint research unit, Hospice Civils de Lyon, bioMerieux, Centre Hospitalier Lyon Sud, 165 Chemin du Grand Revoyet, 69310 Pierre-Benite, France. [2]EA 7426 Pathophysiology of Injury-induced Immunosuppression, University of Lyon1-Hospices Civils de Lyon-bioMérieux, Hôpital Edouard Herriot, 5 Place d'Arsonval, 69437 Lyon, Cedex 3, France. [3]Institut de Genomique Fonctionnelle de Lyon, Univ Lyon, CNRS UMR 5242, Ecole Normale Superieure de Lyon, Universite Claude Bernard Lyon, 1, 46 allee d'Italie, F-69364 Lyon, France. [4]Hospices Civils de Lyon, Immunology Laboratory, Groupement Hospitalier Edouard Herriot, Lyon, France. [5]Hospices Civils de Lyon, Department of Anaesthesiology and Critical Care Medicine, Groupement Hospitalier Edouard Herriot, Université Claude Bernard Lyon 1, Lyon, France.

Received: 18 January 2018 Accepted: 27 June 2018
Published online: 05 July 2018

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.
2. Villesen P, Aagaard L, Wiuf C, Pedersen FS. Identification of endogenous retroviral reading frames in the human genome. Retrovirology. 2004;1:32.
3. Bolze PA, Mommert M, Mallet F. Contribution of Syncytins and other endogenous retroviral envelopes to human placenta pathologies. Prog Mol Biol Transl Sci. 2017;145:111–62.
4. Bannert N, Kurth R. Retroelements and the human genome: new perspectives on an old relation. Proc Natl Acad Sci U S A. 2004;101(Suppl 2):14572–9.
5. Perot P, Bolze P-A, Mallet F. From viruses to genes: syncytins. In: Witzany G, editor. Viruses, Essential Agents of Life. Netherlands: Springer; 2012. p. 325–61.
6. Vargiu L, Rodriguez-Tome P, Sperber GO, Cadeddu M, Grandi N, Blikstad V, Tramontano E, Blomberg J. Classification and characterization of human endogenous retroviruses; mosaic forms are common. Retrovirology. 2016;13:7.
7. Sperber GO, Airola T, Jern P, Blomberg J. Automated recognition of retroviral sequences in genomic data–RetroTector. Nucleic Acids Res. 2007;35(15):4964–76.
8. Mager DL, Medstrand P. Retroviral Repeat Sequences. Encyclopedia of Life Sciences, 2005, John Wiley & Sons, Ltd. www.els.net.
9. Gifford R, Tristem M. The evolution, distribution and diversity of endogenous retroviruses. Virus Genes. 2003;26(3):291–315.
10. Jern P, Sperber GO, Blomberg J. Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. Retrovirology. 2005;2:50.
11. Hughes JF, Coffin JM. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat Genet. 2001;29(4):487–9.
12. Hurst TP, Magiorkinis G. Epigenetic control of human endogenous retrovirus expression: focus on regulation of long-terminal repeats (LTRs). Viruses. 2017;9(6)
13. Balada E, Vilardell-Tarres M, Ordi-Ros J. Implication of human endogenous retroviruses in the development of autoimmune diseases. Int Rev Immunol. 2010;29(4):351–70.
14. Madeira A, Burgelin I, Perron H, Curtin F, Lang AB, Faucard R. MSRV envelope protein is a potent, endogenous and pathogenic agonist of human toll-like receptor 4: relevance of GNbAC1 in multiple sclerosis treatment. J Neuroimmunol. 2016;291:29–38.
15. Antony JM, van Marle G, Opii W, Butterfield DA, Mallet F, Yong VW, Wallace JL, Deacon RM, Warren K, Power C. Human endogenous retrovirus glycoprotein-mediated induction of redox reactants causes oligodendrocyte death and demyelination. Nat Neurosci. 2004;7(10):1088–95.
16. Katoh I, Kurata S. Association of endogenous retroviruses and long terminal repeats with human disorders. Front Oncol. 2013;3:234.
17. Cohen CJ, Lock WM, Mager DL. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. Gene. 2009;448(2):105–14.
18. Long Q, Bengra C, Li C, Kutlar F, Tuan D. A long terminal repeat of the human endogenous retrovirus ERV-9 is located in the 5′ boundary area of the human beta-globin locus control region. Genomics. 1998;54(3):542–55.
19. Schulte AM, Lai S, Kurtz A, Czubayko F, Riegel AT, Wellstein A. 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. 1996;93(25):14759–64.
20. Ting CN, Rosenberg MP, Snow CM, Samuelson LC, Meisler MH. Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. Genes Dev. 1992;6(8):1457–65.
21. Blond JL, Lavillette D, Cheynet V, Bouton O, Oriol G, Chapel-Fernandes S, Mandrand B, Mallet F, Cosset FL. 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. 2000;74(7):3321–9.
22. Mangeney M, de Parseval N, Thomas G, Heidmann T. The full-length envelope of an HERV-H human endogenous retrovirus has immunosuppressive properties. The Journal of general virology. 2001;82(Pt 10):2515–8.
23. Mangeney M, Renard M, Schlecht-Louf G, Bouallaga I, Heidmann O, Letzelter C, Richaud A, Ducos B, Heidmann T. Placental syncytins: genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins. Proc Natl Acad Sci U S A. 2007;104(51):20534–9.

24. Magin C, Lower R, Lower J. cORF and RcRE, the rev/rex and RRE/RxRE homologues of the human endogenous retrovirus family HTDV/HERV-K. J Virol. 1999;73(11):9496–507.
25. Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, Martin L, Ware CB, Blish CA, Chang HY, et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature. 2015;522(7555):221–5.
26. Santoni FA, Guerra J, Luban J. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. Retrovirology. 2012;9:111.
27. Fort A, Hashimoto K, Yamada D, Salimullah M, Keya CA, Saxena A, Bonetti A, Voineagu I, Bertin N, Kratz A, et al. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nat Genet. 2014;46(6):558–66.
28. Gimenez J, Montgiraud C, Pichon JP, Bonnaud B, Arsac M, Ruel K, Bouton O, Mallet F. Custom human endogenous retroviruses dedicated microarray identifies self-induced HERV-W family elements reactivated in testicular cancer upon methylation control. Nucleic Acids Res. 2010;38(7):2229–46.
29. Perot P, Mugnier N, Montgiraud C, Gimenez J, Jaillard M, Bonnaud B, Mallet F. Microarray-based sketches of the HERV transcriptome landscape. PLoS One. 2012;7(6):e40194.
30. Becker J, Perot P, Cheynet V, Oriol G, Mugnier N, Mommert M, Tabone O, Textoris J, Veyrieras JB, Mallet F. A comprehensive hybridization model allows whole HERV transcriptome profiling using high density microarray. BMC Genomics. 2017;18(1):286.
31. Mavrommatis B, Young GR, Kassiotis G. Counterpoise between the microbiome, host immune activation and pathology. Curr Opin Immunol. 2013;25(4):456–62.
32. Hurst TP, Magiorkinis G. Activation of the innate immune response by endogenous retroviruses. The Journal of general virology. 2015;96(Pt 6):1207–18.
33. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016;351(6277):1083–7.
34. Andersson ML, Medstrand P, Yin H, Blomberg J. Differential expression of human endogenous retroviral sequences similar to mouse mammary tumor virus in normal peripheral blood mononuclear cells. AIDS Res Hum Retrovir. 1996;12(9):833–40.
35. Krieg AM, Gourley MF, Klinman DM, Perl A, Steinberg AD. Heterogeneous expression and coordinate regulation of endogenous retroviral sequences in human peripheral blood mononuclear cells. AIDS Res Hum Retrovir. 1992;8(12):1991–8.
36. Balestrieri E, Pica F, Matteucci C, Zenobi R, Sorrentino R, Argaw-Denboba A, Cipriani C, Bucci I, Sinibaldi-Vallebona P. Transcriptional activity of human endogenous retroviruses in human peripheral blood mononuclear cells. Biomed Res Int. 2015;2015:164529.
37. Wallace TA, Downey RF, Seufert CJ, Schetter A, Dorsey TH, Johnson CA, Goldman R, Loffredo CA, Yan P, Sullivan FJ, et al. Elevated HERV-K mRNA expression in PBMC is associated with a prostate cancer diagnosis particularly in older men and smokers. Carcinogenesis. 2014;35(9):2074–83.
38. Garcia-Montojo M, Dominguez-Mozo M, Arias-Leal A, Garcia-Martinez A, De las Heras V, Casanova I, Faucard R, Gehin N, Madeira A, Arroyo R, et al. The DNA copy number of human endogenous retrovirus-W (MSRV-type) is increased in multiple sclerosis patients and is influenced by gender and disease severity. PLoS One. 2013;8(1):e53623.
39. Mameli G, Poddighe L, Mei A, Uleri E, Sotgiu S, Serra C, Manetti R, Dolei A. Expression and activation by Epstein Barr virus of human endogenous retroviruses-W in blood cells and astrocytes: inference for multiple sclerosis. PLoS One. 2012;7(9):e44991.
40. Bhardwaj N, Maldarelli F, Mellors J, Coffin JM. HIV-1 infection leads to increased transcription of human endogenous retrovirus HERV-K (HML-2) proviruses in vivo but not to increased virion production. J Virol. 2014; 88(19):11108–20.
41. Johnston JB, Silva C, Holden J, Warren KG, Clark AW, Power C. Monocyte activation and differentiation augment human endogenous retrovirus expression: implications for inflammatory brain diseases. Ann Neurol. 2001;50(4):434–42.
42. Young GR, Mavrommatis B, Kassiotis G. Microarray analysis reveals global modulation of endogenous retroelement transcription by microbes. Retrovirology. 2014;11:59.
43. Larsson E, Venables P, Andersson AC, Fan W, Rigby S, Botling J, Oberg F, Cohen M, Nilsson K. 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. 1997;11(Suppl 3):142–4.
44. Serra C, Mameli G, Arru G, Sotgiu S, Rosati G, Dolei A. In vitro modulation of the multiple sclerosis (MS)-associated retrovirus by cytokines: implications for MS pathogenesis. J Neurovirol. 2003;9(6):637–43.
45. Brudek T, Christensen T, Aagaard L, Petersen T, Hansen HJ, Moller-Larsen A. B cells and monocytes from patients with active multiple sclerosis exhibit increased surface expression of both HERV-H Env and HERV-W Env, accompanied by increased seroreactivity. Retrovirology. 2009;6:104.
46. Maliniemi P, Vincendeau M, Mayer J, Frank O, Hahtola S, Karenko L, Carlsson E, Mallet F, Seifarth W, Leib-Mosch C, et al. Expression of human endogenous retrovirus-w including syncytin-1 in cutaneous T-cell lymphoma. PLoS One. 2013;8(10):e76281.
47. Tolosa JM, Schjenken JE, Clifton VL, Vargas A, Barbeau B, Lowry P, Maiti K, Smith R. The endogenous retroviral envelope protein syncytin-1 inhibits LPS/PHA-stimulated cytokine responses in human blood and is sorted into placental exosomes. Placenta. 2012;33(11):933–41.
48. Holder BS, Tower CL, Forbes K, Mulla MJ, Aplin JD, Abrahams VM. Immune cell activation by trophoblast-derived microvesicles is mediated by syncytin 1. Immunology. 2012;136(2):184–91.
49. Zeng M, Hu Z, Shi X, Li X, Zhan X, Li XD, Wang J, Choi JH, Wang KW, Purrington T, et al. MAVS, cGAS, and endogenous retroviruses in T-independent B cell responses. Science. 2014;346(6216):1486–92.
50. Cavaillon JM, Adib-Conquy M. Bench-to-bedside review: endotoxin tolerance as a model of leukocyte reprogramming in sepsis. Crit Care. 2006;10(5):233.
51. Allantaz-Frager F, Turrel-Davin F, Venet F, Monnin C, De Saint Jean A, Barbalat V, Cerrato E, Pachot A, Lepape A, Monneret G. Identification of biomarkers of response to IFNg during endotoxin tolerance: application to septic shock. PLoS One. 2013;8(7):e68218.
52. Turrel-Davin F, Venet F, Monnin C, Barbalat V, Cerrato E, Pachot A, Lepape A, Alberti-Segui C, Monneret G. mRNA-based approach to monitor recombinant gamma-interferon restoration of LPS-induced endotoxin tolerance. Crit Care. 2011;15(5):R252.
53. Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, Smit AF, Finn RD. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res. 2013;41(Database issue):D70–82.
54. Chistiakov DA, Bobryshev YV, Nikiforov NG, Elizova NV, Sobenin IA, Orekhov AN. Macrophage phenotypic plasticity in atherosclerosis: the associated features and the peculiarities of the expression of inflammatory genes. Int J Cardiol. 2015;184:436–45.
55. Conley AB, Piriyapongsa J, Jordan IK. Retroviral promoters in the human genome. Bioinformatics. 2008;24(14):1563–7.
56. Oja M, Peltonen J, Blomberg J, Kaski S. Methods for estimating human endogenous retrovirus activities from EST databases. BMC bioinformatics. 2007;8(Suppl 2):S11.
57. Prudencio M, Gonzales PK, Cook CN, Gendron TF, Daughrity LM, Song Y, Ebbert MTW, van Blitterswijk M, Zhang YJ, Jansen-West K, et al. Repetitive element transcripts are elevated in the brain of C9orf72 ALS/FTLD patients. Hum Mol Genet. 2017;26(17):3421–31.
58. Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. Nature. 1980;284(5757):601–3.
59. Maksakova IA, Mager DL, Reiss D. Keeping active endogenous retroviral-like elements in check: the epigenetic perspective. Cellular and molecular life sciences : CMLS. 2008;65(21):3329–47.
60. Leung DC, Lorincz MC. Silencing of endogenous retroviruses: when and why do histone marks predominate? Trends Biochem Sci. 2012;37(4):127–33.
61. Mameli G, Astone V, Arru G, Marconi S, Lovato L, Serra C, Sotgiu S, Bonetti B, Dolei A. Brains and peripheral blood mononuclear cells of multiple sclerosis (MS) patients hyperexpress MS-associated retrovirus/HERV-W endogenous retrovirus, but not human herpesvirus 6. The Journal of general virology. 2007;88(Pt 1):264–74.
62. Kowalczyk MJ, Danczak-Pazdrowska A, Szramka-Pawlak B, Zaba R, Osmola-Mankowska A, Silny W. Human endogenous retroviruses and chosen disease parameters in morphea. Postepy dermatologii i alergologii. 2017;34(1):47–51.
63. Bergallo M, Loiacono E, Galliano I, Montanari P, Peruzzi L, Tovo PA, Coppo R. HERV-K and W expression in peripheral mononuclear cells of children with Henoch-Schonlein purpura and relation with TLRs activation. Minerva Pediatr. 2017;
64. Contreras-Galindo R, Kaplan MH, He S, Contreras-Galindo AC, Gonzalez-Hernandez MJ, Kappes F, Dube D, Chan SM, Robinson D, Meng F, et al. HIV infection reveals widespread expansion of novel centromeric human endogenous retroviruses. Genome Res. 2013;23(9):1505–13.

65. Lee YJ, Jeong BH, Park JB, Kwon HJ, Kim YS, Kwak IS. The prevalence of human endogenous retroviruses in the plasma of major burn patients. Burns : journal of the International Society for Burn Injuries. 2013;39(6):1200–5.
66. Lee KH, Rah H, Green T, Lee YK, Lim D, Nemzek J, Wahl W, Greenhalgh D, Cho K. Divergent and dynamic activity of endogenous retroviruses in burn patients and their inflammatory potential. Exp Mol Pathol. 2014;96(2):178–87.
67. Kwon DN, Lee YK, Greenhalgh DG, Cho K. Lipopolysaccharide stress induces cell-type specific production of murine leukemia virus type-endogenous retroviral virions in primary lymphoid cells. The Journal of general virology. 2011;92(Pt 2):292–300.
68. Biswas SK, Lopez-Collazo E. Endotoxin tolerance: new mechanisms, molecules and clinical significance. Trends Immunol. 2009;30(10):475–87.
69. Seeley JJ, Ghosh S. Molecular mechanisms of innate memory and tolerance to LPS. J Leukoc Biol. 2017;101(1):107–19.
70. Lu YC, Yeh WC, Ohashi PS. LPS/TLR4 signal transduction pathway. Cytokine. 2008;42(2):145–51.
71. Barrett LW, Fletcher S, Wilton SD. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cellular and molecular life sciences : CMLS. 2012;69(21):3613–34.
72. Brkic Z, Maria NI, van Helden-Meeuwsen CG, van de Merwe JP, van Daele PL, Dalm VA, Wildenberg ME, Beumer W, Drexhage HA, Versnel MA. Prevalence of interferon type I signature in CD14 monocytes of patients with Sjogren's syndrome and association with disease activity and BAFF gene expression. Ann Rheum Dis. 2013;72(5):728–35.
73. Power D, Santoso N, Dieringer M, Yu J, Huang H, Simpson S, Seth I, Miao H, Zhu J. IFI44 suppresses HIV-1 LTR promoter activity and facilitates its latency. Virology. 2015;481:142–50.
74. Carneiro MW, Fukutani KF, Andrade BB, Curvelo RP, Cristal JR, Carvalho AM, Barral A, Van Weyenbergh J, Barral-Netto M, de Oliveira CI. Gene expression profile of high IFN-gamma producers stimulated with Leishmania braziliensis identifies genes associated with cutaneous Leishmaniasis. PLoS Negl Trop Dis. 2016;10(11):e0005116.
75. Novakovic B, Habibi E, Wang SY, Arts RJ, Davar R, Megchelenbrink W, Kim B, Kuznetsova T, Kox M, Zwaag J, et al. beta-glucan reverses the epigenetic state of LPS-induced immunological tolerance. Cell. 2016;167(5):1354–68. e1314
76. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–80.
77. Wang X, Liu Z, Wang P, Li S, Zeng J, Tu X, Yan Q, Xiao Z, Pan M, Zhu F. Syncytin-1, an endogenous retroviral protein, triggers the activation of CRP via TLR3 signal cascade in glial cells. Brain Behav Immun. 2017;67:324–34.
78. Weinberger L, Ayyash M, Novershtern N, Hanna JH. Dynamic stem cell states: naive to primed pluripotency in rodents and humans. Nat Rev Mol Cell Biol. 2016;17(3):155–69.
79. Balada E, Ordi-Ros J, Vilardell-Tarres M. Molecular mechanisms mediated by human endogenous retroviruses (HERVs) in autoimmunity. Rev Med Virol. 2009;19(5):273–86.
80. Crow YJ, Manel N. Aicardi-Goutieres syndrome and the type I interferonopathies. Nat Rev Immunol. 2015;15(7):429–40.
81. Picard C, Belot A. Does type-I interferon drive systemic autoimmunity? Autoimmun Rev. 2017;16(9):897–902.
82. Bergallo M, Galliano I, Montanari P, Gambarino S, Mareschi K, Ferro F, Fagioli F, Tovo PA, Ravanini P. CMV induces HERV-K and HERV-W expression in kidney transplant recipients. Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology. 2015;68:28–31.
83. Hotchkiss RS, Monneret G, Payen D. Sepsis-induced immunosuppression: from cellular dysfunctions to immunotherapy. Nat Rev Immunol. 2013;13(12):862–74.
84. Jiang L, Wang Y, Zhu D, Xue Z, Mao H. Alteration of histone H3 lysine 9 dimethylation in peripheral white blood cells of septic patients with trauma and cancer. Mol Med Rep. 2016;14(6):5467–74.
85. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:Article3.
86. Hochberg YBaY. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 1995. 1995;57(1):289–300.

3.2.3 MODULATION CHEZ DES VOLONTAIRES SAINS STIMULES IN VIVO PAR LPS, APPROCHE PAR RNASEQ

L'étude précédente a montré que des HERV étaient modulés dans un modèle *in vitro* de PBMC mimant la réponse de l'hôte suivant un choc septique. A partir d'un jeu de données publiques, nous avons cherché à savoir si on peut détecter une modulation HERV sur des PBMCs d'individus sains, stimulés par LPS, *in vivo.* Le but étant aussi de valider un pipeline RNAseq pour l'analyse du transcriptome HERV. Cette étude a été réalisée en collaboration avec Maria-Paola Pisano, étudiante italienne en thèse de biologie à l'université de Cagliari. Les analyses sont toujours en cours et je montre ici les résultats préliminaires.

3.2.3.1 MATERIEL

Le jeu de données publique utilisé est un jeu de données RNAseq (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87290). Il est composé de PBMCs provenant de 15 volontaires sains, prélevés avant et après stimulation par LPS, in vivo (1ng/kg). Les données ont été générées à l'aide d'un séquenceur Illumina HiSeq 2000, les reads (paired-end) ont une longueur de 100.

3.2.3.2 PIPELINE D'ANALYSE

Comme expliqué précédemment, cibler ou détecter de manière spécifique des HERV est encore aujourd'hui un challenge technique. Il en est de même pour le RNAseq. De nombreux pipelines d'analyse existent pour le RNAseq aujourd'hui, mais la plupart ne prennent pas en compte les éléments répétés du génome. Pour cela, la taille des reads est importante, nous avons cherché dans les données publiées un jeu de données adapté à notre question biologique et dont les caractéristiques techniques étaient favorables à l'analyse d'éléments répétés. Nous avons ensuite optimisé un pipeline d'analyse en assemblant les outils les plus adaptés à chaque étape pour cibler l'analyse d'éléments répétés (Figure 3-10).

Figure 3-10 : Pipeline d'analyse RNAseq. Le contrôle qualité a été réalisé avec FastQC (+Trim Galore), le mapping avec HISAT2 (98,4% des reads ont été correctement mappés sur hervgdb4), le compte des reads avec HTSeq-count et le différentiel d'expression avec DESeq2.

3.2.3.3 RESULTATS

3.2.3.3.1 DESCRIPTION DU TRANSCRIPTOME

Après les étapes de pré-processing, nous avons filtré le jeu de données en enlevant les loci non exprimés. C'est-à-dire les loci HERV qui ne comptaient pas de reads dans plus de 10 échantillons ont été retirés de l'analyse. De cette manière, nous avons retenu 7% des loci de *hervgdb4* pour l'analyse, ce qui est proche du pourcentage de transcriptome HERV observé à la fois dans des PBMCs (modèle ET) et en sang total (cohorte de patients en choc septique, IS) avec la puce HERV-V3. Nous avons ensuite décrit le transcriptome HERV en terme d'éléments actifs (exprimés) par groupe de HERV prototypes ou types de HERV, de la même manière que pour le jeu de données ET (Figure 3-11, Figure 1 de l'article section 3.2.2.4). Bien que les pourcentages d'éléments exprimés soient similaires, on observe des différences selon les groupes de HERV. La proportion de HERV prototypes exprimés est inférieure, 5,9 % contre 9,4% dans ET, et celle d'éléments Dfam (DfamHERV et DfamMALR) est supérieure, 6,9% d'élément exprimés contre 5,5% dans ET. On remarque qu'une proportion moins importante de Gamma-rétrovirus et de Spuma-Epsilon like sont exprimés par rapport aux Betaretrovirus dans ce jeu de données que dans le modèle ET. Par exemple, la famille HERV-H, une des plus exprimée dans le modèle ET (940 probesets, plus de 16% de la famille) possède ici nettement moins d'éléments actifs (672, 7%). Au contraire, le groupe HML-2 possède plus d'éléments exprimés (179, 19%) que dans le modèle ET. Ces différences peuvent s'expliquer par l'utilisation d'outils différents et aussi par des conditions expérimentales différentes.

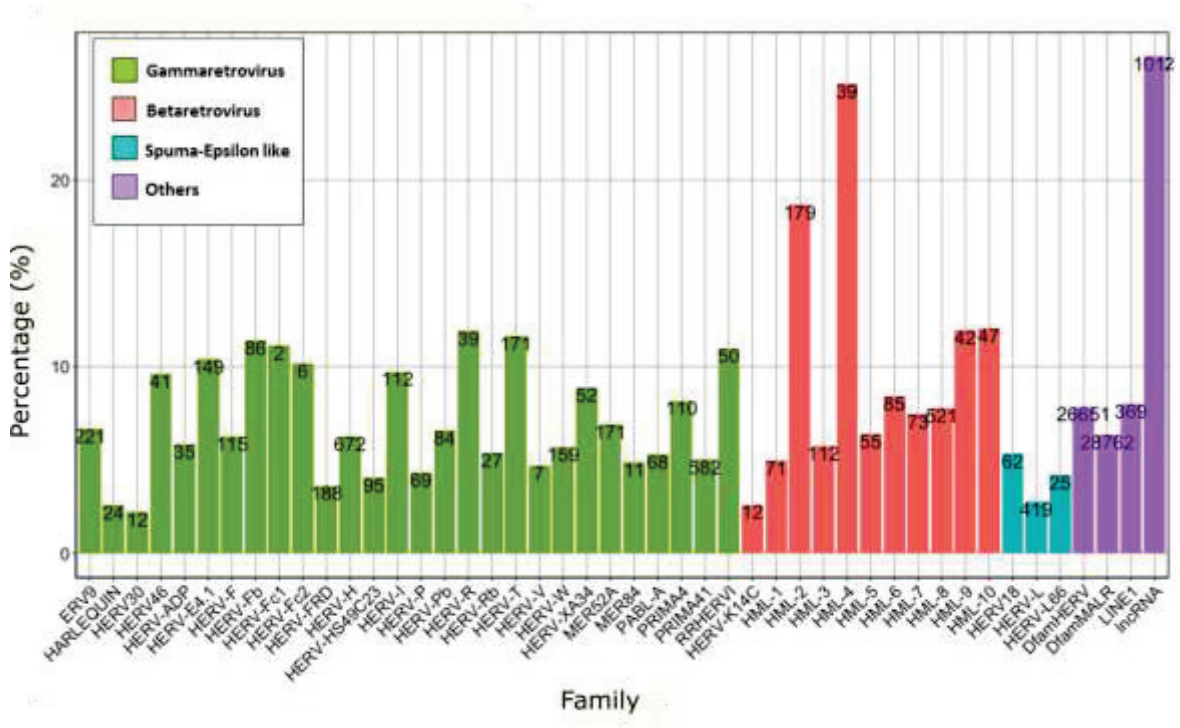


Figure 3-11 : Transcriptome par type de HERV. Comptes et proportions des loci ayant un signal non-nul dans au moins 13 échantillons, groupés par type de HERV. L'axe des x représente chaque type de HERV, chaque type étant coloré par classe de rétrovirus. L'axe des y représente le pourcentage de loci considéré actif, par rapport au nombre total de loci de la base de données hervgdb4. Les chiffres représente le nombre d'éléments actifs correspondant.

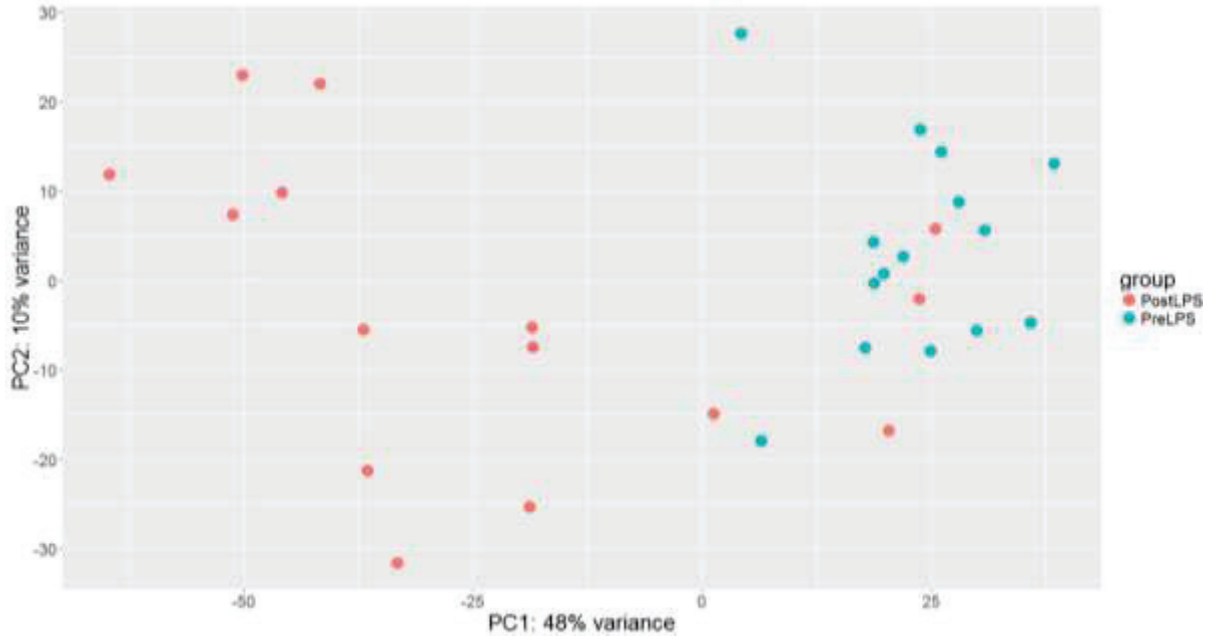


Figure 3-12 : Analyse en composante principale du jeu de données filtré. Chaque point représente un échantillon. La condition (avec/sans stimulation par LPS) est codée en rouge ou bleu, respectivement.

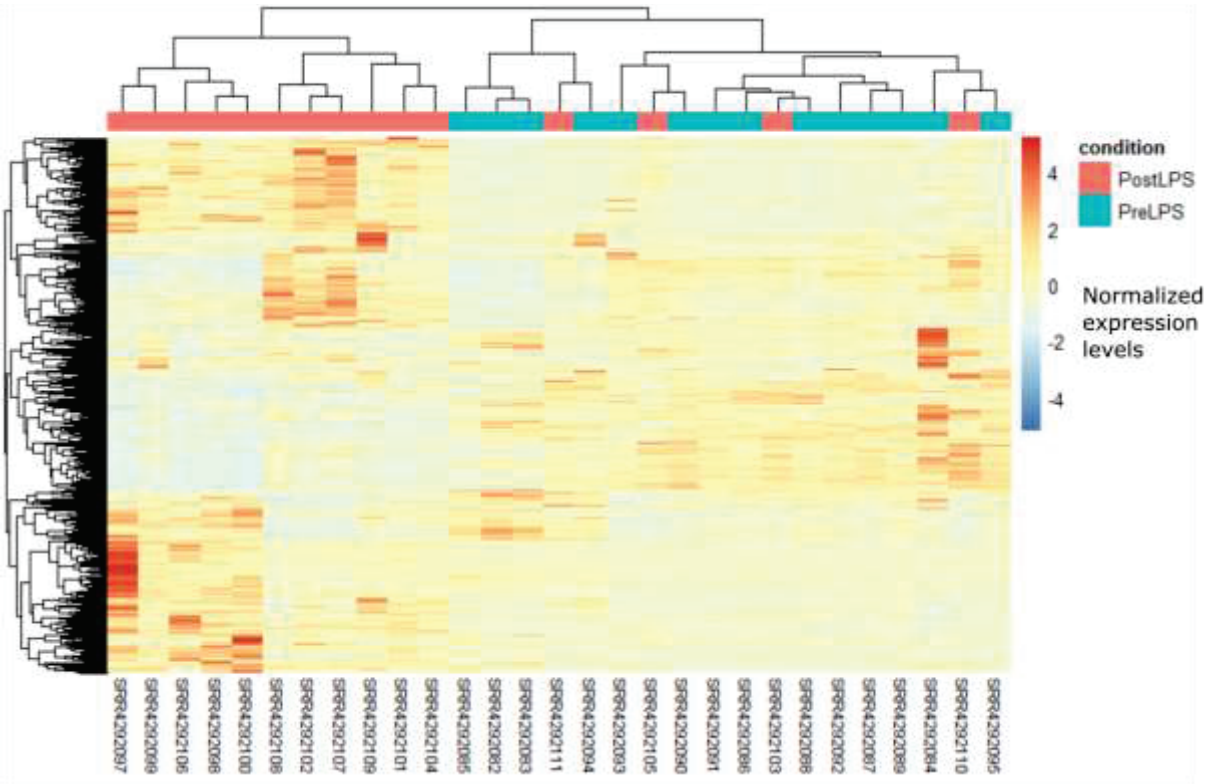


Figure 3-13 : Clustering Hiérarchique des probesets (1%) les plus variants entre les échantillons. La heatmap représente le niveau d'expression de chaque locus pour chaque échantillon par un code couleur allant du bleu (pas exprimé) au rouge (fortement exprimé). Chaque ligne correspond à un locus HERV, chaque colonne à un échantillon. Les loci ont été regroupés par une méthode de clustering hiérarchique (méthode d'agrégation « complète », distance Euclidienne). Les échantillons ont été regroupés par une méthode de clustering hiérarchique (méthode d'agrégation « complète », distance de corrélation de Pearson).

Nous avons d'abord procédé à une approche non supervisée en réalisant une analyse en composante principale (ACP) et un clustering hiérarchique sur les loci les plus variant en terme d'expression. Sur l'ACP (Figure 3-12), on voit que l'axe 1 (48% de la variabilité totale) sépare bien les échantillons en fonction de l'état stimulé ou non. De la même manière, le clustering hiérarchique sur les 1% des loci les plus variants entre échantillons (Figure 3-13) sépare clairement 2 groupes, en fonction de la stimulation. Cela souligne qu'il existe un signal de modulation du transcriptome HERV après stimulation par LPS.

3.2.3.3.2 DIFFERENTIEL D'EXPRESSION

Nous avons ensuite analysé l'expression différentielle des HERV avant et après stimulation au LPS, en utilisant DESeq2. Les résultats sont illustrés par le volcano plot de la Figure 3-14. Plus de 5000 loci sont modulés par une stimulation au LPS, donc 3583 voient leur

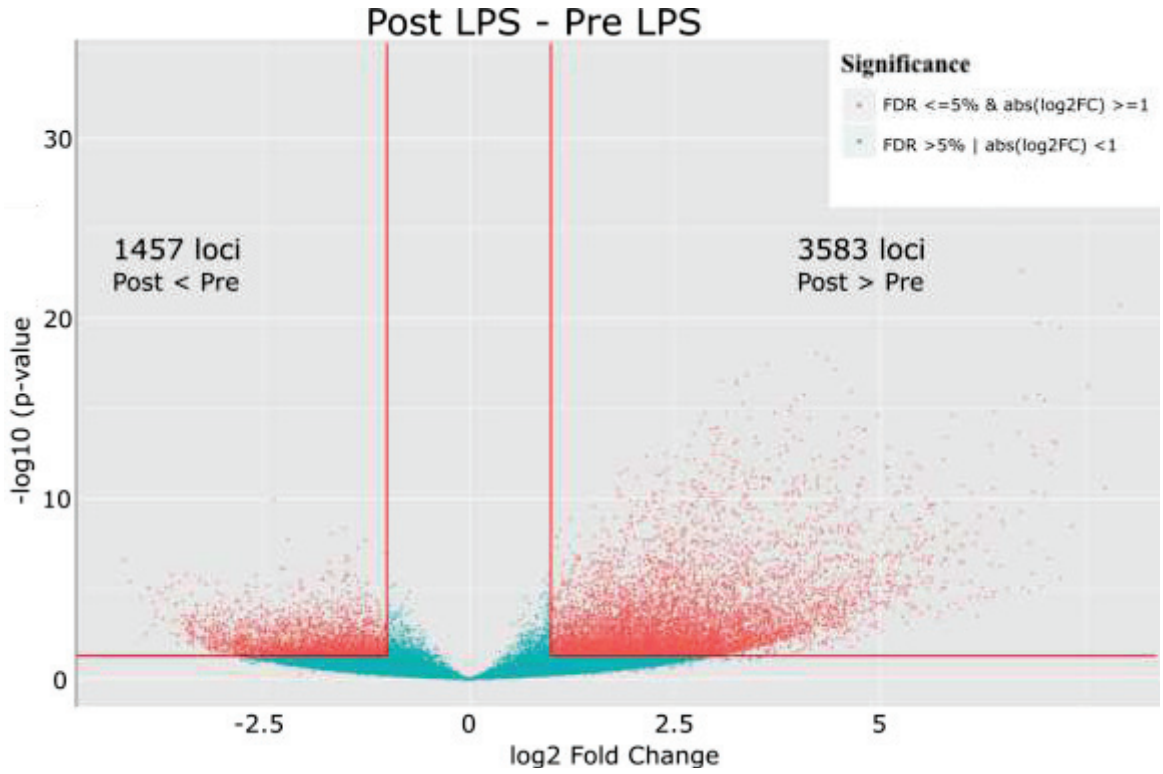


Figure 3-14 : HERV différentiellement exprimés après stimulation au LPS. Volcano plot du modèle testant le différentiel d'expression entre échantillons stimulés ou non par LPS. L'axe des x représente le log2 Fold Change (Stimulé / Non stimulé), l'axe des y représente la significativité statistique, après correction pour le multi-testing par la méthode de Benjamini Hochberg (-log10(p_value)). Les points rouges représentent les loci différentiellement exprimés (valeur absolue du log2 FC supérieure à 1, p-value ajustée inférieure à 0,05).

expression augmenter après stimulation (log2(FC) absolu > 1 et FDR < à 0,05). De manière intéressante, les 230 loci les plus différentiellement exprimés sont tous up-modulés par la stimulation LPS, montrant globalement une réponse transcriptomique HERV à la stimulation

LPS. Il n'y a pas de groupe de HERV spécifiquement enrichi dans les éléments modulés, bien que 22% (n=19) des éléments exprimés du groupe HERV-Fb sont modulés. Pour finir, parmi les 7 HERV les plus différentiellement exprimés, on retrouve des HERV proche des gènes IL1RN, IL1R2, IL18R1 et TNFAIP6, tous des gènes situés sur le chromosome 2 et jouant un rôle dans la réponse immunitaire. TNFAIP6 et IL18R1 ayant un rôle sur la réponse inflammatoire, les autres régulant cette réponse. D'autre part, les HERV proches de IL18R1 ont également été retrouvés modulés par puce dans l'analyse de patients en choc septique, notamment entre patients avec HLA-DR normal et diminué. A partir de tous les HERV modulés par la stimulation LPS, nous avons réalisé une analyse de réseaux fonctionnels (avec Ingenuity Pathway Analysis). Pour cela, nous avons récupéré le gène le plus proche de

chacun des HERV modulés (situés à +/- 10kb), et nous avons recherché un enrichissement fonctionnel au sein de cette liste de gènes. Les 4 réseaux qui sont les plus enrichis sont la voie de réponse à IL10, de réponse à l'IL6, de réponse à TREM1 et la voie de réponse aux Toll-like Receptor (TLR). Les éléments HERV modulés se trouvent donc dans l'environnement proche de gènes directement impliqués dans la réponse immunitaire. La plupart sont probablement embarqués dans l'unité transcriptionnelle des gènes mais on ne peut pas exclure que certain jouent un rôle direct sur l'expression de ces gènes, par exemple en jouant un rôle promoteur, enhancer ou de terminaison de la transcription.

Pour conclure, cette analyse préliminaire a permis d'une part de montrer qu'il est possible d'étudier le transcriptome HERV à partir d'un pipeline RNAseq standard, en exploitant le niveau fin d'annotation de la base *hervgdb4*; d'autre part, nous confirmons une expression de HERV dans les PBMCs, sur des volontaires stimulés ou non au LPS. Les HERV modulés entre ces conditions semblent être régulés par les mêmes contrôles d'expression que pour les gènes. Pour être capable de déterminer si une LTR joue un rôle *en cis* sur son gène voisin, par exemple un rôle promoteur, il serait intéressant de développer une stratégie qui permettrait de cibler spécifiquement les transcrits hybrides LTR-gènes.

3.2.4 MODULATION APRES UN CHOC SEPTIQUE

3.2.4.1 RESUME DE L'ETUDE

Nous avons montré qu'il existe de la modulation d'expression HERV dans le sang total de patients après une agression inflammatoire grave mais avec un outil limité, montré qu'il existe de la modulation d'expression de HERV dans un modèle *ex vivo* mimant la réponse immunitaire innée après un choc septique avec la puce HERV-V3, et nous avons mis en évidence que cette modulation était également observable *in vivo* sur des volontaires sains exposés au LPS. Nous avons ensuite étudié l'expression des HERV, sur des patients en choc septique, avec l'outil le plus adapté à ce jour pour explorer le transcriptome HERV, la puce HERV-V3. Deux cohortes de patients en choc septique sont exploitées dans cette étude. Une cohorte composée de 20 patients, stratifiés par l'expression du marqueur mHLA-DR à J3. Les prélèvements à J1 et J3 après admission en réanimation pour choc septique ont été analysés. Cette cohorte bien définie sert comme cohorte de découverte. La seconde cohorte est composée de 102 patients en choc septique, qui ne sont pas stratifiés et pour laquelle des échantillons à J1, J3 et J6 ont été analysés. Cette cohorte sert de cohorte de validation. Dans cette étude nous décrivons, d'une part, l'expression et la modulation du transcriptome HERV dans le sang total dans la phase initiale d'un choc septique. D'autre part, nous montrons qu'il existe une modulation HERV selon le statut immunitaire des patients, dont le mHLA-DR est le proxy. Enfin, les HERV modulés entre les patients à niveau de mHLA-DR variable à J3 ont été utilisés en tant que signature pronostique. Sur une cohorte de validation de patients en choc septique, non stratifiés sur le statut immunitaire, nous avons montré que cette signature permet de distinguer 2 groupes de patients, séparés selon leur niveau de sévérité. Afin de garder la meilleure balance entre capacité de la signature à séparer les groupes de patients et taille de la signature, nous avons procédé à une réduction de la signature via une approche de classification Random Forest. Ceci nous a permis d'identifier une signature pronostique de 10 HERV.

3.2.4.2 ARTICLE

MODULATION OF LTR-RETROTRANSPOSONS EXPRESSION IN MHLA-DR STRATIFIED SEPTIC SHOCK PATIENTS: A PILOT STUDY

Marine Mommert*, Olivier Tabone*, Audrey Guichard, Guy Oriol, Elisabeth Cerrato, Mélanie Denisot, Valérie Cheynet, Alexandre Pachot, Alain Lepape, Guillaume Monneret, Fabienne Venet, Karen Brengel-Pesce, Julien Textoris, François Mallet

And the MIP Rea Study Group and the REALISM Study Group.

* Contribution à parts égales

Année 2018

En préparation

Title : Modulation of LTR-retrotransposons expression in mHLA-DR stratified septic shock patients: a pilot study

Marine Mommert[1,2,*], Olivier Tabone[2,*], Audrey Guichard[1,2], Guy Oriol[1], Elisabeth Cerrato[2], Mélanie Denisot[2], Valérie Cheynet[1], Alexandre Pachot[2], Alain Lepape[3,4,5], Guillaume Monneret[2,6],Fabienne Venet[2,6], Karen Brengel-Pesce[1], Julien Textoris[2,7], François Mallet[1,2], the MIP Rea Study Group and the REALISM Study Group.

Author Affiliations:

*These authors have contributed equally to this work
[1]Joint research unit, Hospice Civils de Lyon, bioMerieux, Centre Hospitalier Lyon Sud, 165 Chemin du Grand Revoyet, 69310 Pierre-Benite, France.
[2]EA 7426 Pathophysiology of Injury-induced Immunosuppression, University of Lyon1-Hospices Civils de Lyon-bioMérieux, Hôpital Edouard Herriot, 5 Place d'Arsonval, 69437 Lyon Cedex 3, France.
[3] Intensive Care Unit, Centre Hospitalier Lyon Sud, Hospices Civils de Lyon, Pierre Bénite, France.
11
[4]Emerging Pathogens Laboratory, Epidemiology and International Health, International Center for Infectiology Research (CIRI), Lyon, France.
[5]Hospices Civils de Lyon, bioMérieux Joint Research Unit, Groupement Hospitalier Edouard Herriot, Lyon, France
[6]Hospices Civils de Lyon, Immunology Laboratory, Groupement Hospitalier Edouard Herriot, Lyon, France.
[7]Hospices Civils de Lyon, Department of Anaesthesiology and Critical Care Medicine, Groupement Hospitalier Edouard Herriot, Université Claude Bernard Lyon 1, Lyon, France.

Author Email Adresses :

Marine MOMMERT : marine.mommert1@ext-biomerieux.com
Olivier TABONE : olivier.tabone@biomerieux.com
Audrey GUICHARD: audrey.guichard@biomerieux.com
Guy ORIOL : guy.oriol@biomerieux.com
Elisabeth CERRATO : elisabeth.cerrato@biomerieux.com
Mélanie DENISOT: melanie.denisot@gmail.com
Valérie CHEYNET: valerie.cheynet@biomerieux.com
Alexandre PACHOT: alexandre.pachot@biomerieux.com
Alain LEPAPE: alain.lepape@chu-lyon.fr
Guillaume MONNERET : guillaume.monneret@chu-lyon.fr
Fabienne VENET : fabienne.venet@chu-lyon.fr
Karen BRENGEL-PESCE : karen.brengel-pesce@biomerieux.com
Julien TEXTORIS : julien.textoris@biomerieux.com
François MALLET: francois.mallet@biomerieux.com

Corresponding Author: Marine MOMMERT
Laboratoire Commun de Recherche bioMérieux / HCL
Centre Hospitalier Lyon Sud, Bat 3F
165 chemin du Grand Revoyet, 69310 Pierre-Bénite, FRANCE
Tel: +33 472 678 782
Email: marine.mommert1@ext-biomerieux.com

Abstract :

Background : Sepsis is defined as a life-threatening organ dysfunction caused by a dysregulated host response to infection. To date, no biomarker has shown sufficient evidence and ease of application in clinical routine to monitor the complexity of the sepsis-related immune alterations. Multiple levels of interaction of human endogenous retroviruses (HERVs) with the immune response led us to hypothesize that they may be relevant candidates both to contribute to the physiopathology of sepsis and to be part of a molecular signature. In this study, we used a recently described high-density microarray which allows exploring HERVs/MaLRs transcriptome in septic shock patients.

Results: About 6.9% of the HERVs/MalRs repertoire is transcribed in whole blood. Evidence was given that a subset of HERVs/MaLRs and genes were modulated in septic shock patients according to the monocyte HLA-DR (mHLA-DR) expression level measured by flow cytometry, as a proxy of immunosuppression. We then identified a large signature consisting of 193 differentially expressed HERVs/MaLRs probesets further reduced to a smaller 10 HERVs/MaLRs signature. Both signatures, validated in an independent septic shock cohort, identified two groups of patients with different severity features including clinical criteria (like mortality) and severity molecular markers (like CD74 ratio, invariant chain of HLA-DR).

Conclusion: This microarray-based approach unveiled the expression of about 87,912 distinct HERV probesets and identified 764 putative promoter LTRs and 642 putative polyA LTRs in whole blood. HERV/MaLR expression was shown to be tightly modulated in septic shock patients according to mHLA-DR expression. We identified a set of potential molecular biomarkers in septic shock patients partially overlapping immunosuppression (mHLA-DR), and the risk of Health Care associated Infections (HAI) (CD74 ratio), complementary to existing molecular markers of a sepsis patients stratification. We identified a HERV/MaLR signature discriminating patients based on their immunosuppression state and severity.

Keywords: HERV transcriptome, whole blood, sepsis shock patients, HLA-DR expression, biomarkers

Introduction :

Sepsis is defined as a life-threatening organ dysfunction caused by a dysregulated host response to infection, the septic shock is a severe subtype of sepsis [1]. Despite major therapeutic progress due in large part to the introduction of antibiotics and intensive care therapy, sepsis remains a major health issue with a high worldwide prevalence and a high mortality rate up to 30-40%. It is characterized by immune dysfunctions with concomitant excessive pro- and anti-inflammatory responses, leading to organ failure, immunoparalysis and secondary infections. The pathophysiological mechanisms of sepsis are still incompletely understood and the heterogeneity observed in patients makes the selection of the appropriate therapeutic care a major clinical challenge. The early model of an overwhelming inflammatory reaction failed to capture the complex pathophysiology of the syndrome, as proven by the failure of multiple clinical trials with a variety of anti-inflammatory agents [2]. While proinflammatory reactions are held responsible for tissue damage in sepsis, new insights suggest that the immune suppression that compensates inflammation contributes significantly to the late morbidity and mortality [3, 4]. For example, the association of low monocyte HLA-DR expression and high interleukin-10 concentration demonstrated that the anti-inflammatory response dominates after septic shock [5]. To date, the best marker for monitoring immune alterations in critically ill patients (sepsis, trauma, pancreatitis, surgery, burns) remains decreased HLA-DR expression on monocytes measured by flow cytometry as it provides valuable information in terms of mortality prediction or evaluation of risk for secondary infections. In the near future, molecular biology tools may circumvent some drawbacks related to flow cytometry [6].

Human endogenous retroviruses (HERVs) are hypothesized to be relevant candidates both to contribute to the physiopathology of sepsis Retrovirus-like sequences represent the 8,3% of the human genome [7]. They are constituted by about 200,000 HERVs and 240,000 Mammalian Apparent Long terminal Repeat (LTR) retrotransposons (MaLRs). HERVs are remnants of ancestral and independent retroviral infections within germ line.

From multiple endogenization events across evolution, they spread into groups of hundreds to thousands copies mainly through a copy/paste mechanism. HERVs share key structural features with infectious retroviruses, i.e. two LTRs surrounding *gag*, *pol* and *env* genes putatively encoding proteins. HERV expression which is heavily controlled by epigenetic mechanisms has been observed in both inflammation and immunosuppression [8]. Several lines of evidence support the role of HERVs as contributors of the immune response. The insertional polymorphism of HERVs LTR at the HLA locus has been associated with several auto-immune diseases [9]. In addition, HERV propagation might have an influence on the setting, regulation and rewiring of transcriptional networks, as illustrated by the IFN-inducible enhancers which might have been introduced and amplified by MER41 LTR propagation [10]. Thus, in placenta, LTR drive the physiological expression of fusogenic and immunomodulatory Syncytins [11], whereas, in B cell lymphoma, THE1B LTRs have been shown to induce the ectopic expression of CSF1R gene which participate to oncogenic processes [12]. These differences of expression in physiological and pathological contexts are the result of a tight epigenetic control of HERV expression. Indeed, it has been shown that multiple layers of epigenetic mechanisms such as DNA methylation [13, 14] and histone marks [15] control the expression of these transposable elements in tissue and/or differentiation dependent manner. The interaction between HERVs and the immune system can take place through either (i) modulating immune gene expression by their regulatory sequences (LTRs) or (ii) triggering innate immune sensing pathways by expressed HERVs RNAs or proteins. All these data led us to hypothesize that sepsis might be associated with HERV transcriptional reactivation that may modulate the immune response. Moreover, irrespectively of causal or consecutive expression HERV can be part in an informative molecular signature allowing a better stratification of septic shock patients.

First, using an HERV-V3 microarray which allows measuring HERVs at the individual locus level [16], we have recently shown that 5.6% of HERVs/MaLRs are transcriptionally active in peripheral blood mononuclear cells (PBMCs) of healthy volunteers. Furthermore,

the HERVs/MaLRs appeared tightly modulated in inflammatory and immunocompromised settings mimicked by *ex vivo* endotoxin tolerance model [17]. Second, we have performed a retrospective analysis of U133 Plus 2.0 datasets focusing on the 337 probesets targeting HERV loci, derived from 30 burn patients, 332 septic shock patients and 105 trauma patients. We identified 5 HERV loci which are differentially expressed in all patients compared to healthy volunteers [18]. Hence, in this pilot study, we used the HERV-V3 microarray, to investigate the HERVs and MaLRs expression in septic shock patients stratified with mHLA-DR level at day 3, following admission at the intensive care unit (ICU). We aimed to identify whether subsets of the HERVome may be expressed according to the immune status. More precisely, we first provided a global view of the HERV transcriptome in whole blood. Secondly, we showed the modulation of HERVs/MaLRs according to patients status. Finally, we have identified a large and reduced HERVs/MaLRs signature in the pilot cohort and validated it in an independent septic shock cohort, according to severity state.

Materials and Methods:

Septic shock patient cohorts:

Discovery cohort: Immunosepsis cohort subset

A subset of shock septic cohort previously published [21] was selected to performed the pilot study. Twenty patients samples was collected at day 1 and at day 3 or 4 (to simplify named day 3 thereafter). The HLA-DR expression was measured at day 3 or 4 by flow cytometry on the monocyte surface as previously described [22]. Ten patients had a normal expression of HLA-DR at day 3 or 4 (more than 30 % of expression), and 10 patients had a low expression of HLA-DR at day 3 or 4 (less than 30 % of expression). Therefore, patients used in this study are stratified by the HLA-DR expression at day 3-4, as a proxy of immunosuppression state. Table 1 describes the patient characteristics. Five healthy volunteers are also included in this analysis.

Table 1 : Description of the discovery IMMUNOSEPSIS cohort.

Patient characteristics at admission and outcomes according to mHLA-DR expression at day 3. Categorical variables are expressed as n (%) and continuous variables as median [Q1-Q3]. Comparisons between normal and low mHLA-DR expression at day 3 groups were performed with Chi-squared test for qualitative variables and Mann-Whitney for quantitative variables, as appropriate. Values in bold indicate significance at p<0.05. ICU: intensive care unit, SOFA: sequential organ failure assessment, IGS: simplified index of gravity, HLA-DR: major histocompatibility complex.

Variable	Total (n=20)	HLA DR normal at day 3 (n=10)	HLA DR low at day 3 (n=10)	pvalue
Characteristics at admission				
Gender, male, n (%)	13 (65)	8 (80)	5 (50)	0.315
Age	59 [52-68]	59 [54-67]	57 [46-69]	0.665
SOFA Score at day 1	9 [8-12]	8 [6-9]	11 [9-16]	**0.018**
IGSII Score at day 1	45 [34-56]	44 [33-45]	58 [47-68]	0.052
Outcomes				
Infection secondaires, n (%)	2 (10)	0 (0)	2 (20)	0.796
Hospital lenght of stay (days)	34 [21-48]	38,5 [28-58]	31 [14-42]	1.000
ICU lenght of stay (days)	6.50 [5-13]	5.50 [4-20]	8 [5-10]	1.000
Non-survivors at day 28, n (%)	9 (45)	2 (20)	7 (70)	0.052

Validation cohort: MIP-Rea cohort subset

A subset of septic shock patients MIP-Rea (Marqueurs immunitaires prognostiques en Réanimation) cohort previously published [19] was selected to perform the validation analysis. Hundred septic shock patients were selected from which we had samples at day 1, day 3 and day 6. Table 2 describes the patient characteristics. For this cohort the measure of mHLA-DR by flow cytometry is missing. We used the CD74 mRNA expression [19] to determine the proxy of severity at D3 and D1 and the CX3CR1 mRNA expression [20] to determine the proxy of mortality at D1 and D3.

Table 2 : Description of the validation MIP-Rea cohort.

Patient characteristics at admission and outcomes of the projection analysis. Categorical variables are expressed as n (%) and continuous variables as median [Q1-Q3]. ICU: intensive care unit, SOFA: sequential organ failure assessment, SAPSII: simplified acute physiology score II.

Variable	Total (n=100)
Characteristics at admission	
Gender, male, n (%)	67 (67)
Age	70 [59-77]
SOFA Score at day 1	11.5 [9-14]
SAPSII Score	63 [51-73]
Outcomes	
Infection secondaires, n (%)	26 (26)
Hospital lenght of stay (days)	28.5 [18-54]
ICU lenght of stay (days)	14 [10-21]
Non-survivors at day 28, n (%)	23 (23)

Sample collection and RNA extraction:

Peripheral whole blood from ICU patients or healthy volunteers was collected from ICU patients or healthy volunteers was collected in PAXgene^{TM} Blood RNA tubes (PreAnalytix). Samples were stabilized at least 4h at room temperature after collection and frozen at -80°C following the manufacturer's guidelines. For ICU patients, blood was collected at D1, D3 (and D6, for validation MIP-REA cohort), after ICU admission. Total RNA was extracted from whole blood using PAXgene Blood RNA kit (PreAnalytix) according to the manufacturer's instructions. Samples with RNA integrity number ≤ 6 were excluded due to poor quality RNA.

Custom Affymetrix HERV-V3 GeneChip Microarray:

HERV-V3 targets 353,994 loci-elements, represented by 4,410,200 probes. The custom HERV GeneChip can discriminate distinct HERV elements composed of a set of highly informative probesets (located in U3, R, U5 subdomains of solo, 5' and 3' individual LTRs and gag/pol/ env regions) hereafter named 'HERVs prototypes repertoire' and a set of probesets with lower quality annotations (located in the first third and last third of complete

LTR, and every 2.5 kb in the region in between LTRs) hereafter referred to as 'HERV/MaLR_Dfam repertoire'. The custom HERV GeneChip also contains probesets targeting LINE1, lncRNA, viruses, and genes repertoire. The descriptions of the HERVgDB4 database and of the final contents of HERV-V3 microarray are provided in Table S1 [16].

RNA amplification, labeling and hybridization :

The cDNA synthesis and amplification steps were performed from 16 ng of RNA using the Ovation Pico WTA System V2 kit (Nugen) according to the manufacturer's instructions. 5 micrograms of amplified purified DNA were fragmented into 50-200 bp fragments and were 3-labeled using Encore Biotin Module kit (Nugen) according to the manufacturer's instructions. The HERV-V3 microarrays were hybridized 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 the GeneChip fluidic station 450 (Affymetrix). The arrays were finally scanned using the GeneChip scanner 3000 7G (Affymetrix) fluorometric scanner. Images (DAT files) were converted to CEL files using the GCOS software (Affymetrix) [16]. The experimental data generated have been deposited in the National Center for Biotechnology Information (NCBI) and are available in the GEO DataSets site under access number (Accession number: GSE121352).

Microarray Analysis Preprocessing:

CEL files were transformed into matrix, normalized, adjusted for background noise (RMA normalization) and probes were summarized into probesets with command apt-probeset-summarize (V 1.18.0) with rma option. Microarray preprocessing and statistical analysis were performed using R/Bioconductor (R v3.1.2) [23]. Quality assessment was performed through simpleaffy (v2.42.0) [24] and arrayQualityMetrics (v3.22.1) [25]. For quality control, several criteria have been used: RNA quality, images of chips, hybridization spike-in, PolyA, Amplification and fragmentation, intensity signals (before and after normalization), probeset homogeneity (RLE, NUSE plots), correlation plots (before and after

normalization) and Principal Component Analysis. For most criteria, outlier detection have been done by computing Kolmogorov Smirnof (KS) statistic between each array and the pooled data (default threshold with arrayQualityMetrics library). And array was removed of the analysis if it did not pass more than four quality controls. One array was removed using these criteria, it's a chips corresponding to a healthy volunteers samples (four healthy volunteers are used to further analysis). Then data have been normalized, adjusted for background noise and summarized using RMA (Robust Multi-Array) algorithm [26]. Experiment batch effects were removed using COMBAT [27]. Finally a filtering process was done to reduce dataset and gain statistical power for analyzes. A Coefficient of Variation (CV) value of 10% was used to determine the intensity threshold from which the 3rd quartile of intensity ranked by range of values is under CV value. Then, probesets under intensity threshold in more than 68% of all samples (14 samples over 45) were removed. Consequently, a probeset was counted as transcriptionally active whether the normalized signal was observed (i) above a $2^{5.5}$ threshold , corresponding to the minimal intensity level shared by all repertoires showing an acceptable variability (defined as the value of the 75th percentile of the distribution of the coefficient of variation as a function of intensity lower than 10%) and (ii) for at least 14 samples out of 45.

Bioinformatics analysis :

Microarray analysis preprocessing was detailed in supplementary methods. We chose a specific threshold to define LTRs functions for "HERV_Dfam" repertoire. We used the dichotomy of probesets signals targeting. More precisely, to preserve sensitivity and robustness regarding function assignation, we willingly and arbitrarily selected a relatively low expression level cut-off of $2^{4.5}$ for positive signal attribution coupled with a significant fold change between U3 and U5 signals. Therefore, an LTR was referred to as 'promoter' (Pr) in cases where the signal of the U5-associated probeset was (i) over the threshold and (ii) at least 3 times higher than its U3 counterpart, and as 'polyadenylation signal' (pA) if the intensity of its U3-associated probeset was (i) over the threshold and (ii) at least 3 times

higher than its U5 counterpart. An LTR was assigned 'readthrough' (RdT) if both U3 and U5 signals (i) were other threshold and (ii) without significant fold change between its. Finally, a LTR was assigned as silent if U3 and U5 were both under threshold; all other remaining LTRs were classified as undetermined (158 LTRs). Principal component analysis was made on HERV expression matrix (prcomp() function on R). The representation on the 2 first components was made using ggplot2 library. The 2 vectors drawn on the plot are the median of coordinate differences between D1 and D3 for each patient, according to their DR status. To visualize HERVs and genes modulation, hierarchical clustering based on correlation distance with average method was made on the 1% most variable probesets. Then, comparisons between i) healthy volunteers (HV) and septic shock patients at day 1 (D1) or day 3 (D3) and ii) septic shocks patients at day 1 or day 3 between them according to HLA-DR low or normal expression were carried out. For all probesets, for differential expression analysis, moderated t-tests were performed (Limma, v3.22.7 [28] and p-values adjusted for multiple testing using the Benjamini-Hochberg procedure [29]. A probeset was considered statistically significantly differentially expressed when absolute log2 Fold Change (|log2FC|) was over 1 and adjusted-pvalue under 0.05. Graphics were made using ggplot2 (v2.2.0) or pheatmap (v1.0.8).

Pathways analysis:

The Ingenuity Pathways Analysis tool (IPA, Ingenuity® Systems, www.ingenuity.com) was used to assess upstream regulators, canonical pathways, disease and functions. This analysis was performed using both HTA and U133 genes repertoire, the significant pathways was extracted of HTA genes repertoire because the result is similar of U133 repertoire but also more complete and informational that U133 repertoire. Only canonical pathways were analyzed between all conditions in the present study. Canonical pathways were predicted to activated or inhibited for z-scores ≥ 2 and ≤ -2 respectively, a p-value cutoff of 0.05 was used (Fisher's exact test).

Primer Design and Validation:

PCR primers were designed whether possible overlapping probes that allowed their detection onto the chip. Design of primer pairs of locus-specific selected based on HERV-V3 microarray analysis using Primer3 and the NCBI Primer-BLAST software (www.ncbi.nlm.nih.gov/tools/primer-blast) and checked *in silico* at UCSC (https://genome.ucsc.edu). HPLC-purified primers were from Eurogenetec. Specificity and sensitivity of the systems were evaluated on 1 ng of human genomic DNA (Promega) by varying the annealing temperature (Tm) from 52°C to 60°C, and amplification cycles were followed by High Resolution Melting (HRM) using Rotor Gene Q (Qiagen), gel electrophoresis analysis on Bioanalyzer 2100 (Agilent) and product sequencing (GATC Biotech). Systems were validated if primer pairs follow three criteria: (i) one HRM peak, (ii) fragment size corresponding to the expected product and (iii) match to the targeted locus sequence after Sanger sequencing. Primer pairs along with an illustration of the experimental validation scheme were summarized in Figure S5 [30]. For Real-Time PCR, a one annealing temperature (56°C) was selected for all repeat-element systems validated, all criteria were validated for all systems at this temperature, and it allowed an homogenization of the experiment. The 24 primer pairs which satisfied our acceptance criteria were used to amplify the original RNA samples used for microarray experiments.

Real-Time PCR and data analysis:

mRNA expression levels were quantified using RT-qPCR. Total RNA (100ng) was DNAse-treated and reverse transcribed using QuantiTec Reverse Transcription Kit (Qiagen). Reverse-transcriptase-free reactions were carried out to verify the absence of contaminating genomic DNA using the TaqMan Gene Expression Assay Human 18S system Hs03003631-g1 and TaqMan Universal PCR kit (both ThermoFisher). For repeat-elements systems, SYBR green experiments were set up using the Type-it HRM PCR kit in 20 µL final reaction volume with 0.7 µM primers and a 10-fold cDNA dilution (2 ng RNA equivalent). PCR

amplifications were carried out in strip tubes which were closed by caps (all Qiagen). The cDNA amplifications were performed using Rotor Gene Q as follows: a 5 min denaturation step at 95°C, followed by 40 cycles (95°C for 10s, Tm for 30s, 72°C for 10s) and HRM analysis (from 65°C to 95°C, 0,1°C increments every 2s). All reactions were performed in duplicates. Expression of housekeeping genes: Peptidylprlolyl Isomerase B (PPIB) and Ribosomal Protein Lateral Stalk Subunit 0 (RPLP0) was monitored for normalization and other gene expression was investigated using TaqMan Universal PCR Master Mix and TaqMan Gene Expression Assay Human: PPIB (Hs00168719_m1), RPLP0 (Hs00420895_gH) (ThermoFisher). The PCR reactions were performed using Rotor Gene Q in strip tubes with 20 µL final reaction, primers and mix concentration and amplification program were determine according to manufacturer's instructions (Thermofisher). The fold change (FC) was determined using $2^{-\Delta\Delta Ct}$ method. The first ΔCT is the difference in threshold cycle between the target and the geometric mean of PPIB and RPLP0 genes; the ΔΔCT is the difference in ΔCT between the target and the healthy volunteers condition. The final value of the healthy volunteers condition was arbitrary set to one and other values scaled up in order to provide a final relative differential expression.

HERV/MaLR signature on validation cohort

Signature finding:

The signature contains the 193 HERVs/MaLRs probesets modulated at D3 between low HLA-DR and normal HLA-DR patients from discovery cohort, with absolute |log2FC| higher than 1 and adjusted p-value lower than 0.1 [29]. To perform the microarray experiment and pre-processing, the same protocol has been used than described before. The experimental data generated have been deposited in the National Center for Biotechnology Information (NCBI) and are available in the GEO DataSets site under access number (Accession number: GSE121352). Hierarchical clustering has been applied on validation cohort, on this 193 HERV/MaLRs probesets signature. Complete clustering

method was applied, with Euclidean distance for rows and Pearson's correlation distance for columns. The samples on heatmap were annotated with CD74 RT-qPCR value of ratio between D3 and D1. The molecular variable is a proxy of HAI, found on the same patients in previous studies [19, 20], and has been binarized according to the same criteria.

Signature reduction:

Random Forests classification algorithm [31] (R package randomForest v4.6-14) has been applied to sort the 193 HERV probesets by their "importance" in separation between DR normal and DR low patients, as defined in the package. Then, we selected the optimal signature size best discriminating normal and low CD74 ratio patients. We predicted the performances by computing Area Under the Curve (AUC) for signature size from top 2 to top 30 probesets sorted by importance. The highest AUC (0.69) was for a signature of 10 probesets. Using the same criteria as before, hierarchical clustering has been applied on the reduced signature.

Results:

Expression and modulation of the HERV/MaLR transcriptome in whole blood

In order to present an overview of the HERV/MaLR transcriptome in whole blood, we used a subgroup of twenty septic shock patients selected from the IMMUNOSEPSIS cohort previously described [21], and 4 healthy volunteers. Septic shock patients were chosen according to mHLA-DR expression at day 3 (10 patients with low expression and 10 patients with normal expression as defined in Material and Methods), and samples were collected on the day of admission (D1) and on day 3 (D3) after the admission. The transcriptome has been analyzed using the custom Affymetrix HERV-V3 microarray [16] which targets 353,994 HERV/MaLR elements and a set of 1,559 genes. The HERV/MaLR repertoire is subdivided into two sub-repertoires according to the level of knowledge in the literature. Onto the chip, this corresponds to a set of highly informative probesets corresponding to accurately annotated HERV loci called hereafter "HERV_prototypes", and a set of probesets

corresponding to roughly annotated HERVs/MaLR elements called hereafter "HERV/MaLR_Dfam" (detailed in Table S1).

A summary of the absolute counts and the relative abundances of transcriptionally active elements of the HERVs/MaLRs transcriptome is given in Figure 1A, at the probeset level. Notably, although the hybridization assessment quality was equivalent across repeated elements and gene repertoires [16], a higher proportion of gene probesets were transcriptionally active, i.e. 42% (32,310 probesets). Overall, 6.9% of targeted HERVs/MaLRs (87,912 probesets) were transcriptionally active in whole blood, therefore the HERV/MaLRs transcriptome seemed globally silent compared of genes repertoire. More precisely, 11.5% of the well-described "HERV_prototypes" repertoire, and 6.1 % of "HERV/MaLR_Dfam" were expressed. Among the expressed prototype elements, 7.4% belonged to gamma-retrovirus like the HERV-H groups containing the most important number of active elements (1042 expressed probesets corresponding to 19% of the whole group). In beta-retroviridae, 2.1% of elements are expressed, among them, HML-8 and HML-1 groups have the most important count of expressed elements (861 probesets, 14% of the whole group and 216 probesets, 14% of the whole group respectively). Finally, 2% of expressed elements belonged to spuma/epsilon-like retrovirus classes as HERV-L group provides the largest count of expressed probesets (1396 probesets corresponding to 11% of the whole group). At the group level, all well-defined HERV groups had expressed loci in whole blood. Nevertheless, we did not observe significant enrichment or depletion of a specific repertoire, class or HERV group, according to healthy volunteers (even if the study contains only 4 healthy volunteers), normal or low expression of HLA-DR or day after shock (data not shown). Of note, if we compare the "HERV_prototypes", "HERV_Dfam" and "MaLR_Dfam" repertoire transcriptional activities, the "HERV_prototypes" (10,050 expressed probesets corresponding to 11.1%) had higher proportion of expressed loci than Dfam elements (37,427 and 40,435 expressed probesets corresponding to 6.6% and 6.5% for HERVs and MaLRs respectively). Thus the proportion of well-annotated transcribed

probesets was larger than “HERV/MaLR_Dfam” transcribed probesets. Lastly, the same analysis was carried out on a sub-group of 100 septic shock patients selected from MIP-REA cohort previously published [19] (Figure S1). Septic shock patients were chosen from which we had samples at day 1, day 3 and day 6. Of note, 92% of expressed probesets were in common between two cohorts.

The HERV/MaLR transcriptomic activity observed collectively in healthy volunteers and in patients is depicted exhaustively at the structure level (LTR/*gag*/*pol*/*env* regions) and we assessed the LTR functionality by assigning them promoter (Pr), polyA (pA) or silent states. These results were summarize in Figure S2. Firstly, Figure S2A provides a simplified comparative view of the “HERV_prototype” repertoire addressed by the chip and its related transcriptome. Active LTRs represent 77.2% of the HERV transcriptome, whereas they represent 71.5% of the chip. Thus LTRs are more represented in active elements than internal HERV/MaLR regions (22.8% of the HERV transcriptome versus 28.5% on the HERV-V3 chip). Secondly, Figure S2B highlighted LTRs functions, the main attributed function, i.e. observed in most of the conditions, was the silent phenotype (75.8% of LTRs),a promoter (Pr) activity and polyadenylation (pA) signal were assigned to 12.1% and 10.1% of LTRs respectively. Finally, we wanted to know whether each single LTR had the potential to change its state (Figure S2C). As expected, 54.2% of LTRs (2,970 LTRs) were systematically silent in all samples. Interestingly, 34.4% of LTRs shifted from a silent status to a promoter (1,119 Silent/Pr LTRs) or polyadenylation (767 Silent/pA LTRs) function. A significant proportion of LTRs, 7.4%, act as promoter (233 Pr LTRs) or polyA (174 pA LTRs) in all samples, and very few pA/Pr shift occurs. Consequently this observation confirms that the shift from promoter to polyA function is an extremely rare if significant event. The same analysis was performed using the complete HERV/MaLR dataset (Figure S2D). Although results should be considered with caution due to the imprecise annotation of “HERV_Dfam” and “MaLR_Dfam” LTRs (Table S1), the overall trends were similar to those observed at the “HERV_prototypes” repertoire level.

HERVs/MaLRs probesets were expressed in whole blood from septic shock patients and healthy volunteers, we then wanted to know if these probesets were modulated between conditions. We performed a differential expression analysis. The exhaustive dataset of this analysis is presented in Table S2. Volcano plots are depicted at the probeset level for HERVs/MaLRs (Figure 1B) and both genes and HERV/MaLR (Figure S3). A similar amount of up- and down-modulated HERVs/MaLRs probesets was observed between D1 or D3 patients and healthy volunteers. More precisely, when we compared each group of patients samples (low or normal mHLA-DR, D1 or D3) with healthy volunteers, we observed, as expected, numerous modulated probesets, i.e. at D1 1,793 and 1,663 probesets are up- and down-regulated respectively, and at D3, 1,833 and 2,274 probesets are up- and down-regulated respectively (Figure 1C) (adjusted p-value < 0,05, |log2FC| > 1). The same analysis was carried out on genes and HERVs/MaLRs repertoires (Figure S4A). Thus, we have highlighted that HERV/MaLR loci were expressed and modulated in healthy volunteers and septic shock patients whole blood.

HERVs/MaLRs modulation according to immune status of septic shock patients

To gain insight into the modulation of genes and HERV/MaLR expression associated with patients status, we performed a differential expression analysis this time between normal and low mHLA-DR patients as well as a principal component analysis (PCA).

Initially, we wanted to observed if the HERV/MaLR modulation was dynamic over the time between septic shock patients. Mainly, the HERV/MaLR regulation did not seem dynamic, 3,456 and 4,107 distinct HERV/MaLR loci were modulated at D1 and D3 respectively (Figure 2A). However, differences between septic shock patients were revealed by principal component analysis (PCA) and as stated above, the difference observed between healthy volunteers and septic shock patients and according to immune states was validated on PCA. PCA was built upon the whole HERVs/MaLRs dataset expression matrix, showed primarily that all septic shock patients were clearly separated from healthy

volunteers (Figure 2B). Nevertheless, low versus normal mHLA-DR groups appeared to behave differently. More precisely, D1 to D3 orthogonal progressions were observed for the two sub-populations, as illustrated by the median of coordinate differences between D1 and D3 (arrow on Figure 2B); depicting patients with normal mHLA-DR tend to get closer to healthy people. In addition, when we explored differentially expressed HERVs/MaLRs loci referred to as DEL between patients according to day of sampling and mHLA-DR status, the first striking observation relates to the incapacity to discriminate at D1 low versus normal mHLA-DR patients (Figure 2C). On Venn diagram analysis, more HERVs/MaLR elements appeared to be under-expressed at D3 in patients with a low versus normal expression of mHLA-DR marker, more precisely 398 probesets were down-modulated and 83 probesets were up-modulated. The same analysis was carried out on genes and HERVs/MaLRs repertoires (Figure S4).

Concerning the gene expression, pathway analysis of septic shock patients with low versus normal expression of mHLA-DR (Table S3) illustrates that most of the modulated pathways (12/13) are repressed in immunocompromised situation. Although to a lesser extent, the comparison of septic shock patients with healthy volunteers showed also that the majority of modulated pathways are repressed in ill people (12/18). Overall, differential expression analysis of low versus normal mHLA-DR condition identified a down regulation of genes involved in antigen recognition, TCR activation or regulation of B cells such as CD3G, DPP4, CD40LG but also chemokines and cytokines receptor such as CCL5 and IL5RA. Up-regulated genes were IL18R1 and CYP1B1 involved in cytokine signaling and cell proliferation, migration and survival, HGF, IL1R2 and PPARG genes involved in cell growth, apoptosis or NF-κB response (Figure 2D).

Concerning the HERV/MaLR expression, when we compared low versus normal mHLA-DR at D3, out of 87 differentially expressed HERV/MaLR elements, 40 belonged to “HERV_Dfam”, 40 to “MaLR_Dfam”, and 7 to “HERV_prototypes”. Among the 7 “HERV_prototypes”, we can observe that 3 solo LTRs and 4 complete or partial proviruses

462 are differentially expressed (Figure 2D). No group appeared enriched or depleted among the
463 most DELs. Although DEL analysis does not indicate any group or chromosomal enrichment,
464 some characteristic points could be observed. Among the 87 HERVs/MaLRs differentially
465 expressed in mHLA-DR stratified patients, 63 also discriminated septic shock patients at D3
466 and healthy volunteers, and so 24 HERVs/MaLRs reflected specifically the altered immune
467 status. They consist of one ERV9 element in the prototype repertoire and 9 and 14 loci in the
468 “HERV_Dfam” and “MaLR_Dfam” repertoires, respectively.

469 To confirm by RT-qPCR the relevance of the DEL and reduced HERVs/MaLRs
470 molecular signature, we selected 23 HERV/MaLR loci significantly modulated in patients at
471 D3 according to mHLA-DR status. HERV/MaLR locus-specific RT-qPCR systems were
472 meticulously designed and validated on gDNA matrix to secure locus specificity (Figure S5).
473 Figure 2E illustrates two examples of RT-qPCR profiles results, overall we could observe 3
474 HERVs/MaLRs phenotypes allowing to discriminate the immune status of patients. We
475 identified HERVs/MaLRs elements with a lowered expression according to low mHLA-DR
476 condition, such as those presented on Figure 2E. Conversely, we identified HERVs/MaLRs
477 with a significant increase in patients with low mHLA-DR (Figure S6). Interestingly, we found
478 also HERVs/MaLRs elements discriminating normal and low mHLA-DR expression patients
479 from D1 (Figure S6). Consequently, we have identified a link between mHLA-DR status used
480 as a proxy of immune state and the HERVs/MaLRs transcriptome modulation.

481 **HERVs/MaLRs elements: new biomarkers of immune state?**

482 Finally, we sought to evaluate whether the identified HERVs/MaLRs modulated in the
483 discovery cohort were expressed and able to stratify and independent cohort of septic shock
484 patients. HERV/MaLR expression was measured on a subset of the MIPREA sepsis cohort
485 (Table 2) with the HERV-V3 tool as done for the discovery cohort. Initially, we selected the
486 largest signature consisting of the 193 HERV/MaLR probesets (corresponding to 164 distinct
487 loci) which are modulated between normal vs low mHLA-DR patients at D3 (absolute log2FC

higher than 1 and adjusted p-value lower than 0.1). With this signature consisting of 193 probesets we made unsupervised analysis on the MIPREA validation cohort, on samples at D3.

The dendrogram on columns highlighted two well separated groups of patients (Figure 3A). The first group hereafter called cluster 1M, consists of 60 septic shock patients and the second group called cluster 2M includes 40 septic shock patients. We compared different variables between the 2 groups. Cluster 1M is characterized by 27% of patients presenting an HAI and 27% of patients that did not survive. Cluster 2M includes a similar amount of patients presenting an HAI (25%) but much less D28 mortality (18%). Apart from clinical endpoints, molecular markers were proposed to contribute to patients stratification, notably the D3/D1 CD74 ratio proposed as proxy of HAI occurrence [19], the CX3CR1 expression proposed as a proxy of mortality [20] (Table S4). Cluster 1M is characterized by 65% of patients with low CD74 ratio and 80% have low expression of CX3CR1 at D3 Cluster 2M included 33% of patients that have low CD74 ratio, 20% of them have low expression of CX3CR1 at D3. Last, both groups present significantly different SOFA scores (at D3 pvalue 0.003 and at D6 pvalue <0.001) at D3 and D6, confirming cluster 1M contains more severely injured patients (Table S4).

Secondly, we have reduced the 193 HERVs/MaLRs large molecular signature using the random forest method (see Material and method). Thus, we have obtained a reduced molecular signature including 10 HERVs/MaLRs corresponding to 8 distinct loci. This signature shows the same performances on validation cohort for stratification according to CD74 ratio. Similarly, the dendrogram on columns highlighted two well separated groups (Figure 3B). The first group hereafter called cluster 1M', consists of 43 patients and the second group called cluster 2M' includes 57patients. Cluster 1M' is characterized by 26% of patients presenting an HAI and 30% of patients that did not survive. Cluster 2M' includes a similar amount of patients presenting an HAI (26%) but much less D28 mortality (18%). Concerning, molecular markers, cluster 1M' is characterized by 74% of patients with low

CD74 ratio and 81% of them have a low expression of CX3CR1 at D3. Cluster 2M' included 35% of patients that have a low CD74 ratio and 37% of them have a low expression of CX3CR1 at D3. Both groups present significantly different SOFA scores (at D3 pvalue 0.019 and at D6 pvalue <0.001) at least at D3 and D6, suggesting cluster 1M' contains more severely injured patients (Table 3 and extended Supplementary Table S5). Of note, all severe patients identified by the largest molecular signature, are also found in the cluster exhibiting severe criteria with reduced HERVs/MaLRs signature. In addition, this difference observed between two clusters according to CD74 ratio was significant (Figure 3C).

These observations confirmed that the HERV signature identified in mHLA-DR stratified patients helps to separate independent septic shock patients based on their severity and immune status.

Table 3 : Patient characteristics according to cluster in projection analysis

Patient characteristics at admission and outcomes according to cluster of the projection analysis. Categorical variables are expressed as n (%) and continuous variables as median [Q1-Q3]. Comparisons between two clusters were performed with Chi square test for qualitative variables and Mann-Whitney or t tests for quantitative variables, as appropriate. Values in bold indicate significance at p<0.05. SOFA: Sequential Organ Failure Assessment. CNRQ: Calibrated Normalized Relative Quantities.

Variable	cluster 1M' (n=43)	cluster 2M' (n=57)	Total (n=100)	P-value
Characteristics at admission				
Gender (male)	29 (67.44 %)	38 (66.67 %)	67 (67.00 %)	
Age (Years)	71.00 [59.50,77.00]	67.00 [58.00,78.00]	69.50 [59.00,77.25]	0.830
SOFA Score J3	12.00 [8.00,13.50]	8.00 [6.00,12.00]	9.00 [7.00,13.00]	**0.003**
Cell count				
White cells J3, 10^9/L	(n=43)	(n=56)	(n=99)	**<.001**
	18.00 [12.28,22.50]	11.93 [8.66,15.91]	14.67 [9.91,18.74]	
Neutrophils J3, 10^9/L	(n=36)	(n=41)	(n=77)	**<.001**
	17.00 [11.91,23.21]	9.11 [7.10,13.20]	13.04 [8.60,17.21]	
Lymphocytes J3, 10^9/L	(n=36)	(n=41)	(n=77)	
	0.67 [0.40, 0.81]	0.93 [0.64, 1.14]	0.80 [0.59, 1.00]	
Platelets J3, 10^3/mm^3	(n=43)	(n=56)	(n=99)	**<.001**
	92.00 [52.50,156.50]	125.50 [80.00,224.75]	119.00 [72.00,200.00]	
Molecular markers				
CNRQ_CX3CR1_D3	0.12 [0.08,0.21]	0.31 [0.18,0.59]	0.21 [0.10,0.39]	**<.001**
RATIO_D3-D1_CNRQ_CD74	0.84 [0.57,1.28]	1.33 [0.95,1.79]	1.21 [0.68,1.69]	**<.001**
CNRQ_S100A9_D3	11.62 [9.18,15.52]	6.93 [5.42,10.79]	8.80 [6.03,13.26]	**<.001**

To confirm by RT-qPCR the relevance of reduced HERVs/MaLRs molecular signature, we selected the 8 HERVs/MaLRs elements of the reduced signature. HERV/MaLR locus-specific RT-qPCR systems were meticulously designed and validated on gDNA matrix to secure locus specificity (Figure S5 and Figure S6). Figure 3D illustrates two examples of HERVs loci include in reduce molecular signature and validated on Immunosepsis cohort. This HERVs loci had already been found within the differential expression analysis. The other results of RT-qPCR validation were presented in Figure S6.

Discussion:

Although the early mortality rate during the first days after the shock has considerably declined over the past 30 years, 50% to 70% of patients die late, several weeks after the shock. As this may reflect a persistent immunosuppression state [3], there is a need to identify patients who would benefit from immunostimulatory therapies, as recently demonstrated using mHLA-DR expression as a stratification tool for IFN-γ immunostimulatory treatment [33]. Furthermore, flow cytometric measurement of HLA-DR has been shown to be a reliable biomarker for the prediction of death and nosocomial infection in sepsis patients [21, 34, 35] therefore reinforcing the idea that biomarkers of immune states after sepsis are useful. Although highly valuable, the flow cytometric format of mHLA-DR measurement makes it difficult to implement in large multi-centered studies and in routine labs [22]. Conversely, molecular markers may be easier to introduce in routine as illustrated by automated tests with standardized methodologies for pathogen detection [36]. Currently, several mRNA biomarkers appear to be promising candidates as surrogate markers of sepsis induced immunosuppression such as CX3CR1 [20, 37], IL10 and CD74 [19] or other genes signature targeting host immune response [38, 39]. Although these observations are remarkable, they explain only partially the morbidity and the mortality found in sepsis. This report suggests that additional biomarkers are required to allow an optimal stratification of the patients taking into account at the same time the inter-individual variability (individual sub-groups) and the immune status in the very first days after the ICU admission, in order to

deliver an appropriate therapy. These biomarkers have to be informative concerning the clinical outcomes such as (i) the risk of mortality, (ii) the HAI occurrence and (iii) the immune status including inflammation and immunosuppression to allow the discrimination between the two phases.

This study highlights the potential of HERVs/MaLRs elements to complement the biomarkers candidates that will help to stratify patients. First, we described in a discovery cohort the HERVs/MaLRs transcriptome modulation in whole blood of healthy volunteers and septic shock patients. Second, we highlighted the HERVs/MaLRs transcriptome modulation according to patients vs healthy volunteers and also depending on immune state. Last, we have been able to stratify a validation cohort according to severity criteria, with a large then a reduce HERVs/MaLRs molecular signature obtained from the discovery cohort analysis.

HERVs/MaLRs transcriptome in whole blood

We used the HERV-V3 chip to provide a first global view of the HERV transcriptome in whole blood. We observed that about 6.9% of LTR retrotransposons was transcriptionally active, similar to 5.6% we had previously observed ex vivo in PBMCs/endotoxin tolerance model [40]. Of note, 82.4% of the expressed probesets (58 587 probesets) were shared between whole blood and PBMCs.

The observed differences between both studies can be due to (i) the variability among whole blood samples, which appeared larger in blood of sepsis patients than in the endotoxin tolerance model, and/or (ii) the cell type composition such as the presence of neutrophils in whole blood, as well as (ii) the stimuli released from the endothelial environment. We observed a higher proportion of gamma-retroviruses including notably HERV families coding potentially for whole or parts of envelope proteins exhibiting an immunosuppressive domain (ISD) [11], such as HERV-H and HERV-W families previously described on B cells and monocytes in multiple sclerosis [41], or HERV-FRD, HERV-Fc2 and HERV-T families [42]. This is also consistent with group-based PCR approaches in PBMCs [43], and in MDM cell

588 lines [44], as well as with the detection/modulation of expression of MSRV/HERV-W and
589 HERV-H loci in PBMCs of healthy donors and multiple sclerosis patients [41, 45-47]. In
590 addition, we detected the expression of poorly characterized groups such as MER52A and
591 PRIMA 41. MER52A, whose promoter function was evidence in HEK-293T kidney cell [48],
592 appeared negatively controlled by epigenetic factors such as methylation in normal
593 endometrium, stem cell or PBMCs, but not in breast myoepithelial cell [48]. The epigenetic
594 modulation in sepsis [49], could allow its reactivation in PBMCs. PRIMA41 elements bear
595 MER41C/D LTRs. MER41 being recently identified as enhancer for adjacent IFN-induced
596 genes and revealed their involvement in the regulation of essential immune functions,
597 including activation of the AIM2 inflammasome through the transcription factor STAT1
598 binding. Interestingly, STAT1 is one of the 3 genes (STAT1, CCR4 and HLA-DRB1/B3)
599 identified as differentially expressed exclusively between mHLA-DR stratified patients but not
600 between patients and healthy volunteers.

601 **Modulation of HERVs and genes following HLA-DR stratification in septic shock**
602 **patients**

603 After looking at the HERV transcriptome landscape in whole blood, we analyzed
604 whether individual HERVs/MaLRs could be finely regulated between patients and healthy
605 volunteers, according to mHLA-DR status. Principal component analysis illustrated on one
606 hand that HERV/MaLR elements gathered some septic patients with healthy volunteers and
607 on the other hand separates patients according to their immunosuppression status. Overall,
608 we observed a similar amount of up-regulated (and down-regulated elements between septic
609 shock patients and healthy volunteers, as previously shown in other microarray experiment
610 targeting specifically genes in septic shock patients [54]. Such behavior reflecting
611 heterogeneous activation and repression within families of HERVs/MaLRs was found in the
612 highly inflammatory context of burn patients [55][56]. Of note, in mice, the replication
613 competent MuERV are reactivated in experimental polymicrobial peritonitis model mimicking
614 sepsis [57]. The potential of stratification of HERVs/MaLRs elements was confirmed on the

validation cohort. The 2 HERVs/MaLRs molecular signatures (large and reduced) separate patients into two groups, with distinct severity criteria such as: SOFA score, HAI and mortality as well as molecular markers CD74 involved in antigen presentation, CX3CR1 involved in adhesion and migration of leukocytes or S100A8/9 involved in cellular process such as cell cycle progression and differentiation. HERV/MaLR element biomarkers are complementary and not overlapping of gene biomarkers.

As this pilot study was designed as a proof of concept study to show that HERV RNA markers could serve as marker of immunosuppression and further stratification tool for immune-stimulatory treatments, larger studies involving higher number of patients are now warranted. Some of the herein identified immunosuppression HERVs/MaLRs markers will thus be evaluated in the REALISM study [58] This study aims to determine the incidence, severity and persistency of innate and adaptive immune alterations in 550 ICU patients. (septic shock, severe trauma/burn and major surgery) compared to 150 age-matched healthy volunteers.

Conclusion:

Development of immunosuppression early after sepsis is well established and there is a growing interest for immunostimulatory treatments in those patients. However, because of patients' heterogeneity, there is a need for robust biomarkers that will help us to stratify patients prior to initiation of those therapies and monitor drug efficacy. In the present study, we used a microarray-based approach unveiling the expression of about 6.9% of HERVs/MaLRs elements in whole blood, putatively associated or contributed to immune response. We identified a panel of 164 HERVs/MaLRs loci that are differentially expressed in septic shock patients stratified by the HLA-DR and showed on an independent cohort that they may help to classify patients according to severity criteria, including molecular markers. The added value of these HEVRs/MaLRs newly identified markers together with classical gene (e.g. CX3CR1) should now be evaluated in a larger cohort of septic patients. Upon

favorable results, we may hypothesize they could serve as a stratification tool prior to immune-stimulatory treatment and to monitor drug efficacy.

FIGURE LEGENDS :

Figure 1: Expression and modulation of the HERV/MaLR transcriptome in whole blood.

(A) Percentages and absolute counts of positive signal-associated probesets within individual groups of “HERV_prototypes” repertoire and “HERV_Dfam” and “MaLR_Dfam” repertoires. A probeset was included as reflecting a significant transcriptional activity whether its normalized intensity was over an intensity threshold of $2^{5.5}$ in at least 14 out of the 45 samples (for all conditions). This conservatory threshold was defined as the minimal intensity level shared by all repertoires that showed an acceptable variability, i.e. the 75^{th} percentile of the distribution of the variation coefficient as a function of intensity should be lower than 10%. “HERV_prototypes” groups were grouped by classes of retroviruses namely gammaretrovirus (green), betaretrovirus (red) and spuma-epsilon like retrovirus (blue). “HERV and MaLR Dfam” repertoires were depicted as a global homogeneous entity (purple). (B) Volcano plots derived from the HERVs/MaLRs differential expression analysis, on the left for septic shock patients at day 1 vs healthy volunteers (HV) and on the right for septic shock patients at day 3 vs HV. The x-axis represents the log2 fold changes values and the y-axis the -log10 adjusted pvalues. Each point represents these values for a probeset. Colored points show the significantly modulated probesets (adjusted pvalue < 0.05, log2FC < -1 (red) or log2FC > 1 (green)). (C) Venn diagrams from HERVs/MaLRs differential expression analyses, comparisons between patients and HV. The samples are grouped according to mHLA-DR status of patients (normal or low) and day after admission (D1 or D3). The green numbers represent the up-modulated probesets, the red numbers represent the down-modulated probesets.

Figure 2: HERVs/MaLRs modulation according to immune status of septic shock patients.

(A) Venn diagrams from HERVs/MaLRs differential expression analyses, comparisons between septic shock patients. The samples are grouped according to mHLA-DR status of patients (normal or low) and day after admission (D1 or D3). The green numbers represent the up-modulated probesets, the red numbers represent the down-modulated probesets. (B) Principal component analysis made from HERV expression matrix. Healthy volunteers, patients with normal expression of mHLA-DR and patients with low expression of mHLA-DR are indicated in grey, salmon and blue respectively. The 2 vectors drawn on the plot are the median of coordinate differences between D1 and D3 for each patient, according to their DR status. (C) Volcano plots derived from the HERVs/MaLRs differential expression analysis, for septic shock patients at day 1 vs day 3. The x-axis represents the log2 fold changes values and the y-axis the -log10 adjusted pvalues. Each point represents these values for a probeset. Colored points show the significantly modulated probesets (adjusted pvalue < 0.05, log2FC < -1 (red) or log2FC > 1 (green)). (D) The table present the number of statistically significantly differentially expressed elements, at locus level (DELs) for HERVs/MaLRs and differentially expressed genes (DEGs). Down-modulated loci are in red, up-modulated loci are in green. For HERVs/MaLRs elements, the name, number of differentially expressed probesets (between brackets) and chromosomal locations (in italic) are indicated (hg38 version of genome). For genes, the current gene symbol and the number of differentially expressed probesets (between brackets) are indicated. (E) This figure illustrates the RT-qPCR of 2 HERVs elements which decreased at day 3 in septic shock patients with a low expression of mHLA-DR. Expression was measured using mRNA derived from whole blood of septic shock patients or healthy volunteers used in the discovery microarray experiment. All PCR reactions were performed in duplicates for each conditions. Expression of housekeeping genes PPIB and RPLP0 was monitored for normalization. The fold change (FC) is determined using $2^{-\Delta\Delta Ct}$ method. The final value of the healthy volunteers condition was arbitrary set to one and other values scaled up in order to provide a final relative differential expression (data were represented by a median and using log2 scale).

Statistically significant differences between two conditions are marked (wilkoxon signed rank test, **: p-value < 0.05 and *: pvalue < 0.1).

<u>Figure 3:</u> HERVs/MaLRs elements: can be used as new biomarkers of immune state?

(A) Projection of the large HERVs/MaLRs signature on septic shock patients validation cohort. The 193 probesets HERVs/MaLRs selected in signature are those modulated at D3 between HLA-DR low and HLA-DR normal patients from discovery cohort, with absolute |log2FC| higher than 1 and adjusted p-value lower than 0.1 On validation cohort previously described [19], hierarchical clustering has been made on these 193 probesets using correlation distance and average method. Samples are annotated according to CD74 ratio. (B) Projection of reduced HERVs/MaLRs signature on septic shock patients validation cohort**.** The 10 probesets HERVs/MaLRs selected after reduction of the large HERVs/MaLRs signature including 193 HERVs/MaLRs probesets modulated at D3 between normal and low HLA-DR patients from discovery cohort, with absolute |log2FC| higher than 1 and adjusted p-value lower than 0.1 On validation cohort previously described [19], hierarchical clustering has been made on the 10 probesets from HERV/MaLR signature using correlation distance and average method. Several criteria are indicated on this heatmap: CX3CR1 expression, patients death at day 28, HAI, CD74 ratio and the day. (C) Box plots showing the ratio D3/D1 of the CD74 value measured by RT-qPCR in two clusters of MIP-REA cohort analysis present in Figure 3A. The difference between this 2 clusters is statistical significant. (D) This figure illustrates the RT-qPCR of 2 HERVs elements include in reduce molecular signature and previously identify in differential expression analysis. Expression was measured using mRNA derived from whole blood of septic shock patients or healthy volunteers used in the discovery microarray experiment. All PCR reactions were performed in duplicates for each conditions. Expression of housekeeping genes PPIB and RPLP0 was monitored for normalization. The fold change (FC) is determined using $2^{-\Delta\Delta Ct}$ method. The final value of the healthy volunteers condition was arbitrary set to one and other values scaled up in order to provide a final relative differential expression (data were represented by a median and using

log2 scale). Statistically significant differences between two conditions are marked (wilkoxon signed rank test, **: p-value < 0.05 and *: pvalue < 0.1).

SUPPLEMENTARY LEGENDS :

Figure S1: Description of HERVs/MaLRs transcriptome in whole blood of septic shock patients belonging to MIP-Rea sub-cohort.

Percentages and absolute counts of positive signal-associated probesets within individual groups of "HERV_prototypes" repertoire and "HERV_Dfam" and "MaLR_Dfam" repertoires. A probeset was included as reflecting a significant transcriptional activity whether its normalized intensity was over an intensity threshold of $2^{5.5}$ in at least 14 out of the 45 samples (for all conditions). This conservatory threshold was defined as the minimal intensity level shared by all repertoires that showed an acceptable variability, i.e. the 75^{th} percentile of the distribution of the variation coefficient as a function of intensity should be lower than 10%. HERV prototypes groups were grouped by classes of retroviruses namely gammaretrovirus (green), betaretrovirus (red) and spuma-epsilon like retrovirus (blue). "HERV and MaLR Dfam" repertoires were depicted as a global homogeneous entity (purple).

Figure S2: Genomic, transcriptomic and functional projections of the "HERV_prototype" repertoire.

(A) A HERV structure relative proportions, in the whole "HERV_Prototype" repertoire (HERV-V3 chip, Table S1b) and in the expressed elements (Transcriptome, Table S1c). Solo LTR, 5' LTR, 3' LTR and proviral genes account for 100% of the HERV chip. (B) LTR features. The descriptive table summarizes the assignment of features. [a,b,c,d] Loss of information from HERV database to understandable functions. [a] Summary of Table S1a, [b] Summary of Table S1b, [c] Enumeration of LTRs whose function is attributable, i.e. defined as LTR combining U3 and U5 adjacent structures on the genome and existing probesets onto the chip allowing a discrimination between U3 and U5 expression signals, [d] Enumeration of LTRs whose function is attributed using both $2^{4.5}$ positive threshold and a fold change of 3 between U3 and U5

747 regions. More precisely, to preserve sensitivity and robustness regarding function
748 assignation, we willingly selected a lower expression level cut-off of $2^{4.5}$ for positive signal
749 attribution. Thus, the LTR was referred to as promoter (Pr), polyadenylation signal (pA),
750 read-through (RdT) or Silent. All other remaining LTRs were classified as undetermined. (C)
751 Specialization of LTRs features. Number of LTRs from the "HERV_prototypes" repertoire
752 according to all the combinations of functions observed in each of the 45 whole blood
753 samples. Silent LTR- Pr (promoter), pA (polyadenylation signal)- RdT(read-through). The
754 obtained combinations could either contain one (e.g. Pr), two (e.g. Silent/Pr), three (e/g.
755 Silent/pA/Pr) or four functions (Silent/pA/Pr/RdT). 987 LTRs were excluded of the analysis as
756 classified at least once as undetermined. (D) Specialisation of LTR features on the whole
757 dataset. Number of LTRs from all HERV repertoires exhibiting all the combinations of the
758 features determined in each of the 44 whole blood samples. The combinations obtained
759 could contain a single function (eg Pr), two functions (eg Silent/Pr), three functions (eg
760 Silent/pA/Pr) or four functions (Silent/pA/Pr/RdT) ; 47,277 LTRs were excluded from the
761 analysis as classified at least once as undetermined because they did not meet the threshold
762 and fold change criteria. Silent LTR, Pr (promoter), pA (polyadenylation signal)- RdT (read-
763 through).

764 **<u>Figure S3:</u> Volcano plot from differential expression analyses of HERVs/MaLRs and**
765 **genes expression.**

766 Differential gene and HERV/MaLR expression analysis. Volcano plots derived from the
767 differential expression analysis were presented according to condition (A) D1 vs HV, (B) D3
768 vs HV, (C) D3 vs D1 and (D) D3 mHLA-DR low vs D3 mHLA-DR normal. The x-axis
769 represents the log2 fold changes values and the y-axis the -log10 adjusted pvalues. Each
770 point represents these values for a probeset. Colored points show the significantly modulated
771 probesets (adjusted pvalue < 0.05, log2FC < -1 (red) or log2FC > 1 (green)).

Figure S4: Venn diagrams from differential expression analyses of HERVs/MaLRs and genes expression.

(A) Comparisons between patients and HV. (B) Comparisons between normal and low mHLA-DR patients at D1 and D3. The samples are grouped according to mHLA-DR status of patients (normal or low) and Day after admission (D1 or D3). The green numbers represent the up-modulated probesets, the red numbers represent the down-modulated probesets.

Figure S5: Selection, design and quality criteria for the design of locus specific qPCR systems, illustrated with the 060400302-HERV0489uL locus, and PCR systems obtained.

Starting where possible from the probe regions of annotated HERV elements, locus-specific PCR primers were designed. This figure shows an example of acceptance criteria of HERV candidate locus acceptance criteria in order to use these primers targeting the locus of interest on a qPCR validation experiment. (A) The table represents primers sequences for this locus, the expected Tm and the expected size. (B) Amplification curve and high-resolution melting curve obtained on 1 ng of genomic DNA at 56°C are showed. These data are produced using from Rotorgene® software. (C) The size of the PCR product is checked on Bioanalyzer. The size of this obtained fragment is indicated by an arrow. The criteria for acceptance criteria is ± 10 pb of a difference between the obtained size and expected size. (D) Sequencing of PCR products with forward and reverse primer was performed by GATC® company. The alignment of sequences obtained after sequencing on the sequence of interest locus is built using Geneious® software. In dark blue and grey, show the sequence of interest locus and the region of the same locus, respectively, are indicated. The position of primers is represented in green (F1 and R1 for forward and reverse respectively). The location of HERV-V3 probesets is indicated in blue and pink (antisense and sense respectively). Colored bases show differences between sequences, N in sequences represents an undetermined base. (E) The list of all validated loci in RT-qPCR in order to

compare results between chip and RT-qPCR technologies and 2 systems that exhibit conflicting profiles between microarray and RT-qPCR. The associated Primers reverse and forward associated is primers are also indicated as well as the Tm and the length of experimental amplicon.

Figure S6: RT-qPCR validation of 22 microarray-based identified HERV/MaLR elements differentially expressed according to immune state.

Comparative median intensity of individual HERV/MaLR loci (A) and HERVs/MALRs loci exhibit conflicting profiles (B), obtained using microarray and RT-qPCR (log2 scale). Intensity level obtained on HERV-V3 microarray according to immune state presented by boxplot (first column). RT-qPCR expression was measured in technical duplicate using mRNA derived from septic shock patients and healthy volunteers samples used in the discovery chip experiment (second column). Statistically significant differences between two conditions are marked (wilkoxon signed rank test, **: p-value < 0.05 and *: pvalue < 0.1).

Table S1: Detection of the HERV transcriptome in PBMCs.

"Tabs_summary_elmt_new" tab: a Number of distinct genomic HERV loci included in the HERV-gDB4 database (cf. "Proto versus Dfam" tab. From this database, a first set of 45,374 highly informative probesets was previously designed, those probesets being located where possible in U3, R, U5 subdomains of solo, 5' and 3' individual LTRs and gag/pol/env regions. This set of 29,083 elements belonging to the so-called "HERV_prototypes" repertoire (see "Proto versus Dfam" tab below), which includes 42 groups. A second set of 283,641 probesets was designed, those probesets being located in the first third and last third of the complete LTR, and every 2.5 kb in the region between LTRs, containing unannotated *gag pol env* genes and collectively labelled as int in Repbase. This set of 168,278 elements classified as the "HERV_Dfam" repertoire (see "Proto versus Dfam" tab below). A third set of 311,286 probesets corresponds to 227,174 elements classified as the "MaLR_Dfam" repertoire. A slight difference in the HERV-gDB4 element count compared to Becker J. [16] is due to

824 database curation. b Number of distinct genomic HERV loci represented on the HERV-V3
825 chip. Differences between database and chip reflect the success in designing HERV-specific
826 probes using the PEHM algorithm [16]. Probes are ultimately assembled into probesets to
827 discriminate between individual genomic HERV and MaLR sequences. The difference in
828 count with respect to Becker J. et al. [16] is due to database curation. c Number of active
829 elements in whole blood samples of septic shock patients or healthy volunteers. After the
830 experiments were normalised using the RMA method and an arbitrary positive threshold was
831 applied (value = $2^{5.5}$), elements that are active in at least 14 out of 45 samples are
832 enumerated. d One element can be composed of several probesets. e Subsets of complete
833 or partial proviruses. f When a single element is labelled with annotations from distinct
834 groups, the chimeric loci are enumerated in the “chimeric” column. All elements of the
835 HERVFc1 group supporting annotation from several groups were thus included in the
836 “chimeric” column. Nevertheless, probesets targeting HERVFc1-labelled regions were used
837 to reflect HERVFc1 group transcriptomic activity (cf. Figure 1).

838 "Proto versus Dfam" tab: this drawing represents schematically the dual strategy used to
839 construct the HERV-gDB4 database (described in [16]). In brief, on one hand, HERV
840 elements were retrieved from the human genome using prototype sequences (the locus or
841 set of loci which maintain the largest open reading frames for *gag pol env* genes within a
842 proviral structure flanked by two complete LTR sequences; see below) and similarity search
843 and, on the other hand, HERV and MaLR elements were reconstructed from highly
844 fragmented information (based on the Hidden Markov Model) contained in the DFAM
845 database. A redundancy removal step discarded the HERV elements collected and
846 annotated with the similarity search strategy from the DFAM subset. As the result of the
847 juxtaposition of adjacent elements, DFAM-derived elements were named according to the
848 name of the largest region. Hence, some HERV_DFAM elements share the same group
849 name with “HERV_prototypes”; nevertheless, because of the inherently poor information, and
850 as it represents a minority of elements with regard to prototype group content (less than 5%),

such elements were not pooled within the prototype groups and examined collectively in the DFAM repertoire.

"Proto-Class I", "Proto-Class II, "Proto-Class III" tabs: coordinates of the prototype sequences used to collect HERV using the similarity search, defining each prototype group name and allowing high-level annotation corresponding to 5'LTR U3, R and U5 domains, *gag, pol and env* genes, and 3'LTR U3, R and U5 domains.

"Aliases" tab: list of the common names of HERVs used in this and other publications and their corresponding DFAM/Repbase names

Table S2: Differential expression of HERV-V3 chip repertoires induced by immune states changes in whole blood samples.

The results of differential expression analysis from microarray are shown in this table. Four tabs are exhibited: "Prototypes", including the probesets from Prototype repertoire; "Dfam", including the probesets from "HERV and MaLR Dfam" repertoires; "Genes", including the probesets from HTA, U133 and Opti repertoires; and "all" including all probesets from filtered dataset. We found in each tab, the columns are the probeset names, its corresponding repertoire, the 3rd column is specific for each tab, genomic coordinates (human genome version: GRCh38), average expression, Fisher test statistic, p-value, adjusted p-value, log2 fold change for all comparisons between D1 vs HV, D3 vs HV, D3 vs D1, D1low vs HV, D1normal vs HV, D3low vs HV, D3normal vs HV, D1low vs D1normal, D3low vs D3normal, D1low vs D3low and D1normal vs D3normal, and the last three columns corresponding to the adjusted p-value for the all comparisons set out above. In "Prototypes", the 3rd column hervgdb4 group name corresponds to the common name given to HERV/MaLR groups as defined in Table S1; in "Dfam", it corresponds to the group name with RepBase nomenclature given in hervgdb4 database; and in "Genes", it corresponds to the alias of gene.

For the differential expression study, data were probeset was differentially expressed if the: adjusted p-value < 0.05 or < 0.1 for HERVs/MaLRs projection and, the |log2FC| > 1 for the corresponding comparison.

Table S3: Ingenuity pathways analysis results

The tables represents canonical pathways identified using Ingenuity Pathways Analysis tool (www.ingenuity.com) and signals derived from HTA probesets contained on the HERV-V3 chip. Canonical pathways predicted to be significantly activated (orange) or inhibited (blue) between septic shock patients at day 3 vs healthy volunteers (left) and septic shock patients at day 3 with a normal expression of mHLA-DR vs with a low expression of mHLA-DR are depicted (z-score ≥ 2 and z-score ≤ 2; p-value cut off of 0.05, Fisher's exact test).

Table S4: Patient characteristics according to cluster in projection analysis with the large 193 HERVs/MaLRs molecular signature

Patient characteristics at admission and outcomes according to cluster of the projection analysis. Categorical variables are expressed as n (%) and continuous variables as median [Q1-Q3]. Comparisons between two clusters were performed with Chi square test for qualitative variables and Mann-Whitney or t tests for quantitative variables, as appropriate. Values in bold indicate significance at p<0.05, just significance values are represented in this table. SOFA: Sequential Organ Failure Assessment. CNRQ: Calibrated Normalized Relative Quantities.

Table S5: Patient characteristics according to cluster in projection analysis with the reduced 10 HERVs/MaLRs molecular signature

Patient characteristics at admission and outcomes according to cluster of the projection analysis. Categorical variables are expressed as n (%) and continuous variables as median [Q1-Q3]. Comparisons between two clusters were performed with Chi-squared test for qualitative variables and Mann-Whitney for quantitative variables, as appropriate. Values in

bold indicate significance at p<0.05, only statistically significant variables are represented in this table. SOFA: Sequential Organ Failure Assessment. CNRQ: Calibrated Normalized Relative Quantities.

References:

1. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM *et al*: **The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)**. *Jama* 2016, **315**(8):801-810.
2. Marshall JC: **Why have clinical trials in sepsis failed?** *Trends in molecular medicine* 2014, **20**(4):195-203.
3. Hotchkiss RS, Monneret G, Payen D: **Sepsis-induced immunosuppression: from cellular dysfunctions to immunotherapy**. *Nature reviews Immunology* 2013, **13**(12):862-874.
4. Leentjens J, Kox M, Koch RM, Preijers F, Joosten LA, van der Hoeven JG, Netea MG, Pickkers P: **Reversal of immunoparalysis in humans in vivo: a double-blind, placebo-controlled, randomized pilot study**. *American journal of respiratory and critical care medicine* 2012, **186**(9):838-845.
5. Monneret G, Finck ME, Venet F, Debard AL, Bohe J, Bienvenu J, Lepape A: **The anti-inflammatory response dominates after septic shock: association of low monocyte HLA-DR expression and high interleukin-10 concentration**. *Immunology letters* 2004, **95**(2):193-198.
6. Monneret G, Venet F: **Monocyte HLA-DR in sepsis: shall we stop following the flow?** *Critical care* 2014, **18**(1):102.
7. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.
8. Hurst TP, Magiorkinis G: **Epigenetic Control of Human Endogenous Retrovirus Expression: Focus on Regulation of Long-Terminal Repeats (LTRs)**. *Viruses* 2017, **9**(6).
9. Andersson G, Svensson AC, Setterblad N, Rask L: **Retroelements in the human MHC class II region**. *Trends in genetics : TIG* 1998, **14**(3):109-114.
10. Chuong EB, Elde NC, Feschotte C: **Regulatory evolution of innate immunity through co-option of endogenous retroviruses**. *Science* 2016, **351**(6277):1083-1087.
11. Bolze PA, Mommert M, Mallet F: **Contribution of Syncytins and Other Endogenous Retroviral Envelopes to Human Placenta Pathologies**. *Progress in molecular biology and translational science* 2017, **145**:111-162.
12. Lamprecht B, Walter K, Kreher S, Kumar R, Hummel M, Lenze D, Kochert K, Bouhlel MA, Richter J, Soler E *et al*: **Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma**. *Nature medicine* 2010, **16**(5):571-579, 571p following 579.
13. Gimenez J, Montgiraud C, Oriol G, Pichon JP, Ruel K, Tsatsaris V, Gerbaud P, Frendo JL, Evain-Brion D, Mallet F: **Comparative methylation of ERVWE1/syncytin-1 and other human endogenous retrovirus LTRs in placenta tissues**. *DNA research : an international journal for rapid publication of reports on genes and genomes* 2009, **16**(4):195-211.
14. Gimenez J, Montgiraud C, Pichon JP, Bonnaud B, Arsac M, Ruel K, Bouton O, Mallet F: **Custom human endogenous retroviruses dedicated microarray identifies self-induced HERV-W family elements reactivated in testicular cancer upon methylation control**. *Nucleic acids research* 2010, **38**(7):2229-2246.
15. Trejbalova K, Blazkova J, Matouskova M, Kucerova D, Pecnova L, Vernerova Z, Heracek J, Hirsch I, Hejnar J: **Epigenetic regulation of transcription and splicing of syncytins, fusogenic glycoproteins of retroviral origin**. *Nucleic acids research* 2011, **39**(20):8728-8739.

16. Becker J, Perot P, Cheynet V, Oriol G, Mugnier N, Mommert M, Tabone O, Textoris J, Veyrieras JB, Mallet F: **A comprehensive hybridization model allows whole HERV transcriptome profiling using high density microarray**. *BMC genomics* 2017, **18**(1):286.
17. Mommert M, Tabone O, Oriol G, Cerrato E, Guichard A, Naville M, Fournier P, Volff JN, Pachot A, Monneret G *et al*: **LTR-retrotransposon transcriptome modulation in response to endotoxin-induced stress in PBMCs**. *BMC genomics* 2018, **19**(1):522.
18. Olivier Tabone MM, Camille Jourdan, Elisabeth Cerrato, Alexandre Pachot, Guillaume Monneret, Fabienne Venet, François Mallet, Julien Textoris: **Endogenous retroviruses transcriptional modulation after severe inflammatory injuries**.
19. Peronnet E, Venet F, Maucort-Boulch D, Friggeri A, Cour M, Argaud L, Allaouchiche B, Floccard B, Aubrun F, Rimmele T *et al*: **Association between mRNA expression of CD74 and IL10 and risk of ICU-acquired infections: a multicenter cohort study**. *Intensive care medicine* 2017, **43**(7):1013-1020.
20. Friggeri A, Cazalis MA, Pachot A, Cour M, Argaud L, Allaouchiche B, Floccard B, Schmitt Z, Martin O, Rimmele T *et al*: **Decreased CX3CR1 messenger RNA expression is an independent molecular biomarker of early and late mortality in critically ill patients**. *Critical care* 2016, **20**(1):204.
21. Landelle C, Lepape A, Voirin N, Tognet E, Venet F, Bohe J, Vanhems P, Monneret G: **Low monocyte human leukocyte antigen-DR is independently associated with nosocomial infections after septic shock**. *Intensive care medicine* 2010, **36**(11):1859-1866.
22. Monneret G, Venet F, Meisel C, Schefold JC: **Assessment of monocytic HLA-DR expression in ICU patients: analytical issues for multicentric flow cytometry studies**. *Critical care* 2010, **14**(4):432.
23. Huber W, Carey VJ, Gentleman R, Anders S: **Orchestrating high-throughput genomic analysis with Bioconductor**. 2015, **12**(2):115-121.
24. Wilson CL, Miller CJ: **Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis**. *Bioinformatics* 2005, **21**(18):3683-3685.
25. Kauffmann A, Gentleman R, Huber W: **arrayQualityMetrics--a bioconductor package for quality assessment of microarray data**. *Bioinformatics* 2009, **25**(3):415-416.
26. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics (Oxford, England)* 2003, **4**(2):249-264.
27. Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods**. *Biostatistics (Oxford, England)* 2007, **8**(1):118-127.
28. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Statistical applications in genetics and molecular biology* 2004, **3**:Article3.
29. Hochberg YBaY: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society Series B (Methodological)* 1995, **Vol. 57, No. 1 (1995), pp. 289-300**.
30. Perot P, Mullins CS, Naville M, Bressan C, Huhns M, Gock M, Kuhn F, Volff JN, Trillet-Lenoir V, Linnebacher M *et al*: **Expression of young HERV-H loci in the course of colorectal carcinoma and correlation with molecular subtypes**. *Oncotarget* 2015, **6**(37):40095-40111.
31. Ho TK: **Random Decision Forests**.
32. Fontaine M, Planel S, Peronnet E, Turrel-Davin F, Piriou V, Pachot A, Monneret G, Lepape A, Venet F: **S100A8/A9 mRNA induction in an ex vivo model of endotoxin tolerance: roles of IL-10 and IFNgamma**. *PloS one* 2014, **9**(6):e100909.
33. Monneret G, Venet F: **Sepsis-induced immune alterations monitoring by flow cytometry as a promising tool for individualized therapy**. *Cytometry Part B, Clinical cytometry* 2016, **90**(4):376-386.

34. Monneret G, Lepape A, Voirin N, Bohe J, Venet F, Debard AL, Thizy H, Bienvenu J, Gueyffier F, Vanhems P: **Persisting low monocyte human leukocyte antigen-DR expression predicts mortality in septic shock**. *Intensive care medicine* 2006, **32**(8):1175-1183.
35. Pachot A, Monneret G, Brion A, Venet F, Bohe J, Bienvenu J, Mougin B, Lepape A: **Messenger RNA expression of major histocompatibility complex class II genes in whole blood from septic shock patients**. *Critical care medicine* 2005, **33**(1):31-38; discussion 236-237.
36. Meyers L, Ginocchio CC, Faucett AN, Nolte FS, Gesteland PH, Leber A, Janowiak D, Donovan V, Dien Bard J, Spitzer S *et al*: **Automated Real-Time Collection of Pathogen-Specific Diagnostic Data: Syndromic Infectious Disease Epidemiology**. *JMIR public health and surveillance* 2018, **4**(3):e59.
37. Pachot A, Cazalis MA, Venet F, Turrel F, Faudot C, Voirin N, Diasparra J, Bourgoin N, Poitevin F, Mougin B *et al*: **Decreased expression of the fractalkine receptor CX3CR1 on circulating monocytes as new feature of sepsis-induced immunosuppression**. *Journal of immunology* 2008, **180**(9):6421-6429.
38. Sweeney TE, Shidham A, Wong HR, Khatri P: **A comprehensive time-course-based multicohort analysis of sepsis and sterile inflammation reveals a robust diagnostic gene set**. *Science translational medicine* 2015, **7**(287):287ra271.
39. Scicluna BP, van Vught LA, Zwinderman AH, Wiewel MA, Davenport EE, Burnham KL, Nurnberg P, Schultz MJ, Horn J, Cremer OL *et al*: **Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study**. *The Lancet Respiratory medicine* 2017, **5**(10):816-826.
40. Mommert Marine OT, Guy Oriol, Cerrato Elisabeth, Audrey Guichard, Paola Fournier, Alexandre Pachot, Guillaume Moneret,Fabienne Venet, Karen Brengel-Pesce, Julien Textoris, François Mallet: **LTR-retrotransposons transcriptome modulation in response to endotoxin-induced stress in PBMCs**.
41. Brudek T, Christensen T, Aagaard L, Petersen T, Hansen HJ, Moller-Larsen A: **B cells and monocytes from patients with active multiple sclerosis exhibit increased surface expression of both HERV-H Env and HERV-W Env, accompanied by increased seroreactivity**. *Retrovirology* 2009, **6**:104.
42. de Parseval N, Lazar V, Casella JF, Benit L, Heidmann T: **Survey of Human Genes of Retroviral Origin: Identification and Transcriptome of the Genes with Coding Capacity for Complete Envelope Proteins**. *Journal of virology* 2003, **77**(19):10414-10422.
43. Balestrieri E, Cipriani C, Matteucci C, Capodicasa N, Pilika A, Korca I, Sorrentino R, Argaw-Denboba A, Bucci I, Miele MT *et al*: **Transcriptional activity of human endogenous retrovirus in Albanian children with autism spectrum disorders**. *The new microbiologica* 2016, **39**(3):228-231.
44. Johnston JB, Silva C, Holden J, Warren KG, Clark AW, Power C: **Monocyte activation and differentiation augment human endogenous retrovirus expression: Implications for inflammatory brain diseases**. *Annals of neurology* 2001, **50**(4):434-442.
45. Antony JM, Deslauriers AM, Bhat RK, Ellestad KK, Power C: **Human endogenous retroviruses and multiple sclerosis: innocent bystanders or disease determinants?** *Biochimica et biophysica acta* 2011, **1812**(2):162-176.
46. Mameli G, Astone V, Arru G, Marconi S, Lovato L, Serra C, Sotgiu S, Bonetti B, Dolei A: **Brains and peripheral blood mononuclear cells of multiple sclerosis (MS) patients hyperexpress MS-associated retrovirus/HERV-W endogenous retrovirus, but not Human herpesvirus 6**. *The Journal of general virology* 2007, **88**(Pt 1):264-274.
47. Kowalczyk MJ, Danczak-Pazdrowska A, Szramka-Pawlak B, Zaba R, Osmola-Mankowska A, Silny W: **Human endogenous retroviruses and chosen disease parameters in morphea**. *Postepy dermatologii i alergologii* 2017, **34**(1):47-51.
48. Zhang B, Xing X, Li J, Lowdon RF, Zhou Y, Lin N, Zhang B, Sundaram V, Chiappinelli KB, Hagemann IS *et al*: **Comparative DNA methylome analysis of endometrial carcinoma**

reveals complex and distinct deregulation of cancer promoters and enhancers. *BMC genomics* 2014, **15**:868.

49. Saeed S, Quintin J, Kerstens HH, Rao NA, Aghajanirefah A, Matarese F, Cheng SC, Ratter J, Berentsen K, van der Ent MA *et al*: **Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity**. *Science* 2014, **345**(6204):1251086.
50. Doolittle WF, Sapienza C: **Selfish genes, the phenotype paradigm and genome evolution**. *Nature* 1980, **284**(5757):601-603.
51. Maksakova IA, Mager DL, Reiss D: **Keeping active endogenous retroviral-like elements in check: the epigenetic perspective**. *Cellular and molecular life sciences : CMLS* 2008, **65**(21):3329-3347.
52. Leung DC, Lorincz MC: **Silencing of endogenous retroviruses: when and why do histone marks predominate?** *Trends in biochemical sciences* 2012, **37**(4):127-133.
53. Perot P, Mugnier N, Montgiraud C, Gimenez J, Jaillard M, Bonnaud B, Mallet F: **Microarray-based sketches of the HERV transcriptome landscape**. *PloS one* 2012, **7**(6):e40194.
54. Allantaz-Frager F, Turrel-Davin F, Venet F, Monnin C, De Saint Jean A, Barbalat V, Cerrato E, Pachot A, Lepape A, Monneret G: **Identification of biomarkers of response to IFNg during endotoxin tolerance: application to septic shock**. *PloS one* 2013, **8**(7):e68218.
55. Lee YJ, Jeong BH, Park JB, Kwon HJ, Kim YS, Kwak IS: **The prevalence of human endogenous retroviruses in the plasma of major burn patients**. *Burns : journal of the International Society for Burn Injuries* 2013, **39**(6):1200-1205.
56. Lee KH, Rah H, Green T, Lee YK, Lim D, Nemzek J, Wahl W, Greenhalgh D, Cho K: **Divergent and dynamic activity of endogenous retroviruses in burn patients and their inflammatory potential**. *Experimental and molecular pathology* 2014, **96**(2):178-187.
57. Cho K, Chiu S, Lee YK, Greenhalgh D, Nemzek J: **Experimental polymicrobial peritonitis-associated transcriptional regulation of murine endogenous retroviruses**. *Shock* 2009, **32**(2):147-158.
58. Rol ML, Venet F, Rimmele T, Moucadel V, Cortez P, Quemeneur L, Gardiner D, Griffiths A, Pachot A, Textoris J *et al*: **The REAnimation Low Immune Status Markers (REALISM) project: a protocol for broad characterisation and follow-up of injury-induced immunosuppression in intensive care unit (ICU) critically ill patients**. *BMJ open* 2017, **7**(6):e015734.

Figure 1

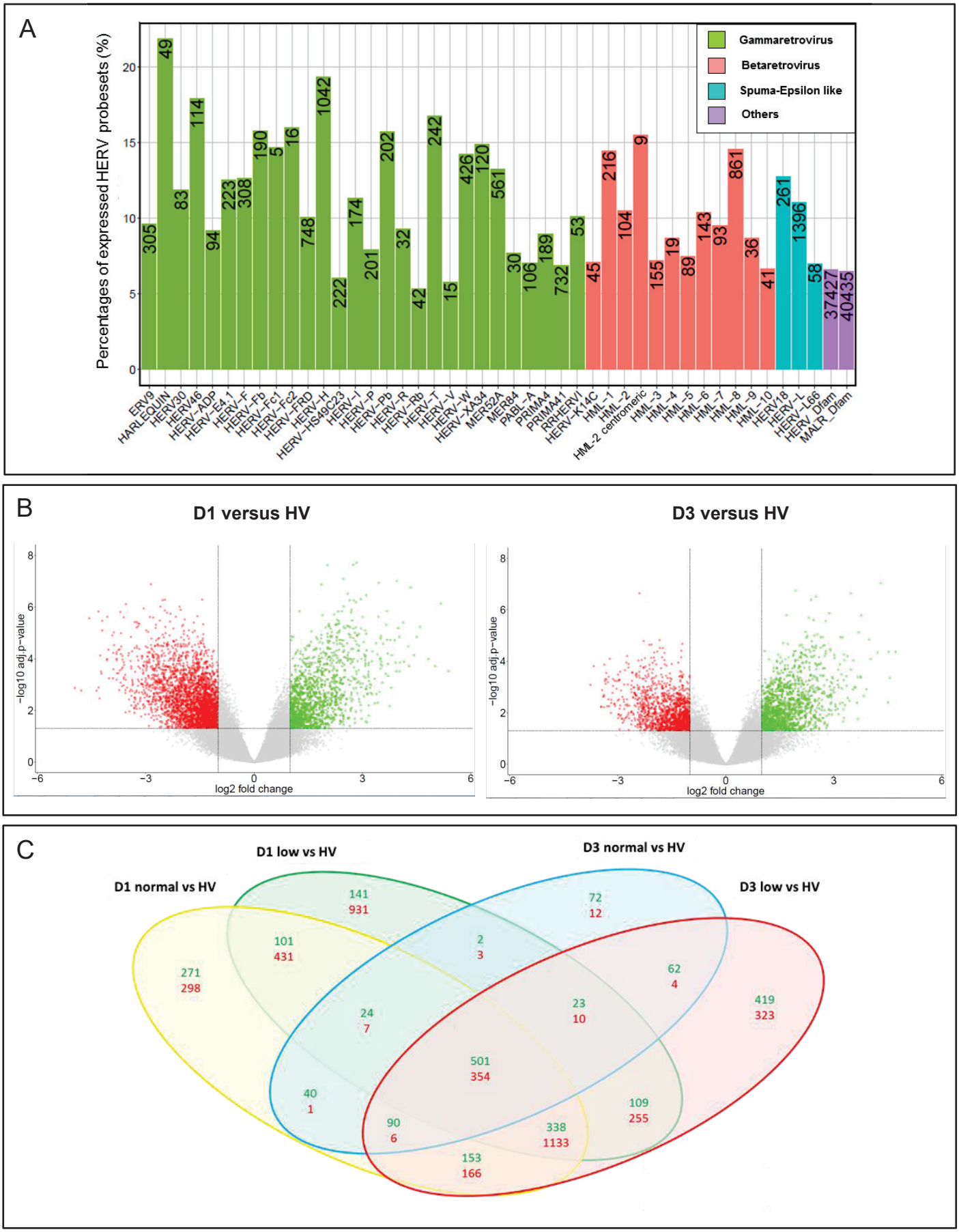
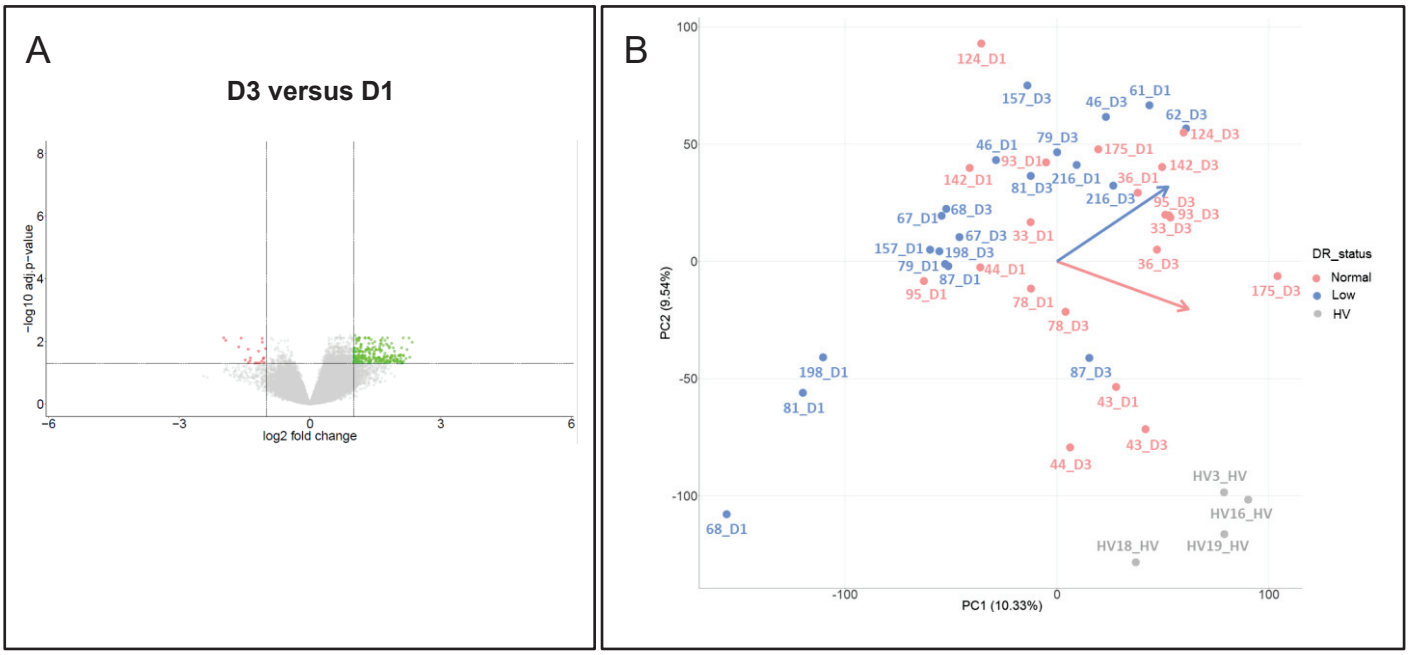
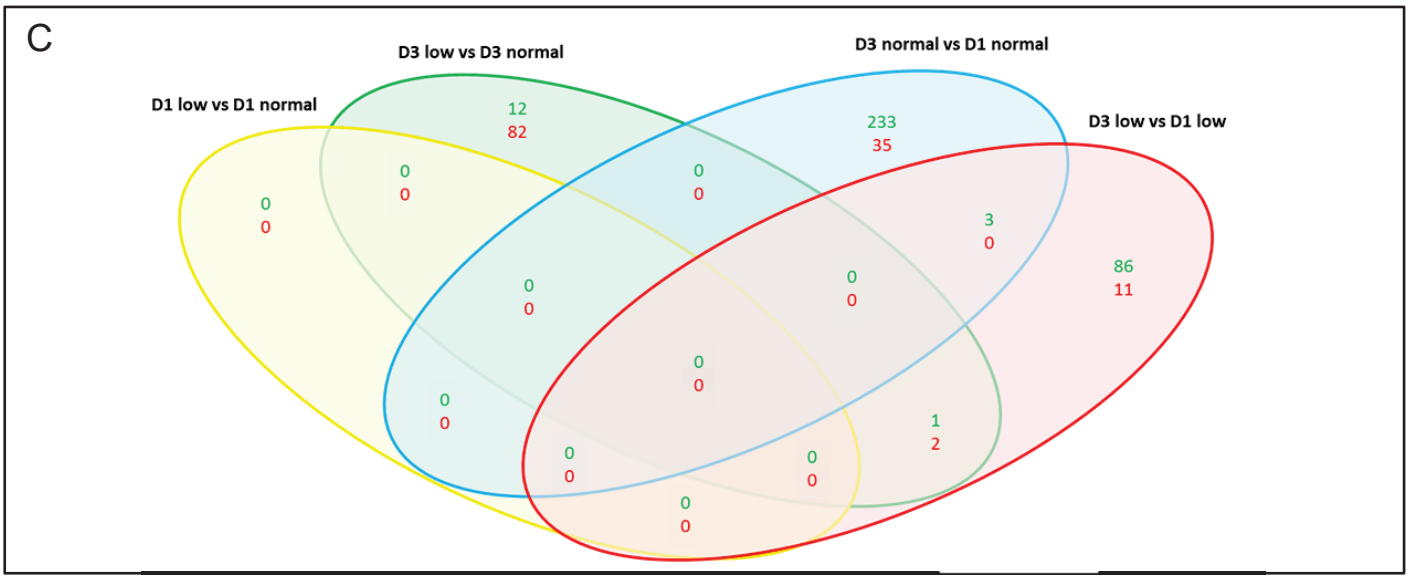

A
Percentages of expressed HERV probesets (%)
Gammaretrovirus
Betaretrovirus
Spuma-Epsilon like
Others
ERV9
HARLEQUIN
HERV30
HERV46
HERV_ADP
HERV_E4_1
HERV_F
HERV_Fb
HERV_Fc1
HERV_Fc2
HERV_FRD
HERV_H
HERV_HS49C23
HERV_I
HERV_P
HERV_Pb
HERV_R
HERV_Rb
HERV_T
HERV_V
HERV_W
HERV_XA34
MER52A
MER84
PABL_A
PRIMA4
PRIMA41
RRHERVI
HERV_K14C
HML_1
HML_2
HML_2 centromeric
HML_3
HML_4
HML_5
HML_6
HML_7
HML_8
HML_9
HML_10
HERV18
HERV_L
HERV_L66
HERV_Dfam
MALR_Dfam
B
D1 versus HV
D3 versus HV
−log10 adj.p−value
log2 fold change
C
D1 low vs HV
D3 normal vs HV
D1 normal vs HV
D3 low vs HV

Figure 2





D

D3 low vs D3 normal	Down-modulated	Up-modulated
Number of HERVs/ MaLRs DELs	76	11
Top 5 highly significant modulated HERVs/MaLR DELs	MSTA *chr13:73935391-73935941*	LTR103b_Mam *chr10:15067610-15067896*
	MLT1C *chr1:150683901-150684283*	MLT1K (2) *chr5:14562791-14563322*
	MLT2C1 *chr12:9038254-9038598*	LTR40a *chr5:14551189-14551685*
	MLT2D *chr5:35867894-35868273*	MLT1C *chr5:14521067-14521516*
	MLT1L *chr7:114967110-114967600*	LTR16E2 *chr10:109173062-109173519*
Number of DEGs	47	10
Top 5 highly significant modulated DEGs	IL5RA (2)	HGF (10)
	CCL5 (13)	IL1R2 (12)
	CD3G (24)	IL18R1 (18)
	DPP4 (1)	CYP1B1 (16)
	CD40LG (2)	PPARG (3)

E

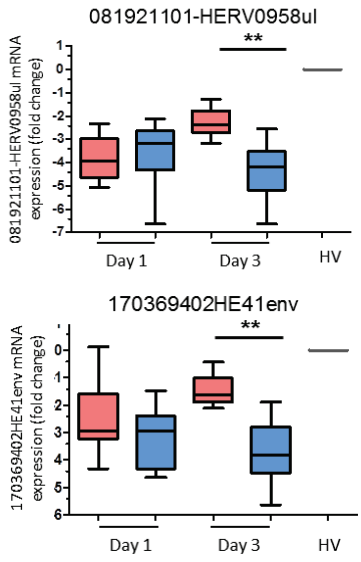


Figure 3

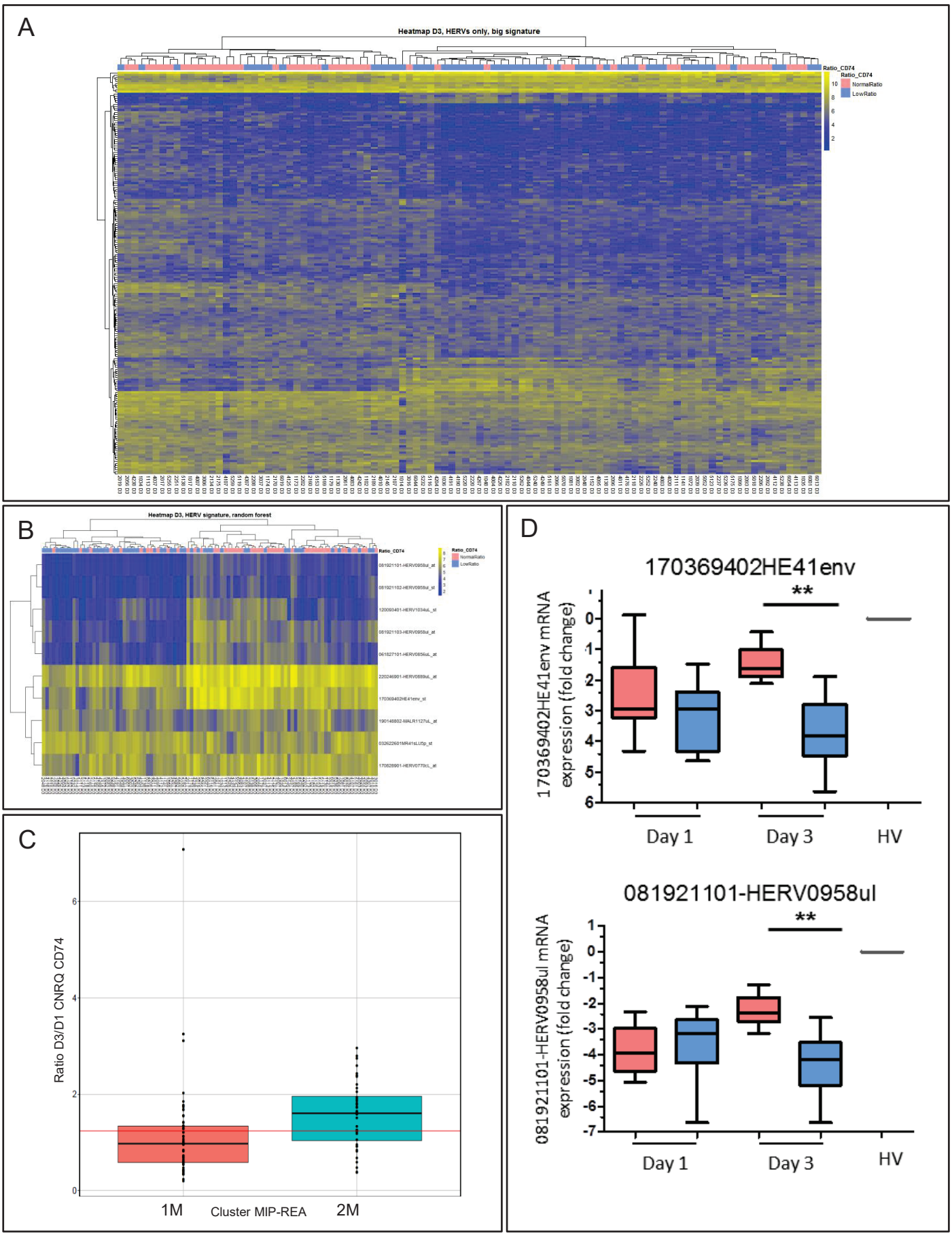

A
Heatmap D3, HERVs only, big signature
Ratio_CD74
NormalRatio
LowRatio
B
Heatmap D3, HERV signature, random forest
C
Ratio D3/D1 CNRQ CD74
1M
Cluster MIP-REA
2M
D
170369402HE41env
170369402HE41env mRNA expression (fold change)
**
Day 1
Day 3
HV
081921101-HERV0958ul
081921101-HERV0958ul mRNA expression (fold change)
**
Day 1
Day 3
HV

Figure S1

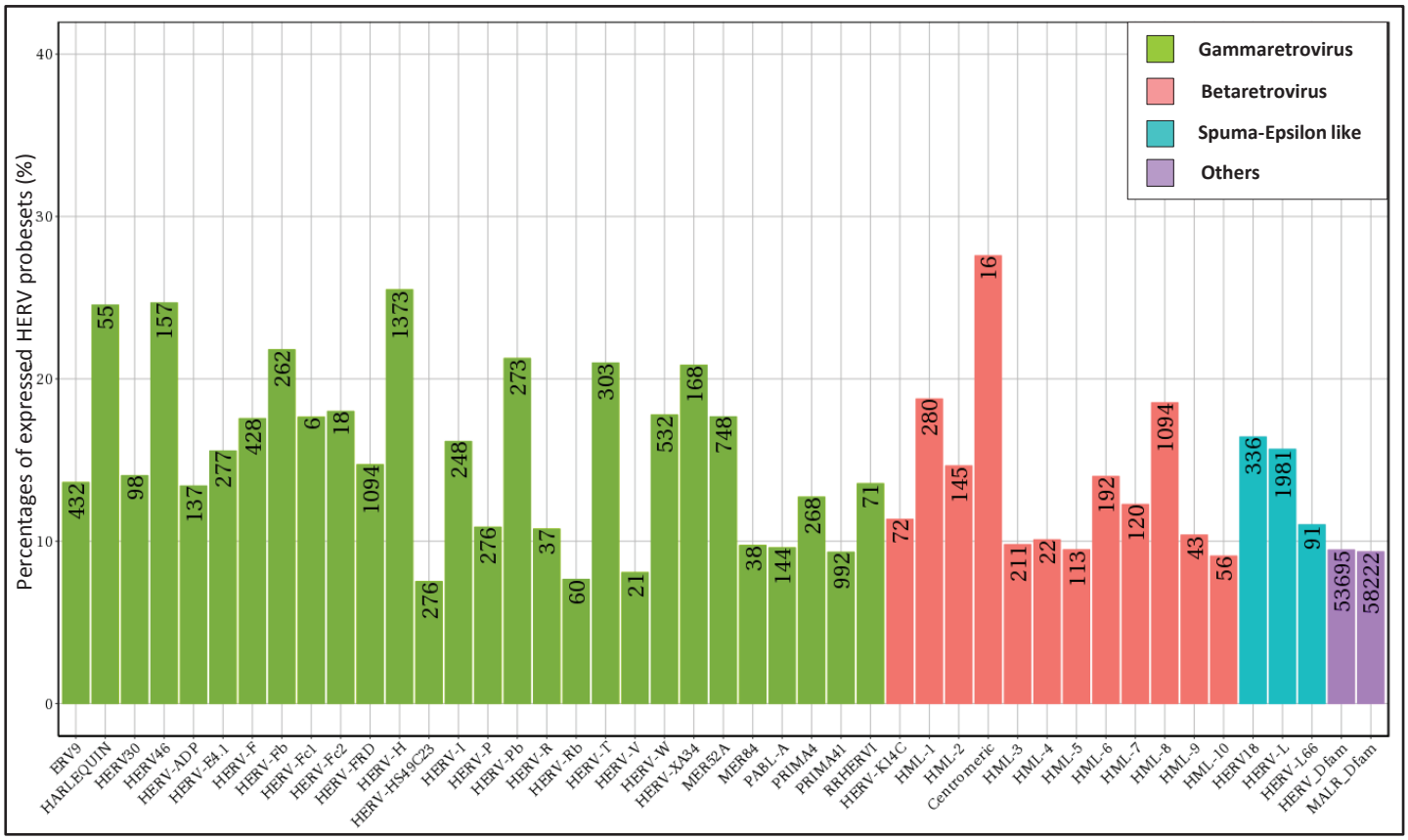
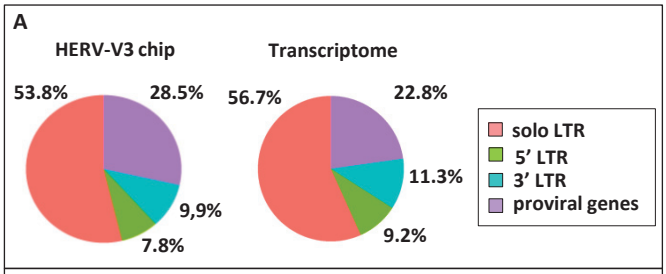

Gammaretrovirus
Betaretrovirus
Spuma-Epsilon like
Others
Percentages of expressed HERV probesets (%)
0
10
20
30
40
ERV9
HARLEQUIN
HERV30
HERV46
HERV-ADP
HERV-E4.1
HERV-F
HERV-Fb
HERV-Fc1
HERV-Fc2
HERV-FRD
HERV-H
HERV-HS49C23
HERV-I
HERV-P
HERV-Pb
HERV-R
HERV-Rb
HERV-T
HERV-V
HERV-W
HERV-XA34
MER52A
MER84
PABL-A
PRIMA4
PRIMA41
RRHERVI
HERV-K14C
HML-1
HML-2
Centromeric
HML-3
HML-4
HML-5
HML-6
HML-7
HML-8
HML-9
HML-10
HERV18
HERV-L
HERV-L66
HERV_Dfam
MALR_Dfam
432
55
98
157
137
277
428
262
6
18
1094
1373
276
248
276
273
37
60
303
21
532
168
748
38
144
268
992
71
72
280
145
16
211
22
113
192
120
1094
43
56
336
1981
91
53695
58222

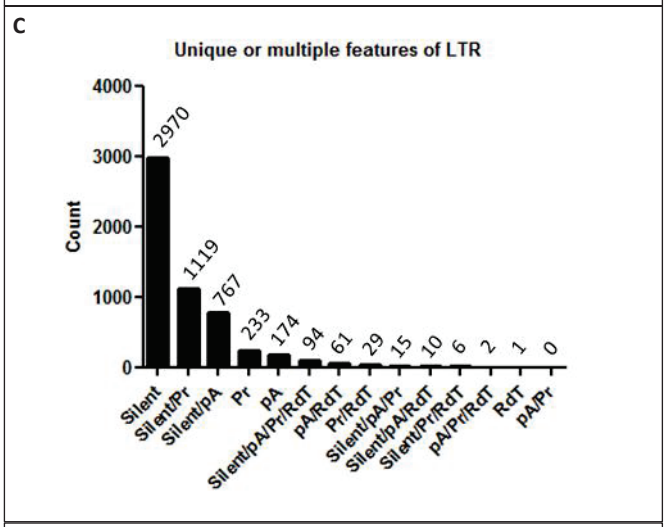
Figure S2



B

LTR features

	Solo LTR	5'LTR	3'LTR	Total (%)
Genome[a]	19 722	3817	4728	
Chip[b]	15 475	2245	2835	
Attributable LTR[c]	5430	418	646	
Attributed LTR[d]	5290	404	642	
Promoter	629	44	91	12.1%
polyAdenylation	517	48	77	10.1%
Readthrough	109	9	7	2%
Silent	4035	303	467	75.8%



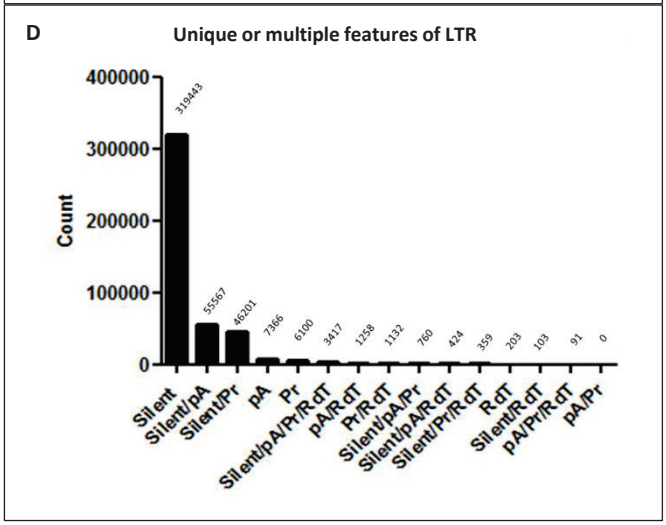


Figure S3

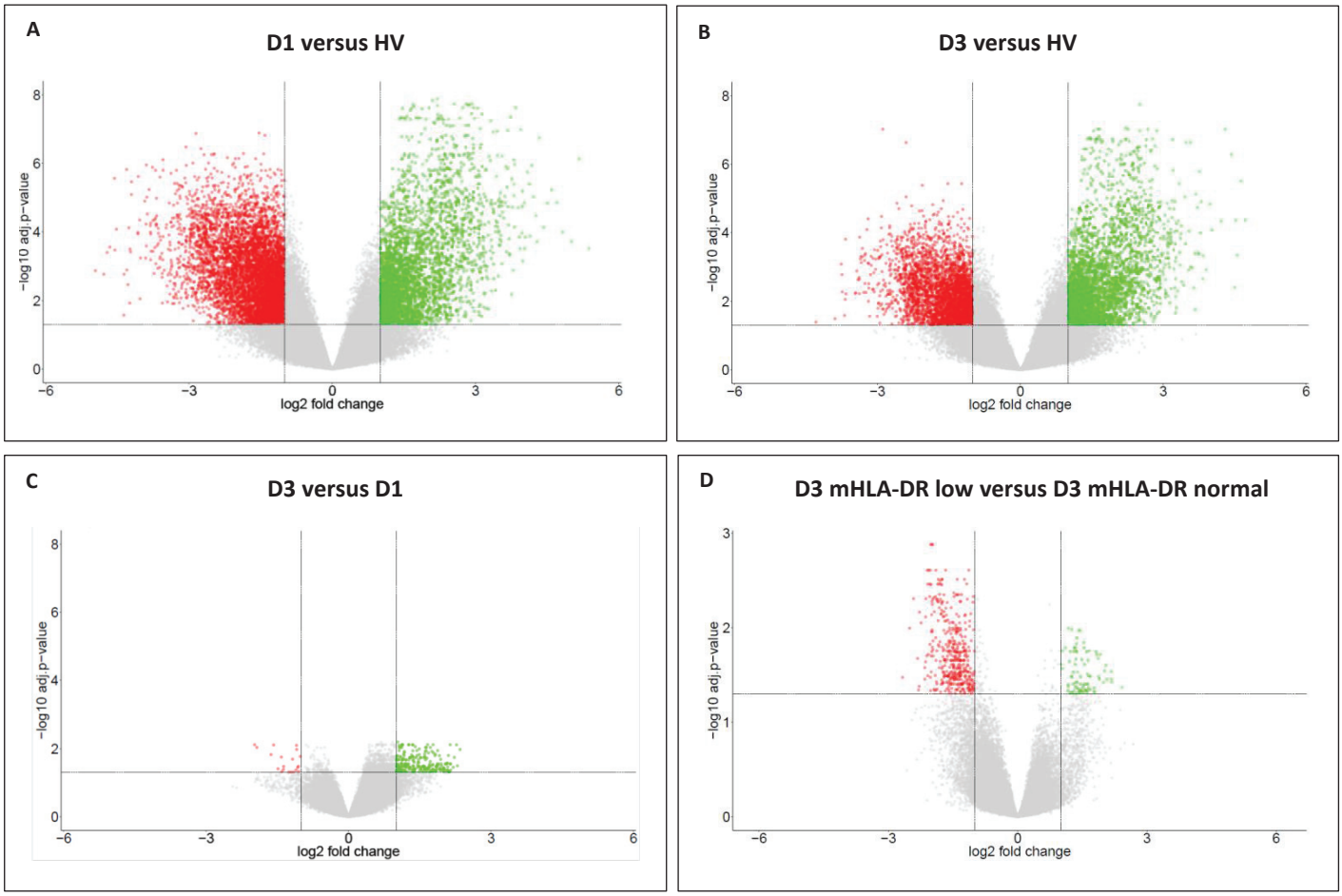
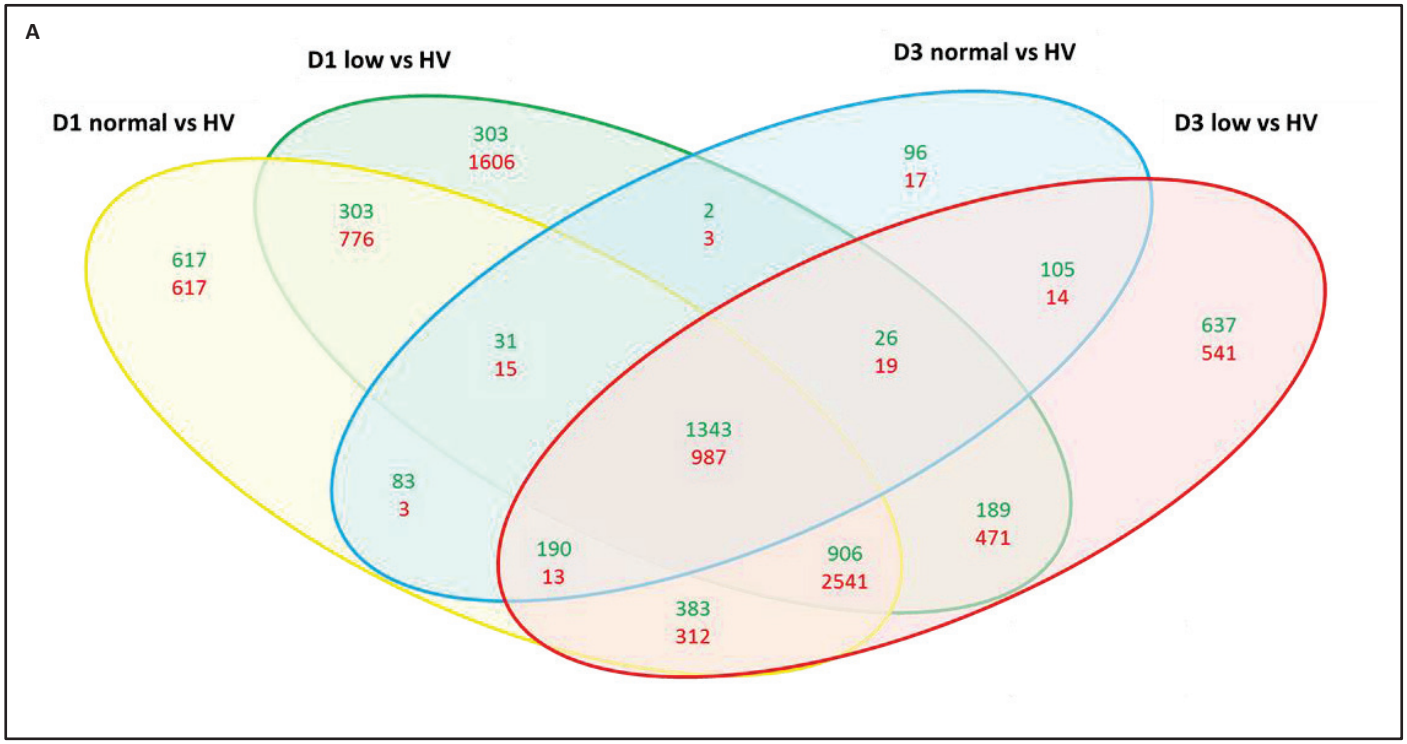

A
D1 versus HV
−log10 adj.p−value
log2 fold change
B
D3 versus HV
−log10 adj.p−value
log2 fold change
C
D3 versus D1
−log10 adj.p−value
log2 fold change
D
D3 mHLA-DR low versus D3 mHLA-DR normal
−log10 adj.p−value
log2 fold change

Figure S4


A
D1 low vs HV
D3 normal vs HV
D1 normal vs HV
D3 low vs HV
303
1606
96
17
303
776
2
3
617
617
105
14
637
541
31
15
26
19
1343
987
83
3
189
471
190
13
906
2541
383
312

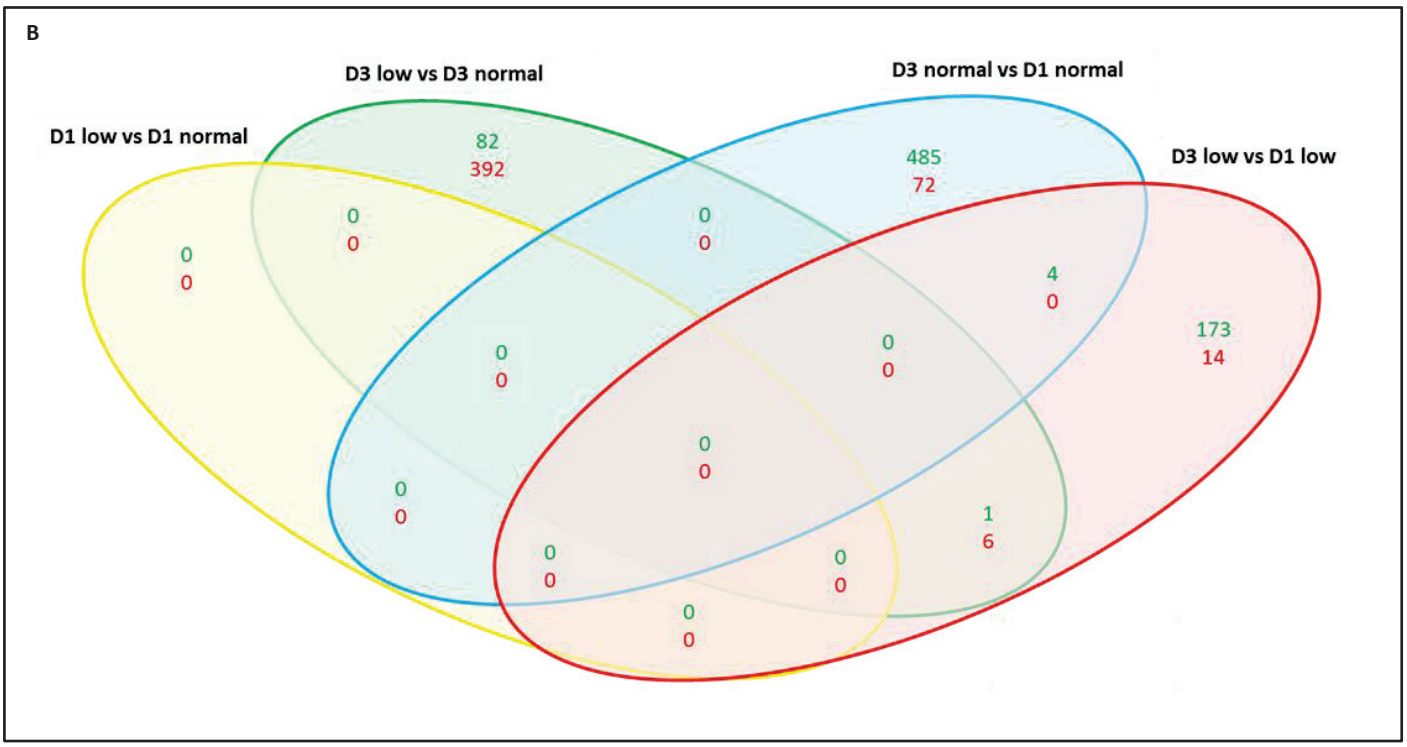
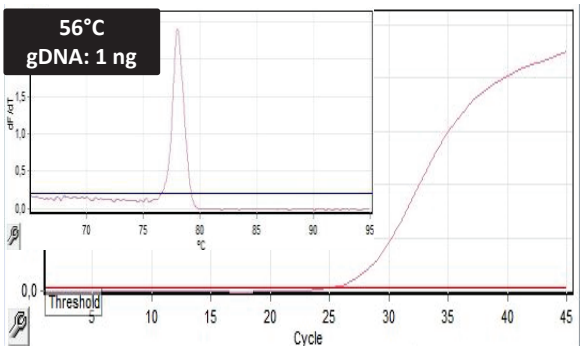

B
D3 low vs D3 normal
D3 normal vs D1 normal
D1 low vs D1 normal
D3 low vs D1 low
82
392
485
72
0
0
0
0
0
0
4
0
173
14
0
0
0
0
0
0
0
0
1
6
0
0
0
0
0
0

Figure S5

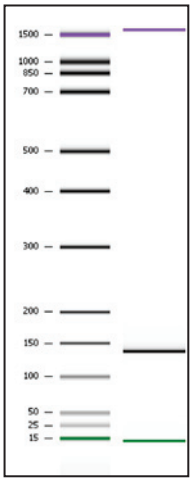
A

System	Primer forward	Primer reverse	Expected Tm	Expected size
060400302-HERV0489uL	CTGGGATACTGCACTCATTCATGT	AGAGGGTCTCAAAAGGAATCATGG	60.5	127 pb

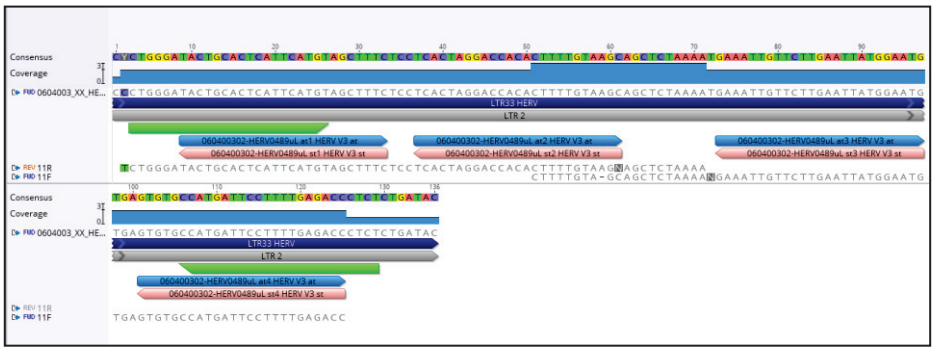
B



C



D



E

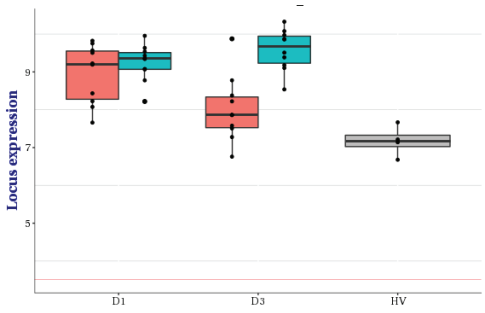
Name of probesets	Primer Forward	Primer Reverse	Tm experimental	Length amplicon experimental (pb)
021460102-HERV0599uL_st	CACACAGAAAGGGACCCTGG	TCACAGCTTGAGAATGTGGTAGG	56°C	241
042486301-MALR1015uL_st	ACAAAGCGAGGACTTAATGGCT	CACAAGCTATCTGAACCCCATGA	56°C	123
050286002-MALR1003uL_at	GACATTTAGGGCCTAGCCGG	CAGAAGATATCCACAGCCGAGT	69,3/59,6	137
050286701-HERV0513uL_at	GGGTGGCTGCATCCTATGG	CTGGTCAGGAAAAAATTTGCCTTC	56°C	153
050287402-MALR1022uL_at	ACATGACATTGTCTGAACTTTGGG	TAGGACCATGCAGATACTAGTGAC	56°C	146
050734202-MALR1019uL_st	TCTTCTGCCCACCTCACTCA	CTTGGGTTTAAATCCAGTCCGC	56°C	88
061827101-HERV0856uL_at	ACTGTGCTTGATGTGCCCA	CCATGCTTGCTCATTCACACA	56°C	127
071588201-MALR1023uL_st	TTGCTTGTTGCTCTCTGAACATTT	CACAGGCTGACCTCCAAACC	56°C	154
081921101-HERV0958ul_at	AAGTTGAACCCAAGGGTATCCA	AGACCCATCATTGCTAGCAGAC	56°C	242
081921102-HERV0958ul_st	TCACTCCTGTGGAATTAGTTCTTGT	AAAAGGATTATGCTAGGTGCTGTTG	56°C	177
081921103-HERV0958ul_at	GGTAGATGGGGCAAGTTGAAGTA	AGAGCAACAGCCTGAGGCTA	56°C	247
100223002-HERV0378uL_st	TCTCTTTCCCCATCACTACCACT	GAAGTAGTCACAGCTTCTGGCA	56°C	97
110076702-HERV0489uL_st	TTTGCTGGTACTTCCCTGTGG	GTGTCCCAGATACACTCTCCATC	56°C	118
120093401-HERV1034uL_st	GCAAGGAATTGGCTCATGGC	ATAGACAGCAGCTTCAGCAGC	56°C	111
121467702-HERV0490uL_at	TGAGTGCTTCTATTTGGGAGAGC	CTCATCGTTCATTCTACCCTCCTT	56°C	138
130037501-MALR1023uL_st	TGGCTATGGGATCCAGTTCTTG	TTCCCAGTGACAGTTTTTCCCTC	56°C	75
130360601-HERV0808cL_st	TCTGTACTGGTTGCCCCAAC	CGTGCCAGGCCTCTAATACTTTT	56°C	135
130894002-HERV0421uL_at	GCATGCTTCTGTGGCCAGA	TGGCATCCCATTCCCATTTCC	56°C	105
141107102-MALR1019uL_at	AGGGAAGACCCCAAGATGATG	CATGCAAAGTCCAACGAGAGG	56°C	111
141115102-HERV0608uL_at	GCATAGACTCACCCGTAGCAG	CATAAAGCAGGAGCCAATGCAG	56°C	110
160627301-MALR1014uL_st	TTCCTACATAACAAACAGCCCCA	GGCAGCTGAGATCCGTTGA	56°C	138
170369402HE41env_at	TGGCTCATAGGGATTCCAGACT	AGCAAGTTGTCAAGAGCCAATCT	56°C	178
170828901-HERV0770cL_at	GCCTGAGCGTTTTTATGCTGG	AGTACACAGGTCTTGCCGAGA	56°C	119
190148802-MALR1127uL_at	GACTGGTTTTGAACTCCTAGGCT	CACCCCTCCCAAATGCATATGT	56°C	183
190478501-MALR1003cL_st	GCATGTTGGTTTGTGTCTTGGT	CAGCGTGCAGTGGTGTGA	56°C	116
220246901-HERV0889uL_at	AAATGAACACCTGGGCAAACAA	TGCCTAACGTAAGCACCCATC	56°C	126
220247002-HERV0797uL_at	GCCAGAGAGGCATAATGAAGCA	GATTCTAAGCCTCCCCCTCATTT	56°C	129
032622601MR41sLU5p_st	TTGCAGGAAAGGGGTGCAG	TTACTCTATGGACTCGCCCTGA	56°C	96
100468701-MALR0820cL_st	ATAAAGGAATACCTGAGGCTGGG	CACTGAGCTCGGCCTCTTC	56°C	221

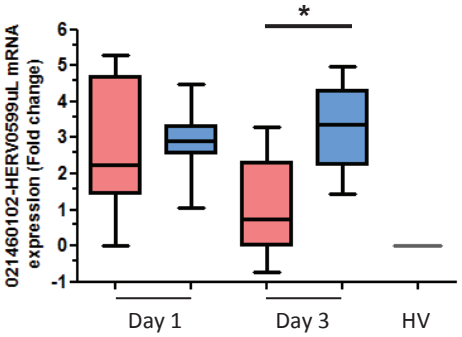
Figure S6 (1)

Corresponding profiles

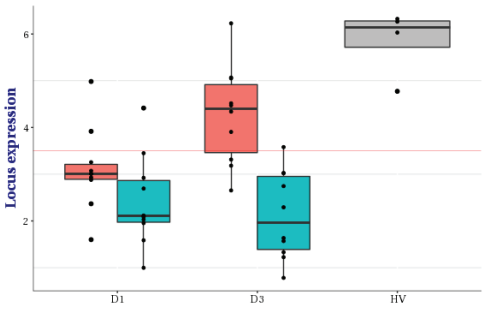
Microarray profiles **RT-qPCR profiles**

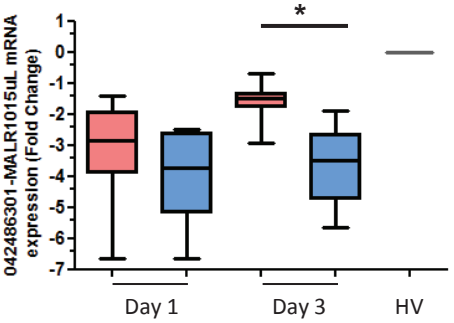
021460102-HERV0599uL_st



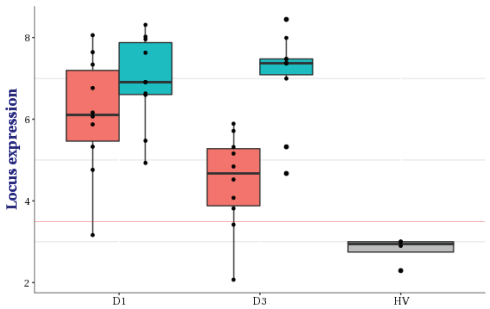


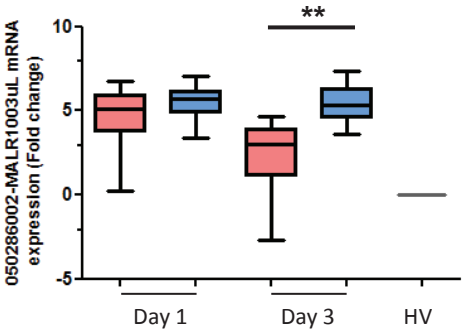
042486301-MALR1015uL_st



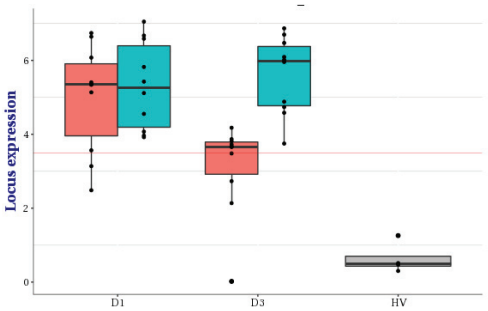


050286002-MALR1003uL_at





050286701-HERV0513uL_at



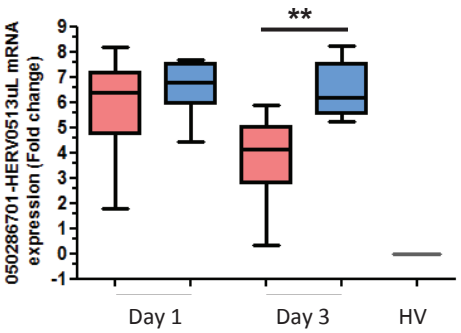
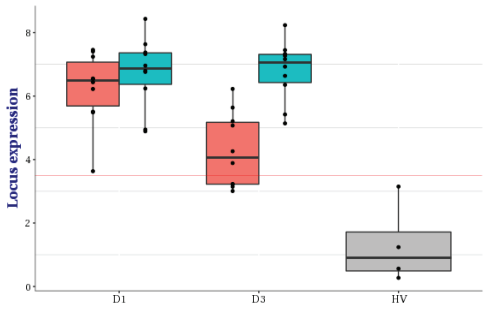


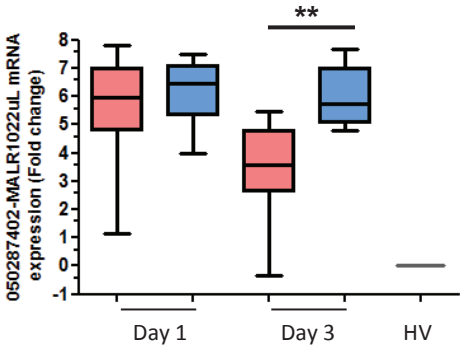
Figure S6 (2)

Microarray profiles

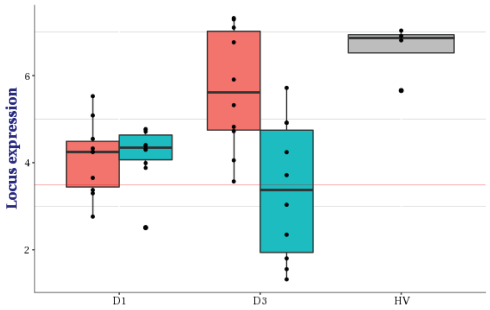
RT-qPCR profiles

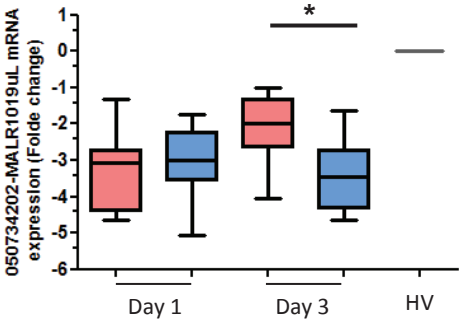
050287402-MALR1022uL_at



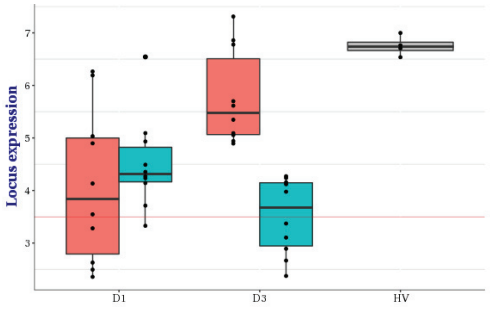


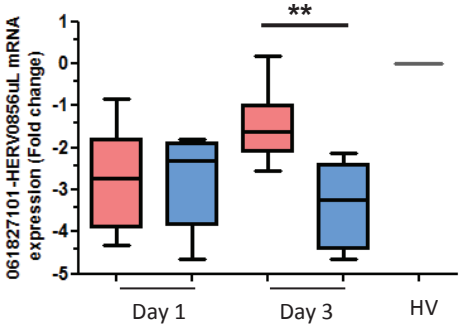
050734202-MALR1019uL_st



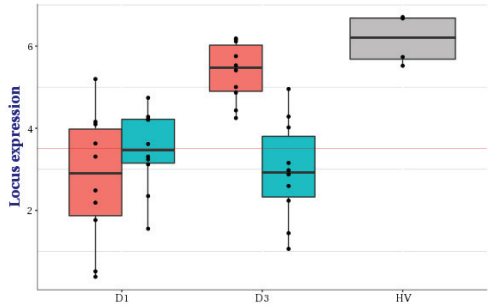


061827101-HERV0856uL_at





071588201-MALR1023uL_st



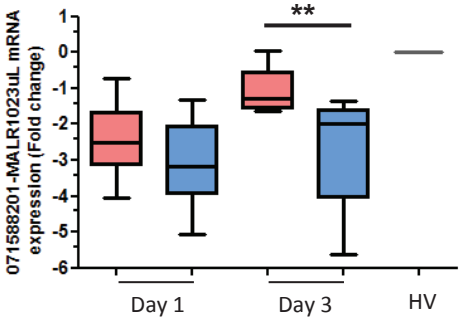
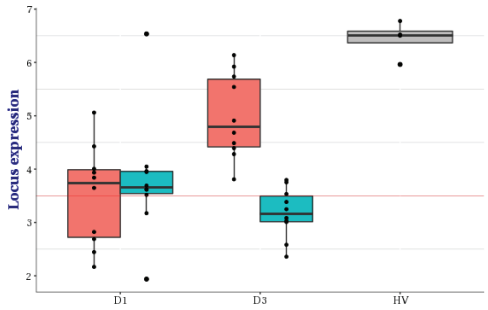
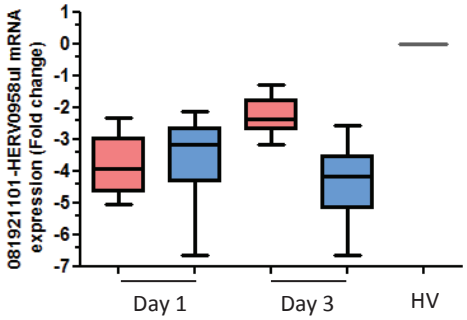


Figure S6 (3)

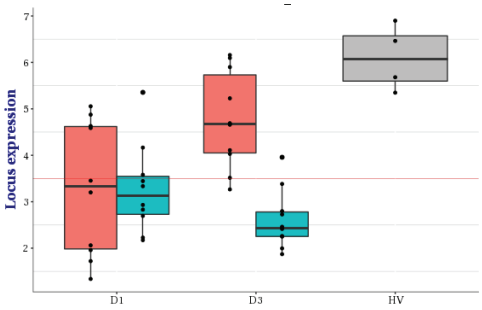
Microarray profiles | **RT-qPCR profiles**

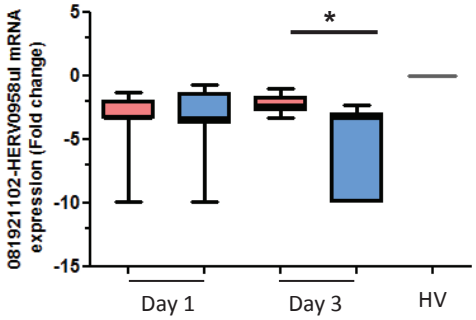
081921101-HERV0958uI_at



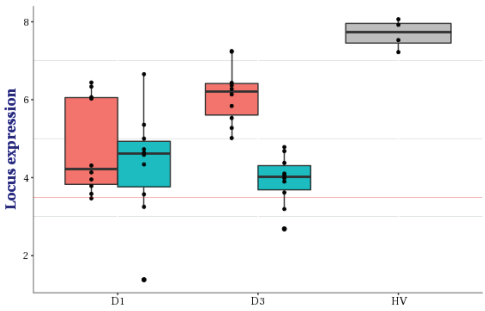


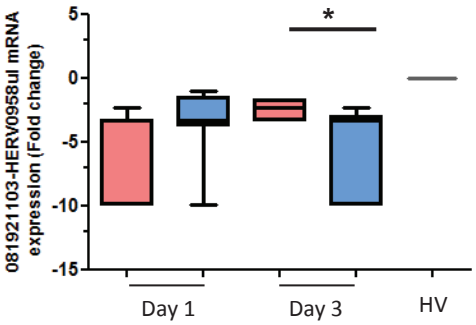
081921102-HERV0958uI_st



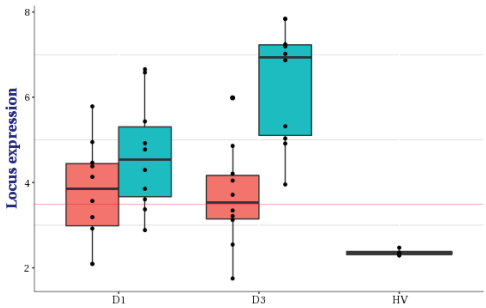


081921103-HERV0958uI_at





100223002-HERV0378uL_st



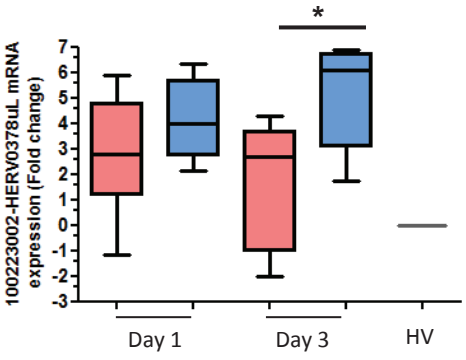
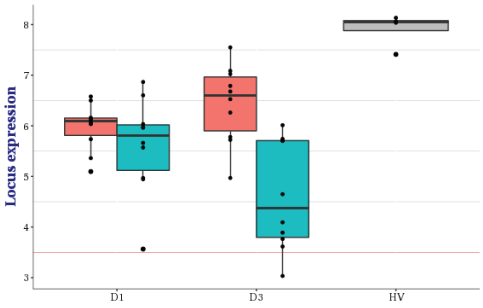
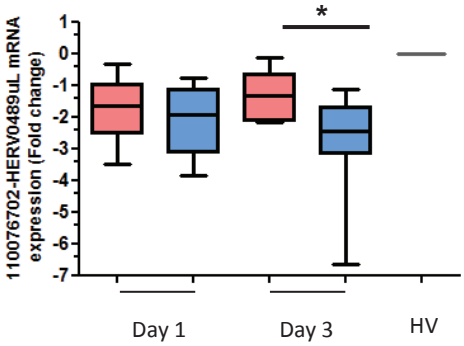


Figure S6 (4)

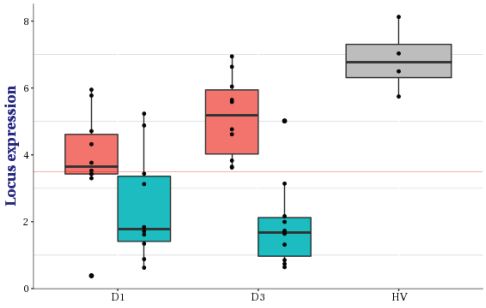
Microarray profiles | **RT-qPCR profiles**

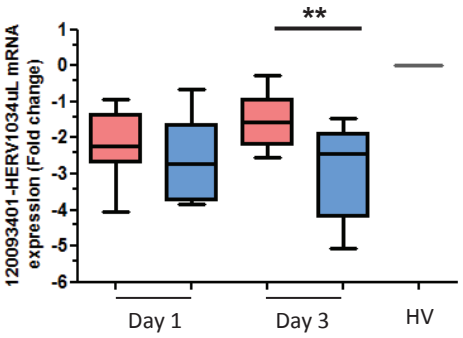
110076702-HERV0489uL_st



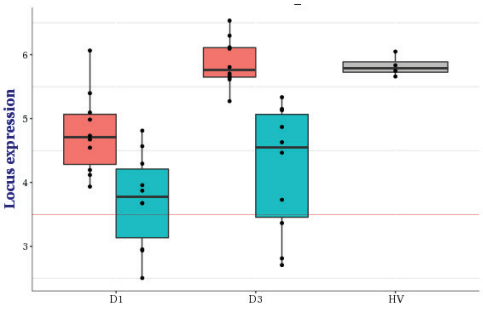


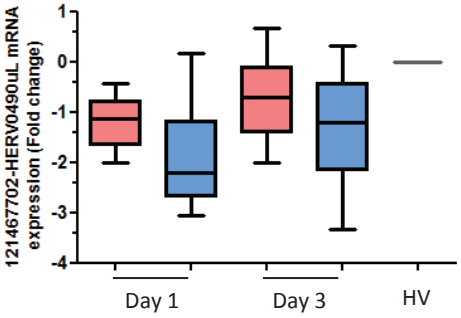
120093401-HERV1034uL_st



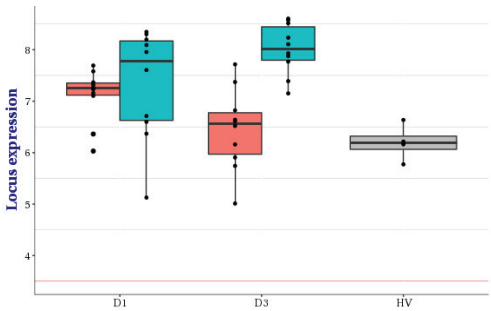


121467702-HERV0490uL_at





130037501-MALR1023uL_st



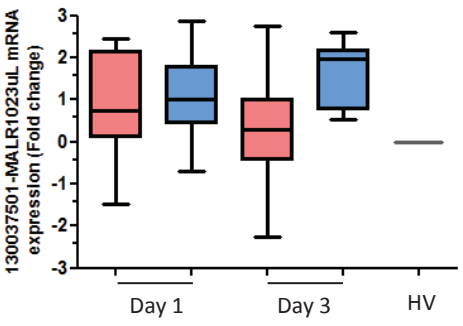
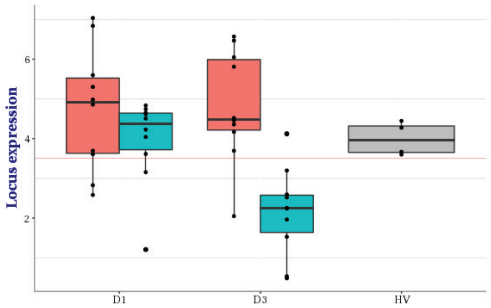
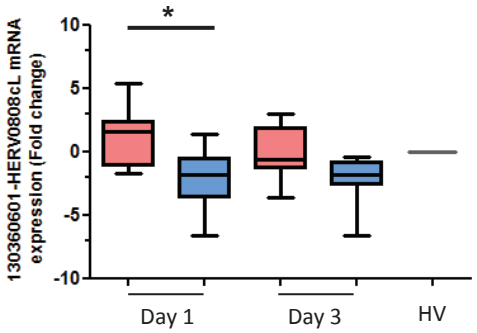


Figure S6 (5)

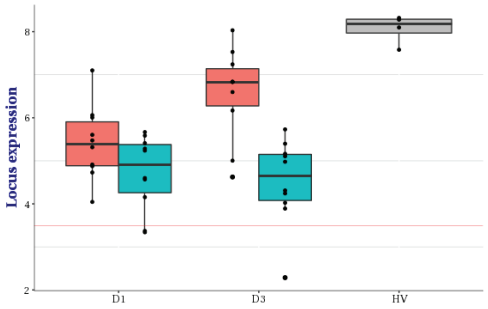
Microarray profiles | **RT-qPCR profiles**

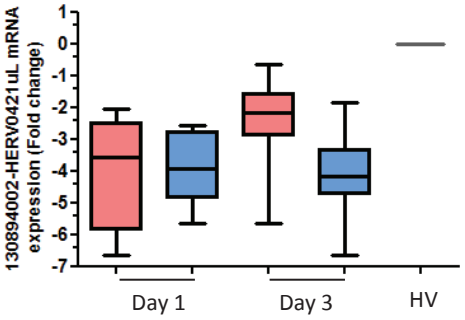
130360601-HERV0808cL_st



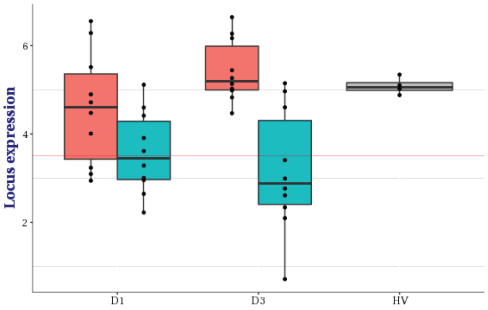


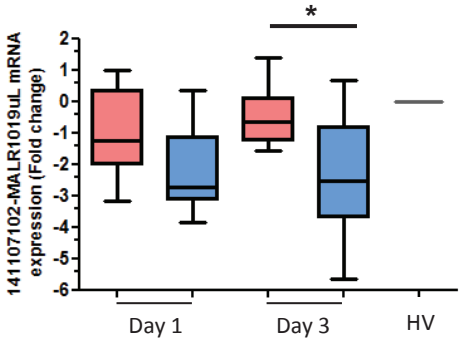
130894002-HERV0421uL_at



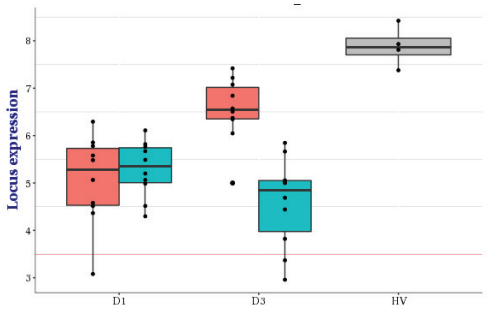


141107102-MALR1019uL_at





141115102-HERV0608uL_at



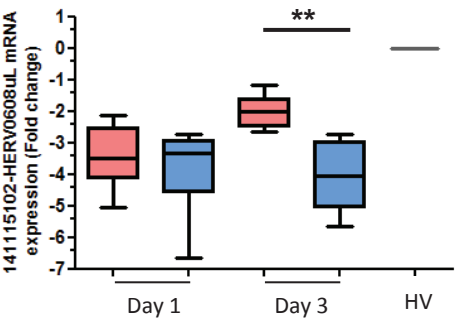
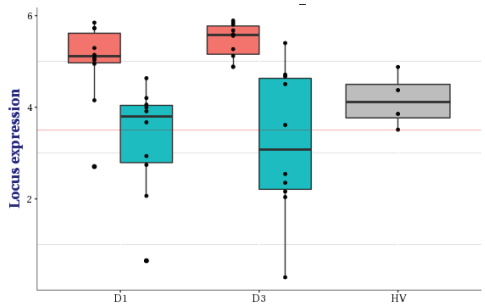


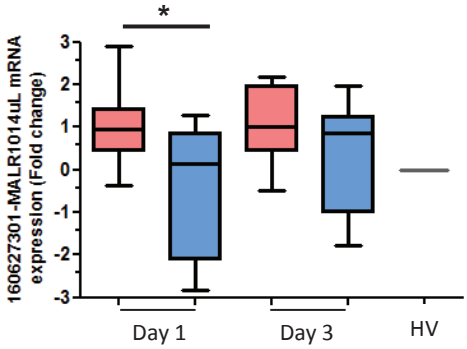
Figure S6 (6)

Microarray profiles

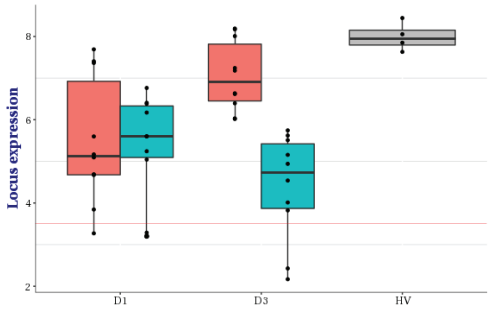
RT-qPCR profiles

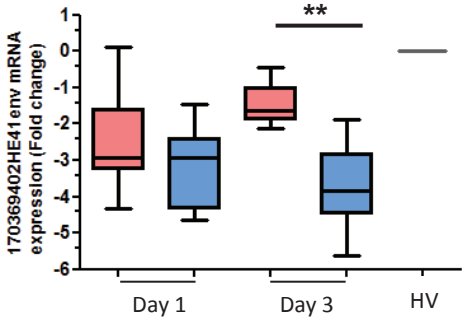
160627301-MALR1014uL_st



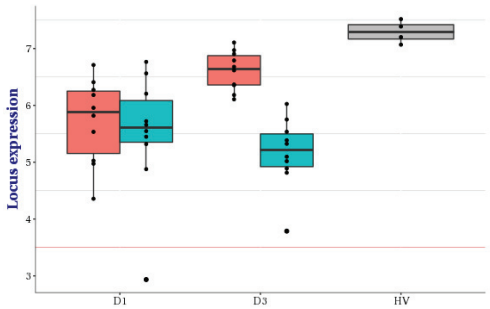


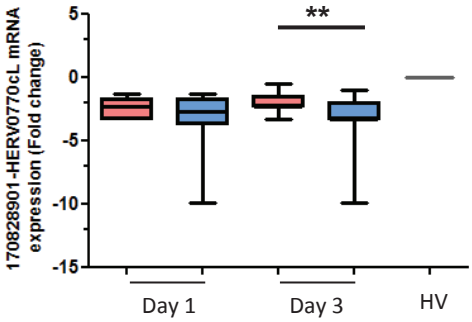
170369402HE41env_at



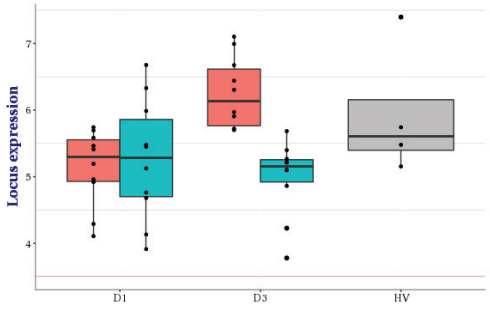


170828901-HERV0770cL_at





190148802-MALR1127uL_at



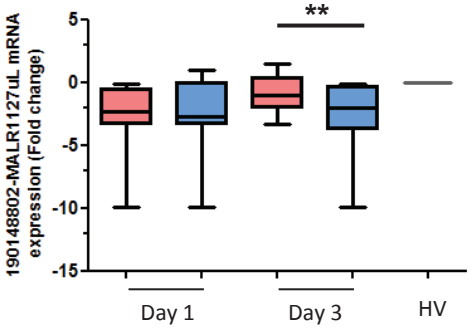
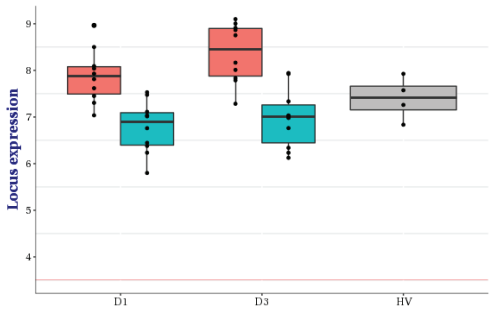


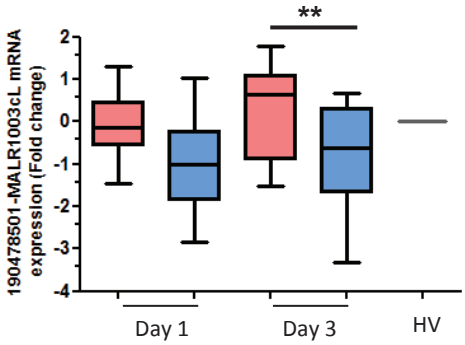
Figure S6 (7)

Microarray profiles

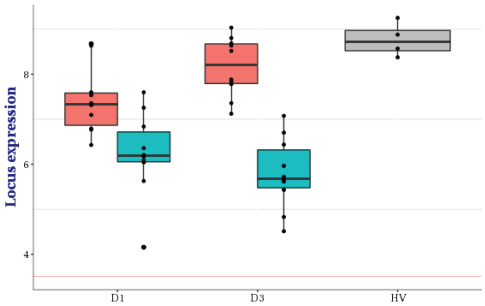
RT-qPCR profiles

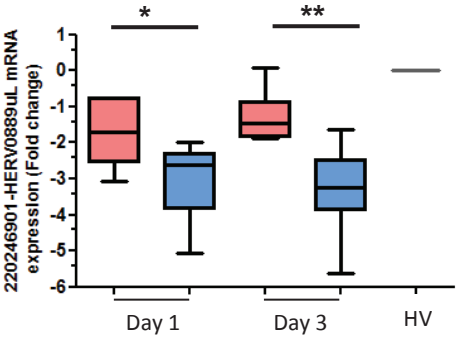
190478501-MALR1003cL_st



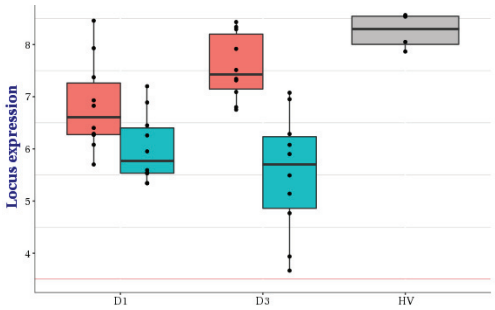


220246901-HERV0889uL_at





220247002-HERV0797uL_at



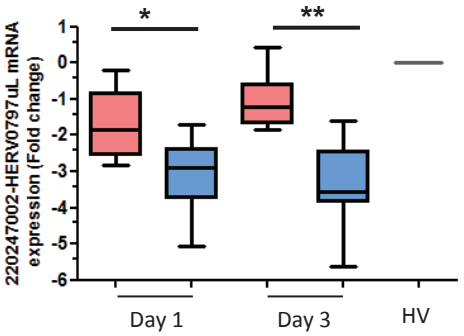


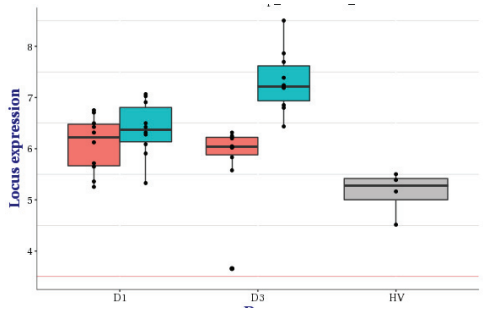
Figure S6 (8)

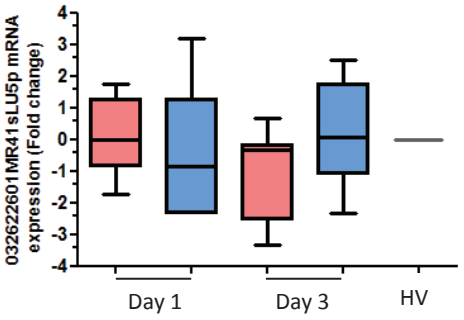
Conflicting results

Microarray profiles

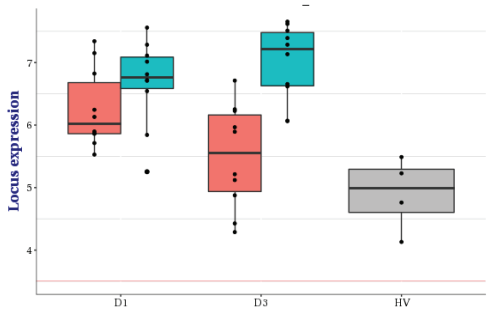
RT-qPCR profiles

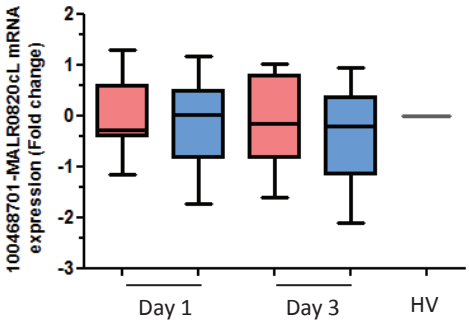
032622601MR41sLU5p_st





100468701-MALR0820cL_st





3.2.5 INTERSECTIONS DES HERV MODULES ENTRE LES JEUX DE DONNEES

Bien que les différents jeux de données analysés avec la puce HERV-V3 comportent des différences expérimentales majeures, et que les questions posées sont différentes, nous avons voulu savoir si des loci HERV sont trouvés modulés dans plusieurs jeux de données. De tels loci représenteraient de bons candidats, en tant que marqueurs de l'état d'immunodépression induite par le sepsis.

Nous avons vérifié les éventuelles intersections des HERV différentiellement exprimés dans les 3 jeux de données de la puce HERV-V3 : le modèle d'endotoxine tolérance (ET), la cohorte Immunosepsis de 20 patients en choc septique stratifiés sur le mHLA-DR (IS) et la cohorte MIP-rea de 102 patients en choc septique (MIP). Globalement, on voit que les intersections entre ET et MIP sont les plus importantes par rapport aux autres (Figure 3-15 A). En prenant les 2000 probesets les plus différentiellement exprimés dans chacun des jeux de données, l'intersection entre ET et MIP est de 164 probesets (8,2%), celle entre IS et MIP est de 105 probesets (5,3%) et l'intersection entre ET et IS est de 81 probesets (1,1%). Quand on fait l'intersection entre les 3 jeux de données, seuls 9 probesets sont en commun pour des listes de taille 2000. Ceci pourrait s'expliquer par la nature différente des échantillons des jeux de données, IS et MIP étant des échantillons de sang total provenant de patients en choc septique, alors que les échantillons de ET sont des PBMCs sur lesquels un modèle d'endotoxine tolérance a été réalisé. De plus les questions posées sont différentes, dans le modèle ET nous avons 3 conditions de stimulation différentes prises à partir des mêmes échantillons (NS, LPS, ET) que nous avons comparées entre elles, dans IS nous avons comparé des patients à statut immunitaire différent (patients à mHLA-DR bas versus normal), et dans MIP nous avons utilisé un proxy de la survenue d'infections secondaires (HAI), le ratio de CD74 entre J3 et J1. Cependant, de manière surprenante, c'est l'intersection entre ET et MIP qui est la plus importante.

En prenant les seuils habituels (abs(log2(FC)) > 1 et p-value < 0,05, Figure 3-15 B), aucun probeset n'est en commun entre les 3 jeux de données, et un nombre faible de probesets en commun entre 2 jeux de données (55/2992, 2%). On remarque également que le jeu de données MIP est celui qui comporte le moins de loci différentiellement exprimés (132 contre 2434 pour ET et 481 pour IS). Avec des seuils moins stricts (valeur absolue du fold change en

log2 supérieure à 0,5 et p-value ajustée inférieure à 0,1, Figure 3-15 C), c'est cette fois-ci IS qui comporte le moins d'éléments différentiellement exprimés (973 contre 8637 pour ET et 1104 pour MIP). On obtient cette fois 38 probesets (HERV ou gène) en commun dans les 3 jeux de données. En détail, ces éléments sont principalement constitués de 34 probesets ciblant 6 gènes différents (MMP9, IL1R2, CYP1B1, STAT4, IL18R1, CX3CR1), et aussi de 4 probesets ciblant 4 HERV/MaLR loci différents (LTR21B, MLT1E3, MSTD et MER61B). De manière intéressante, ces 4 loci possèdent le même profil d'expression sur le modèle ET (NS sous exprimé par rapport à LPS et ET), alors qu'ils peuvent avoir des profils différents dans IS (Moins exprimés chez les patients avec mHLA-DR bas que mHLA-DR haut pour 3 HERV sur les 4, données non montrées). Ces 4 éléments appartiennent à des groupes différents et sont situés sur des chromosomes différents (chr1, chr5, chr10 et chr21). La LTR21B se situent dans la 3'UTR d'un transcrit (non –codant) du gène GBP5, impliqué dans l'activation de l'inflammasome et a donc un rôle dans l'immunité innée. La LTR du groupe MER61B se situent dans un intron du gène SAMSN1, inhibiteur de la prolifération des lymphocytes B. Les 2 autres ne se situent pas à proximité de gènes codants (> 10kb) et restent néanmoins intéressants dans l'optique d'apporter une information transcriptomique nouvelle, différente de celle de gènes.

Au total, une dizaine de HERV modulés entre les différentes conditions d'au moins 2 jeux de données vont être validés dans la cohorte REALISM (NCT02638779). Le but principal de cette étude est de caractériser et de déterminer l'incidence de l'immunodépression induite par une agression chez des patients admis en réanimation (choc septiques, brulés, traumas et chirurgie lourde), dans les 2 mois suivant leur admission en unité de soin intensive. Les HERV sélectionnés pourraient participer à la caractérisation de l'état immunitaire de ces patients. (Article en annexe 2, 6.2 (Rol et al., 2017)). Un Brevet est également en cours de dépôt sur le même thème.

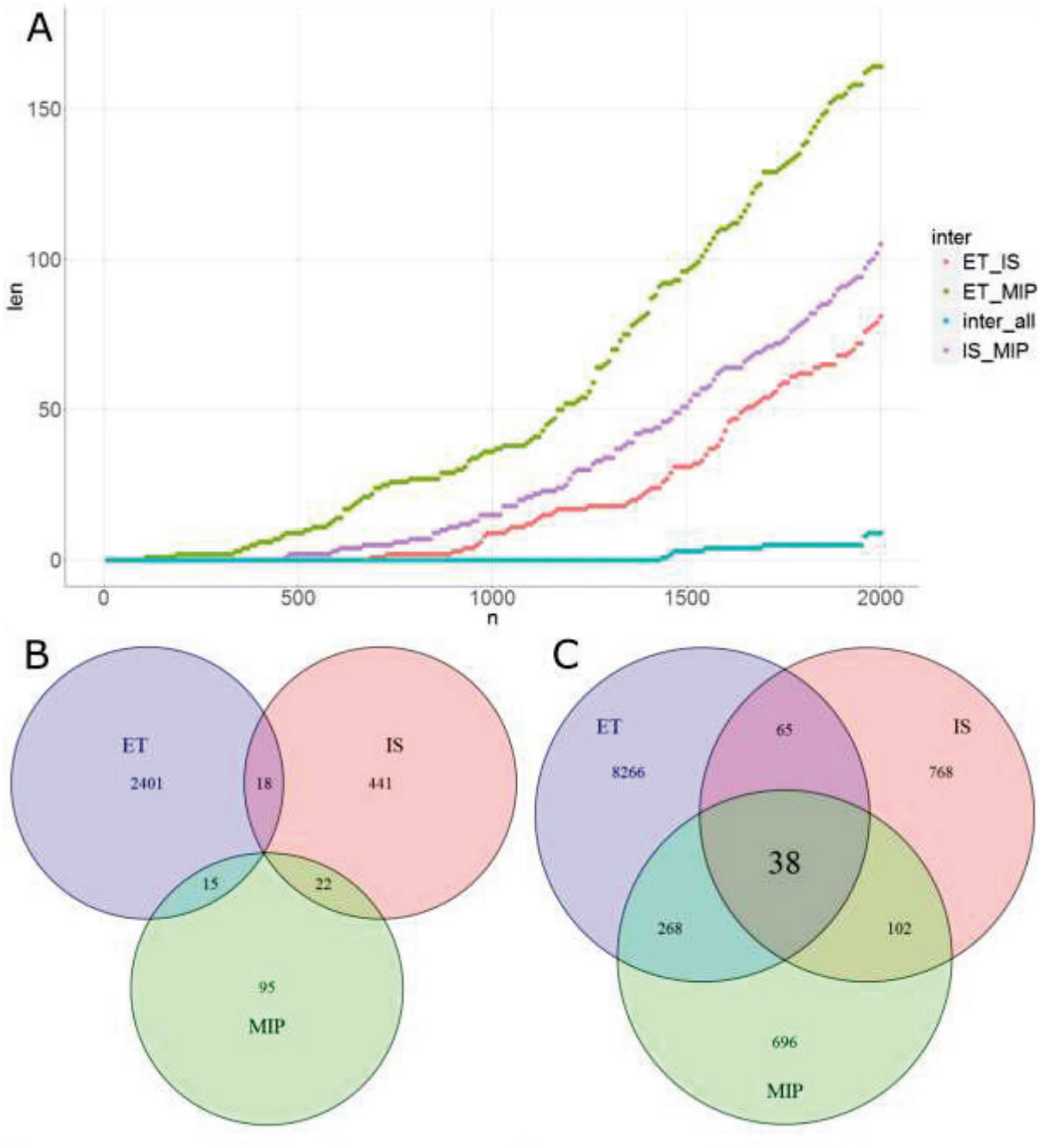


Figure 3-15 : HERV et gènes différentiellement exprimés dans les 3 jeux de données de la puce HERV-V3. A. Taille de l'intersection par paire de jeu de données, ou les 3 ensembles, en fonction de la taille des top listes de loci différentiellement exprimés. Pour des liste de loci différentiellement de taille allant de 1 à 2000, nous avons calculé l'intersection entre jeux de données, 2 par 2 (entre le modèle ET et Immunosepsis en rouge, entre le modèle ET et MIP-rea en vert, et entre Immunosepsis et MIP-rea en violet) ou les 3 jeux de données à la fois (en bleu). **B et C. Diagrammes de Venn montrant le nombre de probesets différentiellement exprimés pour des seuils fixes** (B. abs(log2FC)>1 et p-value ajustée <0,05 et C. abs(log2FC)>0,5 et p-value ajustée <0,1).

3.3 ETUDE DES VARIATIONS EN NOMBRE DE COPIES DES HERV DANS LE GENOME HUMAIN

3.3.1 RESUME DE L'ETUDE

A partir des données d'expression de la puce HERV-V3, nous avons remarqué une variabilité plus importante à intensité égale, de l'expression des HERV, par rapport à celle des gènes (Figure 3-16).

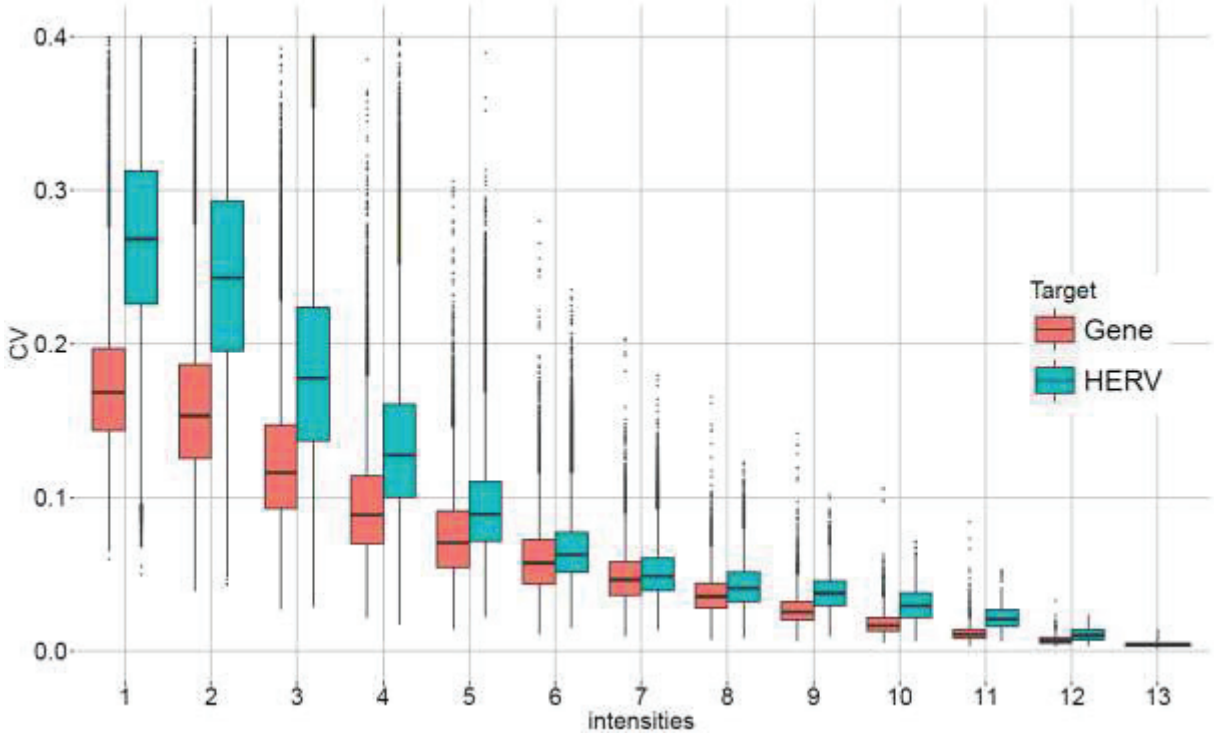


Figure 3-16: Coefficient de variation d'expression en fonction du niveau d'intensité d'expression entre gènes et HERV. L'axe des x représente les niveaux d'intensité d'expression, discrétisés (ex :1 signifie expression comprise entre 0.5 et 1.5). L'axe des y représente le coefficient de variatio (CV). Le CV est définit par l'écart-type divisé par la moyenne.

Au-delà de possibles biais techniques ou analytiques (le nombre plus important de HERV ciblés par rapport aux gène explique à lui seul la variabilité plus importante observée), nous posons l'hypothèse que cette variabilité d'expression des HERV est en partie due à un polymorphisme de présence des éléments considérés dans les génomes des individus étudiés. Cette étude a été réalisée en collaboration avec un étudiant en alternance, en master de bioinformatique à l'université de Rouen, Maxime Bodinier.

Pour répondre à cela, nous avons développé une méthode (HERVdel) qui compare la couverture des reads des données de séquençage génomique au niveau de chaque HERV annoté dans le génome avec sa région génomique environnante (+/- 25kb). Si le niveau de couverture est plus faible au niveau du HERV qu'au niveau de sa région environnante, on considérera que ce HERV est absent du génome de l'individu étudié. Nous évaluons l'ensemble des HERV annotés dans les bases de données publiques (nommé RepBase par simplicité) et notre propre base de données (hervgdb4), sur plus de 2000 génomes d'individus, inclus dans le projet 1000 génomes, regroupés en populations humaines, réparties dans les 5 continents. Nous résumons l'information du polymorphisme en fréquences alléliques d'absence par population humaine (VAF).

Nous montrons dans ce travail que la méthode développée est robuste et fiable, et qu'il existe un polymorphisme de présence des HERV dans le génome. Et notamment, que ce polymorphisme est lié à la population d'appartenance des individus, suggérant que ces éléments censés ne plus être mobiles, ont été en mouvement dans le génome récemment à l'échelle de l'évolution, après les phénomènes migratoires des populations humaines. En accord avec la littérature, on montre que le groupe HML-2, le plus récent des groupes de HERV, est celui possédant le plus de loci avec une grande fréquence d'absence dans les populations humaines.

Les HERV annotés dans le génome humain ne sont donc pas systématiquement présents chez tous les individus. Avant de chercher à connaître les éventuelles influences que les HERV très polymorphiques peuvent avoir sur l'ensemble du transcriptome, ces résultats sont à prendre en compte dans l'analyse du transcriptome de ces éléments, et pourraient en partie expliquer la variabilité d'expression plus importante observée chez les HERV par rapport aux gènes (Figure 3-16).

3.3.2 ARTICLE

HERVS COPY NUMBER VARIATION IN THE HUMAN POPULATION

Olivier Tabone*, M. Bodinier*, M. Mommert, M. Fablet, F. Mallet, J. Textoris

* Contribution à parts égales

Année 2018

En préparation

LTR retrotransposons copy number loss in human populations

ABSTRACT (250 words max)

LTR retrotransposons consist of Human Endogenous Retroviruses (HERVs) and Mammalian-apparent Long-terminal Repeat retrotransposons (MaLR) and they represent today more than 8% of the human genome. In order to know if all annotated LTR retrotransposons are present in all human populations, we analyze whole genome sequencing data. Based on two different annotation databases, regrouping more than 1,500,000 entries, we assess the presence / absence of each locus based on coverage analysis on the 1000 genomes dataset (2691 individuals of phase III divided into 26 populations and 5 super-populations). Validation of our method is performed on simulated data and through comparison to previous published data. We synthetize the results in percentages of absence of each locus in the 26 populations. Surprisingly, we observe an high proportion of loci being often absent in human populations. We also highlight super-population-specific patterns of LTR retrotransposon absence and we show that LTR5 is the most polymorphic group.

INTRODUCTION

LTR retrotransposons consist of Human Endogenous Retroviruses (HERVs) and Mammalian-apparent Long-terminal Repeat retrotransposons (MaLR) and they represent today more than 8% of the human genome. They are remnants of ancient and independent

retroviral infections within germline and have been fixed in our genome million years ago (Young, Stoye, and Kassiotis 2013). These rare events happened several times in evolution, forming several groups or families. Each insertion led to distinct groups or families, each including multiple copies. Current classification annotates more than hundred such groups. As retrotransposons, they are able to duplicate across the genome, via copy/paste mechanism, explaining their number in the genome.

A retrovirus classically consists of internal regions coding for viral proteins (gag, pro, pol, env) flanked by two identical Long Terminal Repeats (LTRs). For HERVs, the accumulation of mutations and recombination events during evolution made most of these elements incomplete, with a majority of solo LTRs, and defective for replication (Young, Stoye, and Kassiotis 2013). Yet, studies suggest that HERVs are involved in genome instability, through homologous recombinations (Campbell et al. 2014). The genome instability across individuals can be reflected by Copy Number Variations (CNVs), with some leading to disease (Rosenfeld et al. 2011; Sun et al. 2000).

Moreover, studies showed unfixed insertions of HERV-K (HML-2) (Wildschutte et al. 2016; Marchi et al. 2014; Belshaw et al. 2004), suggesting that HERVs are still moving in human genome, either by reinfection, duplication or homologous recombinations. A growing number of studies points out normally dormant HERVs being reactivated to drive diseases (Babaian and Mager 2016; Küry et al. 2018). Many examples exist in literature, especially in cancer (Lamprecht et al. 2010), autoimmune or neurological diseases (Suntsova et al. 2015; Christensen 2016). HERV can also play a role in physiological conditions, by driving syncytin (Bolze, Mommert, and Mallet 2017) or rewiring innate

immune pathways (Chuong, Elde, and Feschotte 2016). These studies highlight the role of HERVs on physiopathology, and knowing their presence or absence among human genomes seems crucial to better understand their impact.

Furthermore, the existing studies on structural variants used a large number of sequenced genomes, in order to provide a comprehensive description of common human genetic variation (Consortium 2015). The analyses showed that a typical genome differs from the reference human genome, and highlighted differences in structural variations between human populations (Consortium 2015). However, they did not focus specifically on repeated elements, especially on LTR retrotransposons. In this work, we wonder if known and annotated LTR retrotransposons are conserved in human genomes.

Thus, we studied LTR retrotransposons CNVs on 2691 genomes from 1000 genomes dataset (phase III). We assessed the presence / absence of each annotated locus based on coverage analysis for all individuals. Based on current public annotation of reference genome, it exists 717,659 LTR entries, covering 268 Million base pairs (Mbp) (Smit, AFA, Hubley, R & Green, P. 2013). We also used our own HERV annotation database, Hgdb4 (Becker et al. 2017; Mommert, Tabone et al. 2018). It contains 881,593 LTR retrotransposon entries, covering 223 Mbp. We validated our method on simulated data and through comparison to previously published data. We synthesized the results in percentages of absence of each annotated locus in the 26 populations. We developed a web-based application allowing to explore and visualize the results. We highlight super-population-specific patterns of LTR retrotransposon frequency and we show that a high proportion of loci are often absent in human populations.

64

65 # RESULTS

66 ## Development and validation of the method

67 We exploited the 1000 genome Whole Genome Sequencing (WGS) data to study
68 HERV Copy Number Variation (CNV), by loss of copy. We assessed the presence / absence of
69 each known HERV locus based on coverage analysis of all individuals.

70 More precisely, the method, hereafter called *HervDel*, consists on comparing the
71 sequencing coverage of each locus with its genomic neighborhood. If a given HERV locus,
72 for a given individual, has lower coverage than its genomic environment (+/- 25kb), the
73 HERV is variant relative to the reference. If not, the HERV is conform to the reference
74 (Figure 1A) . Actually, we distinguish 3 genotypes: the HERV is present on both alleles
75 (*HmzR*), deleted on both alleles (*HmzΔ*), and deleted on one allele (*Htz*). The Cumulative
76 Distribution Function (CDF) of the 50kb window coverage centered on the locus allows to
77 compute the probability to observe a lower coverage than the observed HERV coverage
78 (Figure S1). *HervDel* computes this probability, for all annotated HERV loci, and all
79 individuals.

80 These probabilities are then compared to a couple of thresholds (*genotyping*
81 *threshold*), one for discriminating the *HmzΔ* and *Htz* states, the other for discriminating
82 the *Htz* and the *HmzR* states. In order to both validate our method and choose the couple of
83 thresholds leading to the best performances, we simulated WGS dataset. We selected twice
84 chromosome 1 genomic reference sequence (mimicking the 2 alleles) and modified it by

replacing some Hgdb4 HERV loci nucleotide sequences by “N” sequences. We attributed a genotype state for all HERVs loci – 25% of HERV loci on chr1 are masked with N on both alleles, simulating *HmzA* state – another 25% of HERV loci are masked on one allele, simulating *Htz* state. We generated sequencing reads by applying ART (Huang et al. 2012) and we mapped the reads to the GRCh38 reference genome. This way, we are able to evaluate the performances of *HervDel* by generating Receiver Operating Characteristic (ROC) curves (Figure 1B). The blue curve represents the ability of the method to distinguish the *HmzΔ* compared to other states, while the orange curve represents the ability to distinguish *HmzR* compared to other states. The values on the curves are the corresponding probabilities on CDF (Figure S1), at youden index. We choose it as genotyping thresholds for the determination of presence /absence HERV loci in the real dataset. A probability lower than 0.107 means the HERV is absent on both alleles (*HmzΔ*), a probability between 0.107 and 0.275 means the HERV is absent on one allele (*Htz*), a probability greater than 0.275 means the HERV is present on both alleles (*HmzR*).

We compared the thresholds with all observed probabilities, and we obtain for each individual and each HERV locus, the corresponding genotype. Individuals belonging to the same population are grouped together, and for each HERV locus, we compute a Variant Allele Frequency (VAF). It represents the frequency of the absent HERV locus allele in a population. It is given by: $VAF_{popLocus} = \frac{2 \cdot Nb_{Hmz\Delta} + Nb_{Htz}}{2 \cdot Nb_{tot}} \cdot 100$, with Nb_{tot} the total number of individuals in the population, $Nb_{Hmz\Delta}$ the number of homozygous deletions for the given HERV locus in the population and $Nb_{Htz}$the number of heterozygous deletions for the given HERV locus in the population. The VAF, expressed in percentage, is thus a measure of

polymorphism of a HERV locus within a population. A VAF at 0% indicates the presence of the locus in all individuals of the population and a VAF at 100% indicates an absence of the locus in all individuals of the population. An intermediate VAF indicates a genotype heterogeneity within the observed population.

Cleaning data

Before applying *HervDel*, a filtering step is needed. Some HERV loci present particular characteristics, potentially leading to misinterpretations or errors in the method. We excluded 26,389 loci having GC rate lower than 31% or higher than 64% (Figure S2A), 23,890 loci within low coverage or highly variable coverage regions (Figure S2B), and 164,609 loci smaller than 90 bps. In total, 214,888 entries (13.4%) were discarded from the analysis (Figure S2C).

Comparison with existing studies.

We compared the results of *HervDel* with estd214 study (1000 genomes Consortium, 2015) and the study hereafter called Wildschutte (Wildschutte et al. 2016). Among all CNV losses detected in estd214, 11 have genomic coordinates which correspond to a HERV locus (Figure 2A). On these loci, the VAF by super-population are similar between the 3 methods. The VAF are identical across approaches for 7 loci. When the VAF is high, *HervDel* tends to underestimate the VAF compared to estd214 and Wildshutte method. The differences are systematic, no matter the super-population. By counting the number of loci being included in a CNV loss from estd214, no matter the length of the deletion, we observe that the

proportion increases with the VAF, from 16% for all loci, to 36% for loci with VAF higher than 90% (Table S3). The comparison with Wildschutte method on HML-2 loci show a strong correlation of VAF (R^2=0.84, Figure 2B). Again, *HervDel* has lower VAF compared to Wildschutte method. Interestingly, *HervDel* detects polymorphism on 1509 loci whereas Wildschutte method does not. The opposite is also true but to a lesser degree (26 loci).

HERV CNV description.

The median VAF on the whole dataset is at 4.48%, with a majority of loci with a VAF between 2 and 5% (555,588 loci, 39.9%). Depending of the source of annotation, the distributions of VAF are quite different (Figure 3A). The VAF computed from repBase annotation are lower, with median equal to 3.25%, than the VAF from Hgdb4, with a median equal to 5.26%. The locus size distributions are different according to the annotation source, with a median equal to 358bp for repBase and a median equal to 155bp for Hgdb4 (Figure S3). Thus, we represent the VAF as a function of locus size and show that the small loci tend to have higher VAF than the long loci, no matter the annotation source (Figure 3B). Among the 2000 most variant loci from Hgdb4 annotation, the most represented sub-groups are LTR5_Hs and LTR5 (Figure 3C). These are LTRs of the HML-2 group. HERV-L and many MalR groups are also represented in the most polymorphic groups. Moreover, LTR5_Hs, LTR5 and LTR5B are the most variant sub-groups and with the highest VAF (VAF equals to 12.8, 13 and 9.8% respectively). The MalR and HERV-L groups have a median VAF between 4.8 and 6.4%. If we look at the distribution of VAF by super-population, AMR, EUR, SAS and AFR super-populations have similar distributions, having a median VAF between 4.49 and 4.91% (Figure 3D). Unexpectedly, VAF distribution of EAS is different from the

other super populations, with a shift toward lower VAF values. The 3 populations with the lower VAF belong to EAS populations (Figure S4). EAS super-population have lower number of loci with VAF higher than 1% or VAF higher than 5% compared to other super-populations (Table 1). However, EAS super population tends to have more loci with VAF higher than 50%.

Table 1. Number of loci with VAF (deleted) above a given threshold, per super-population

	VAF > 1%	VAF > 5%	VAF > 10%	VAF > 50%
African (AFR)	1,299,212 (93.7%)	604,588 (43.6%)	154,894 (11.2%)	1,575 (0.1%)
American (AMR)	1,310,915 (94.5%)	680,291 (49.0%)	168,874 (12.2%)	1,776 (0.1%)
East Asian (EAS)	1,244,886 (89.8%)	485,755 (35.0%)	147,424 (10.6%)	2,678 (0.2%)
European (EUR)	1,311,327 (94.5%)	646,709 (46.6%)	146,630 (10.6%)	1,528 (0.1%)
South Asian (SAS)	1,288,220 (92.9%)	616,957 (44.5%)	161,446 (11.6%)	2,010 (0.1%)

By making intersections between the super-populations, we observe that a large majority are common between the 5 super-populations, for loci with VAF higher than 1%, 5% and 10% (Figure S5). Interestingly, the proportion of loci, absent in one super-population exclusively, is pretty low (Table 2). But it tends to be higher when VAF threshold increases. Again, EAS super-population has a particular pattern, with very low proportion (0.06%) of exclusive loci for VAF higher than 1%, but which greatly increases (31%) for VAF higher than 50%.

	Super pop	AFR	AMR	EAS	EUR	SAS	Total exclusive
VAF>1%	Exclusive	3697	11573	734	5207	2291	23502
	Total	1282078	1300426	1178121	1281288	1252941	6294854
	Proportion	**0,0028836**	**0,00889939**	**0,000623**	**0,0040639**	**0,0018285**	**0,003733526**
VAF>5%	Exclusive	15328	46640	4851	31879	24396	123094
	Total	588519	660936	466722	632742	601065	2949984
	Proportion	**0,02604504**	**0,07056659**	**0,010394**	**0,0503823**	**0,040588**	**0,041727006**
VAF>10%	Exclusive	8362	16276	9955	11759	13361	59713
	Total	152845	166248	148761	150029	161833	779716
	Proportion	**0,05470902**	**0,09790193**	**0,066919**	**0,0783782**	**0,0825604**	**0,076583012**
VAF>50%	Exclusive	126	86	938	86	214	1450
	Total	1634	1915	3015	1671	2147	10382
	Proportion	**0,07711138**	**0,04490862**	**0,311111**	**0,0514662**	**0,099674**	**0,139664804**

166 **Table 2: Super-population specific polymorphic loci.** Number and proportion of loci
167 being exclusive to a super-population, according to a VAF threshold.

168 On the 500 most variable loci across populations, different VAF patterns appear
169 (Figure 4A). The hierarchical clustering groups populations according to their belonging
170 super-populations almost perfectly. African super-population is the most distant super-
171 population from the others. American populations appear split, Puerto Rican and
172 Colombian populations are grouped with European populations. Interestingly, on several
173 loci, VAF is very high for all super-populations, except for one, meaning the locus is absent
174 in a large majority of cases. To detect HERVs loci with important VAF difference between at
175 least 2 populations, we selected the 40 most inter-population distant loci (Figure 4B).
176 Among these 40 loci, 23 have a VAF higher than 50% in a majority of populations .
177 Interestingly, 29 loci over 40 overlap with a CNV loss from estd214 study (Table S4).
178 Moreover, 18 belong to repBase and 22 to Hgdb4. Over the 22 from Hgdb4, 13 loci overlap

with one of the 18 from repBase. On these 18 loci, 5 are from MER41 super-group (28%, comprised of LTR25, LTR38-int, MER4E1, MER66B and PRIMA41-int groups).

DISCUSSION

We developed a sensitive and robust method, *HervDel*, assessing the presence or absence of annotated HERV on 2,691 human genomes from 26 populations. This is a first global CNV loss analysis exclusively focused on more than 1,500,000 HERV entries. Contrary to what is admitted, we showed that VAF, the allele frequency of HERV absence, is pretty high, meaning that HERV loci are far from being present in all human genomes (Jern and Coffin 2008). The frequencies observed are higher than existing CNV studies on human genomes. While *HervDel* concludes that more than 90% of HERVs entries are variant from the reference in at least 1% of samples, 1000 genomes consortium study (Consortium 2015) concluded to a large majority of variants having a frequency lower than 0.5%. And in a meta-analysis, Zarrei et al. (Zarrei et al. 2015) showed that most CNV losses with at least 1% frequency are rare (0.2%). These differences suggest that HERVs loci evolve differently compared to the rest of the genome. Despite it is admitted that majority of HERV groups are not able to duplicate in human genome anymore, the results suggest a global HERV polymorphism, although mechanisms remain to be understood.

The relationship we show between locus size and allele frequency, can be explained by 2 hypotheses: first, there is an artificial effect of locus size on allele frequency, explained by the high coverage variability leading to higher probability for a small locus to be in a

drop of coverage. Besides, this relationship has already been observed in whole genome analysis (Consortium 2010). But the hierarchical clustering on variant loci groups related populations together, showing that if any, this hypothesis is not exclusively explaining the high polymorphism observed. The second hypothesis is that small HERV loci tend to be more easily deleted in genomes although the mechanisms are not known. This size effect is also illustrated by the higher absence frequencies observed for Hgdb4, which have smaller entries, at HERV functional region level. It may also highlight partial HERV locus deletions that could occur frequently, and not detectable at locus level.

Interestingly, LTRs from HML-2, the most recent group (Bannert and Kurth 2006), appear to be the most polymorphic group. This is in line with existing studies focusing exclusively on this group, which suggests this recently integrated group is still active and is able to retrotranspose (Wildschutte et al. 2016; Marchi et al. 2014). The good correlation with Wildschutte study on HML-2 elements confirms this group is polymorphic, and reinforce the validity of *HervDel*. Despite our method tends to underestimate frequencies compared to other studies, *HervDel* seems also more sensitive than estd214 and Wildschutte studies. This may be explained by the approach we developed, allowing to detect coverage drops in local genomic environment, and avoiding the problems related to between-sample coverage variability. To go further on the analysis of HML-2 group, it would be interesting to focus on annotated whole elements from HML-2 group and verify if only solo LTRs can be detected, proving recent homologous recombination occurred.

Moreover, we highlighted population-specific patterns of VAF variations, suggesting recent duplication events in human evolution ladder. These CNV loss population-patterns

222 are not specific to HERVs. Indeed, the publication from the 1000 Genomes Project
223 Consortium (Consortium 2015) shows populations structures, separating super-
224 populations and highlighting their internal substructure. According to their results, we also
225 show that majority of variant are shared between the super-populations, but when the
226 frequency of absence is high, there are many super-population-exclusive polymorphism.
227 Like in *HervDel*, Colombian and Puerto Rican populations group with European populations.
228 However, the very particular patterns found for East Asian populations have not been
229 observed in previous studies. It seems this super-population have globally fewer number of
230 deleted loci, with lower VAF. But compared to other super-populations, EAS has higher
231 number of loci with very high VAF. It would be interesting to specifically investigate these
232 observations in further studies. And a way to better understand the population-specific
233 patterns would be to investigate data from human close ancestors, like chimpanzees, but
234 also from Australopithecus to homo Neanderthal's genomes (Gibbons 2017).

235 Importantly, we highlight in this work, that many HERVs annotated in reference
236 genome, are actually absent in most of the populations. For example for HERV
237 transcriptome analyses, knowing presence or absence of HERV loci in a population should
238 allow to gain power and accuracy.

239 Finally, we showed that HERVs might be more polymorphic than expected. The real
240 impact of HERV CNV loss is not known; but we know that some HERVs are involved in some
241 physiological or pathological settings. As an HERV which is absent cannot be expressed or
242 modulated (Mommert, Tabone et al. 2018; Lamprecht et al. 2010), or cannot provide
243 transcription factor binding site (TFBS) to neighbor genes (Ito et al. 2017; Chuong, Elde,

and Feschotte 2016), this study can become of importance in future global studies on HERVs. The latter example suggested that loci from MER41 group include TFBS, especially for STAT1. Yet, among the most distant loci between populations, we found an important number of loci-related MER41 group. Sixty percent of these elements carry both probable STAT1 and IRF1 binding sites (data not shown). It suggests that the frequent absence of these elements in many populations could lead to some abnormal gene expression in particular settings, potentially leading to diseases. To conclude, existing HERVs annotations have to be taken with caution when analyzing HERVs, especially with reference genome-based next generation sequencing data, as most individuals seem to be very different from reference genome for many HERVs.

METHODS

Genomic DATA

We retrieved 1000 Genomes Project (Consortium 2010) low coverage data (~4-10x) , in GRCh38 aligned format. (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/data/) . It represents a total of 2,691 samples from 5 super-populations subdivided in 26 populations. A population contains between 66 and 120 samples (Table S1).

HERV loci annotation Data

We target GRCh38 HERV loci from two annotation databases: Hgdb4 (Becker et al. 2017) and RepeatMasker (Smit, AFA, Hubley, R & Green, P. 2013), Open-3.0. http://www.repeatmasker.org. 1996-2010) track, from UCSC Table Browser (Karolchik et al. 2004). RepeatMasker lists 720,177 repBase entries, at LTR and HERV internal regions levels and Hgdb4 lists 881,593 entries, at LTR functional sub-regions and HERV internal regions levels.

Method Development.

Downloaded WGS alignment files were programmatically parsed using python package pysam 0.12. HERV loci GC rate was determined using biopython 1.7 on GRCh38 loaded genome through the package pyfaidx 0.5. Probability to observe a coverage lower than or equal to HERV locus median coverage in 50kb window, was computed dividing the count of base pair coverage lower than or equal to median HERV coverage by the size of 50kbps region centered on HERV locus. The resulting database was stored in a dataframe using pandas 0.22 package.

Performances and threshold definition

WGS read simulation step was performed using ART MountRainer (Huang et al. 2012) with similar parameters to 1000 genomes WGS conditions, specifying Illumina HiSeq 2000 as the sequencing system, paired-read, read length of 100 bps, mean fragment length of 386 bps with a standard deviation of 10 and a read coverage of 3x (for each allele). After a mate-wise concatenation of each created fastq, bwa mem (Li and Durbin 2009) was run with the

same parameters used by the 1000 genomes consortium: a penalty mismatch of 4, a gap open penalties of 6, and a gap extension penalty of 1. Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC), was obtained using scikit-learn 0.19 package. Youden's index was computed for each curve by optimization of adding sensitivity to specificity minus one. The probabilities at Youden's index are 0.107 and 0.275, and as threshold between *HmzΔ* and *Htz* genotypes, and between *Htz* and *HmzR* genotypes respectively for all analyses. All VAF are computed using $VAF_{popLocus} = \frac{2 \cdot Nb_{Hmz\Delta} + Nb_{Htz}}{2 \cdot Nb_{tot}} \cdot 100$ with Nb_{tot} the total number of individual in the population, $Nb_{Hmz\Delta}$ the number of homozygous deletions for the given HERV locus in the population and $Nb_{Htz}$the number of heterozygous deletions for the given HERV locus in the population.

Cleaning Data

Loci with a size below 90 bps, with too low coverage, in highly variable region or with too extreme GC rate were discarded prior running *HervDel*.

For highly variable regions determination, 3 genes were randomly selected in each chromosome (Table S2). For each sample and each chromosome, the coverage standard deviation for the 3 genes was averaged and 2 was added to this value. The obtained value was used as threshold to discard HERV loci within highly variable coverage 50 kbps region (e.g. on chromosome 7, Figure S2B). Similarly, a coverage lower than or equal to 2 was used to discard HERV loci region of too low coverage. The influence of GC rate on coverage was studied running *HervDel* (*genotyping thresholds* of 0.107 and 0.275) on remaining HERV loci (Figure S2A). The loess curve was computed using statsmodel 0.8 package, the derivative was obtained after fitting an univariate spline on the loess curve with scipy 1.0.1 package.

Comparison with other studies

Wildschutte et al. (Wildschutte et al. 2016) VAF data were obtained from the supplementary dataset 2 of the publication. Loci positions were shifted from GRCh37 to GRCh38 position using UCSC tool liftOver. The PNAS loci were then intersected with the ones from our annotation. If multiple loci match within our DB, the longer one was retained. 95 loci intersect with ours. 1000 genomes structural variants file (estd214 study) was retrieved from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/ALL.wgs.integrated_sv_map_v2.20130502.svs.genotypes.vcf.gz. For overlap summary between *HervDel* and estd214, only deletions, intersecting at least 95% of locus size from our annotation were kept (218,216 structural variants). 11 CNVs from estd214 have identical size with a HERV locus (+/- 5%).

HERV CNV description

Hierarchical clustering were made with Euclidean distance and complete method for rows and correlation distance (Pearson) and average method for columns. The K-means clustering was made using scikit-learn 0.19 package.

Shiny web application

A shiny web application (ADRESS) has been done to visualize VAF results through interactive heatmaps (Figure S6). This application allows users to display most variant or most deleted loci, to display a set of loci of interest or all loci within a genome slice. For

functional analyses, name of genes close to HERV loci can be displayed. Data displayed on heatmap can be downloaded.

Python and R packages

Both R (3.4.3) and python (3.6.3) languages were used. From python, packages used for graphical visualization are matplotlib 2.2.0, seaborn 0.8.1 and plotnine 0.3.0. From R, packages used are ggplot2 2.1.1, VennDiagram 1.6.20, pheatmap 1.0.8, Shiny web application (WEB ADDRESS) is based on shiny 1.0.5 package, static heatmap is based on pheatmap 1.0.8 package, interactive heatmap on heatmaply 0.14.1, and tables on the DataTables wrapper package DT 0.2.

DATA ACCESS

ACKNOWLEDGMENTS

Author contributions

DISCLOSURE DECLARATION

REFERENCES

Babaian, Artem, and Dixie L. Mager. 2016. "Endogenous Retroviral Promoter Exaptation in Human Cancer." *Mobile DNA* 7: 24. https://doi.org/10.1186/s13100-016-0080-x.

342 Bannert, Norbert, and Reinhard Kurth. 2006. "The Evolutionary Dynamics of Human
343 Endogenous Retroviral Families." *Annual Review of Genomics and Human Genetics* 7
344 (1): 149–73. https://doi.org/10.1146/annurev.genom.7.080505.115700.

345 Becker, Jérémie, Philippe Pérot, Valérie Cheynet, Guy Oriol, Nathalie Mugnier, Marine
346 Mommert, Olivier Tabone, Julien Textoris, Jean-Baptiste Veyrieras, and François
347 Mallet. 2017. "A Comprehensive Hybridization Model Allows Whole HERV
348 Transcriptome Profiling Using High Density Microarray." *BMC Genomics* 18 (April).
349 https://doi.org/10.1186/s12864-017-3669-7.

350 Belshaw, Robert, Vini Pereira, Aris Katzourakis, Gillian Talbot, Jan Pačes, Austin Burt, and
351 Michael Tristem. 2004. "Long-Term Reinfection of the Human Genome by
352 Endogenous Retroviruses." *Proceedings of the National Academy of Sciences* 101
353 (14): 4894–99. https://doi.org/10.1073/pnas.0307800101.

354 Bolze, P.-A., M. Mommert, and F. Mallet. 2017. "Contribution of Syncytins and Other
355 Endogenous Retroviral Envelopes to Human Placenta Pathologies." *Progress in*
356 *Molecular Biology and Translational Science* 145: 111–62.
357 https://doi.org/10.1016/bs.pmbts.2016.12.005.

358 Campbell, Ian M., Tomasz Gambin, Piotr Dittwald, Christine R. Beck, Andrey Shuvarikov,
359 Patricia Hixson, Ankita Patel, et al. 2014. "Human Endogenous Retroviral Elements
360 Promote Genome Instability via Non-Allelic Homologous Recombination." *BMC*
361 *Biology* 12 (1): 74. https://doi.org/10.1186/s12915-014-0074-4.

362 Christensen, Tove. 2016. "Human Endogenous Retroviruses in Neurologic Disease." *APMIS:*
363 *Acta Pathologica, Microbiologica, et Immunologica Scandinavica* 124 (1–2): 116–26.
364 https://doi.org/10.1111/apm.12486.

365 Chuong, Edward B., Nels C. Elde, and Cédric Feschotte. 2016. "Regulatory Evolution of
366 Innate Immunity through Co-Option of Endogenous Retroviruses." *Science (New*
367 *York, N.Y.)* 351 (6277): 1083–87. https://doi.org/10.1126/science.aad5497.

368 Consortium, The 1000 Genomes Project. 2010. "A Map of Human Genome Variation from
369 Population-Scale Sequencing." *Nature* 467 (7319): 1061–73.
370 https://doi.org/10.1038/nature09534.

371 Consortium, The 1000 Genomes Project. 2015. "A Global Reference for Human Genetic
372 Variation." *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

373 Gibbons, Ann. 2017. "Neandertals Gave 'Lost' African DNA Back to Moderns." *Science* 358
374 (6362): 431–431. https://doi.org/10.1126/science.358.6362.431.

375 Huang, Weichun, Leping Li, Jason R. Myers, and Gabor T. Marth. 2012. "ART: A next-
376 Generation Sequencing Read Simulator." *Bioinformatics (Oxford, England)* 28 (4):
377 593–94. https://doi.org/10.1093/bioinformatics/btr708.

378 Ito, Jumpei, Ryota Sugimoto, Hirofumi Nakaoka, Shiro Yamada, Tetsuaki Kimura, Takahide
379 Hayano, and Ituro Inoue. 2017. "Systematic Identification and Characterization of

Regulatory Elements Derived from Human Endogenous Retroviruses." *PLOS Genetics* 13 (7): e1006883. https://doi.org/10.1371/journal.pgen.1006883.

Jern, Patric, and John M. Coffin. 2008. "Effects of Retroviruses on Host Genome Function." *Annual Review of Genetics* 42 (1): 709–32. https://doi.org/10.1146/annurev.genet.42.110807.091501.

Karolchik, Donna, Angela S. Hinrichs, Terrence S. Furey, Krishna M. Roskin, Charles W. Sugnet, David Haussler, and W. James Kent. 2004. "The UCSC Table Browser Data Retrieval Tool." *Nucleic Acids Research* 32 (Database issue): D493-496. https://doi.org/10.1093/nar/gkh103.

Küry, Patrick, Avindra Nath, Alain Créange, Antonina Dolei, Patrice Marche, Julian Gold, Gavin Giovannoni, Hans-Peter Hartung, and Hervé Perron. 2018. "Human Endogenous Retroviruses in Neurological Diseases." *Trends in Molecular Medicine* 24 (4): 379–94. https://doi.org/10.1016/j.molmed.2018.02.007.

Lamprecht, Björn, Korden Walter, Stephan Kreher, Raman Kumar, Michael Hummel, Dido Lenze, Karl Köchert, et al. 2010. "Derepression of an Endogenous Long Terminal Repeat Activates the CSF1R Proto-Oncogene in Human Lymphoma." *Nature Medicine* 16 (5): 571–79, 1p following 579. https://doi.org/10.1038/nm.2129.

Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform." *Bioinformatics* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Marchi, Emanuele, Alex Kanapin, Gkikas Magiorkinis, and Robert Belshaw. 2014. "Unfixed Endogenous Retroviral Insertions in the Human Population." *Journal of Virology* 88 (17): 9529–37. https://doi.org/10.1128/JVI.00919-14.

Mommert, Marine, Olivier Tabone, Guy Oriol, Elisabeth Cerrato, Audrey Guichard, Magali Naville, Paola Fournier, et al. 2018. "LTR-Retrotransposon Transcriptome Modulation in Response to Endotoxin-Induced Stress in PBMCs." *BMC Genomics* 19 (1): 522. https://doi.org/10.1186/s12864-018-4901-9.

Rosenfeld, Jill A., Yves Lacassie, Dima El-Khechen, Luis F. Escobar, James Reggin, Carolyn Heuer, Emily Chen, et al. 2011. "New Cases and Refinement of the Critical Region in the 1q41q42 Microdeletion Syndrome." *European Journal of Medical Genetics* 54 (1): 42–49. https://doi.org/10.1016/j.ejmg.2010.10.002.

Smit, AFA, Hubley, R & Green, P. 2013. "RepeatMasker." 2015 2013. http://www.repeatmasker.org/.

Sun, Chao, Helen Skaletsky, Steve Rozen, Jörg Gromoll, Eberhard Nieschlag, Robert Oates, and David C. Page. 2000. "Deletion of Azoospermia Factor a (AZFa) Region of Human Y Chromosome Caused by Recombination between HERV15 Proviruses." *Human Molecular Genetics* 9 (15): 2291–96. https://doi.org/10.1093/oxfordjournals.hmg.a018920.

Suntsova, Maria, Andrew Garazha, Alena Ivanova, Dmitry Kaminsky, Alex Zhavoronkov, and Anton Buzdin. 2015. "Molecular Functions of Human Endogenous Retroviruses in Health and Disease." *Cellular and Molecular Life Sciences: CMLS* 72 (19): 3653–75. https://doi.org/10.1007/s00018-015-1947-6.

Wildschutte, Julia Halo, Zachary H. Williams, Meagan Montesion, Ravi P. Subramanian, Jeffrey M. Kidd, and John M. Coffin. 2016. "Discovery of Unfixed Endogenous Retrovirus Insertions in Diverse Human Populations." *Proceedings of the National Academy of Sciences* 113 (16): E2326–34. https://doi.org/10.1073/pnas.1602336113.

Young, George R, Jonathan P Stoye, and George Kassiotis. 2013. "Are Human Endogenous Retroviruses Pathogenic? An Approach to Testing the Hypothesis." *Bioessays* 35 (9): 794–803. https://doi.org/10.1002/bies.201300049.

Zarrei, Mehdi, Jeffrey R. MacDonald, Daniele Merico, and Stephen W. Scherer. 2015. "A Copy Number Variation Map of the Human Genome." *Nature Reviews Genetics* 16 (3): 172–83. https://doi.org/10.1038/nrg3871.

Figure legends

Figure 1: Principle and performances of HervDel. (a.) Schematic representation . Reference genome is represented in black. Three samples genome reference-aligned reads (arrows) and the associated coverage (violet, blue and orange curves) are represented. Coverage at herv1 locus is equal to its environment coverage (Sample1, in violet), faithful to the reference status or lower (Sample2, in blue and Sample3, in orange), variant from reference. HmzR: Homozygous Reference, locus present on both alleles, Htz: Heterozygous, HmzΔ: Homozygous deleted, locus absent on both alleles. **(b.) Performances and threshold selection.** Receiver operating characteristic (ROC) curves representing the performances of detection of HmzΔ versus other genotypes (blue) and HmzR versus other genotypes (orange). Numbers in blue and orange are the probability to have lower coverage at Youden's index maximum value, used as threshold between HmzΔ and Htz genotypes, and between Htz and HmzR genotypes respectively.

AUC: Area Under the Curve.

Figure 2: Comparison of HervDel. (a.) Comparison of the 11 common loci between estd214, Wildschutte and HervDel studies. HervDel and estd214 intersecting loci are represented in x-axis, VAF (%) in y-axis. Each graph represents a super-population. red points are VAF for Wildschutte, green points for estd214 and blue points for HervDel study. **(b.) Correlation between Wildschutte and HervDel.** 95 HML-2 loci intersect between the 2 studies. x-axis represents VAF from Wildschutte study, y-axis VAF from HervDel. Each point represents one locus, for one population. Red line is the line y=x, dotted orange line is the linear regression between the 2 methods. R^2=0.84.

Figure 3: HERV absence frequency description. (a.) VAF distribution according to annotation database. Kernel density estimation of VAF. Blue curve corresponds to RepBase, and orange curve to Hgdb4. **(b.) VAF as a function of locus size.** Bars represent median VAF according to locus size. Blue bars correspond to RepBase, and orange bars to Hgdb4. **(c.) Most polymorphic HERV groups in Hgdb4.** Groups from Hgdb4 with at least 20 loci among the top 2000 loci with highest VAF are represented. Grey bars represent the number of loci within the corresponding group. Orange points represent the median VAF for each group. **(d.) VAF density distribution according to super-population.** Kernel density estimation of VAF. Super-population are color-coded. AMR: American, EUR: European, SAS: South Asian, AFR: African, EAS: East Asian.

Figure 4: Super-population clusterization (a.) VAF of the 500 most variant loci. Heatmap of VAF (from 0% (blue) to 100% (red)). Each line represents a HERV locus, each column represents a population. On the top, super-populations are color-coded. Clusterization was made with Euclidean distance and complete method for rows and correlation distance (Pearson) and average method for columns. **(b.) 40 most distant loci between 2 populations loci.** Each locus was splitted in two clusters according to a k-means method. This heatmap represents the 40 loci with highest inertia, labelled with their genomic position (rows). Each line represents a HERV locus, each column represents a population. On the top, super-populations are color-coded. Clusterization was made with Euclidean distance and complete method for rows and correlation distance (Pearson) and average method for columns.

Supplementary figures

Supplementary figure 1. Principle of implemented HervDel method. (a.) LTR5_Hs locus (chr7:158236790-158237758) and its 2kb surrounding genomic environment coverage. hg38 aligned WGS reads come from HG01707 sample. Red segment is positioned on locus LTR5_Hs, at its median coverage. Yellow dashed line corresponds to 50 kb coverage median, centered on HERV locus. **(b.) Cumulative distribution function of 50kb coverage, centered on LTR5_Hs locus.** Red cross is the point at median coverage of LTR5_Hs (probability = 0.074). Dashed black lines correspond to genotyping thresholds and illustrate the way the method assigns a genotype to observed HERV coverage probabilities . In this example, the locus genotype is HmzΔ.

Supplementary figure 2. Cleaning data. (a.) Deviation of VAF as a function of GC rate.

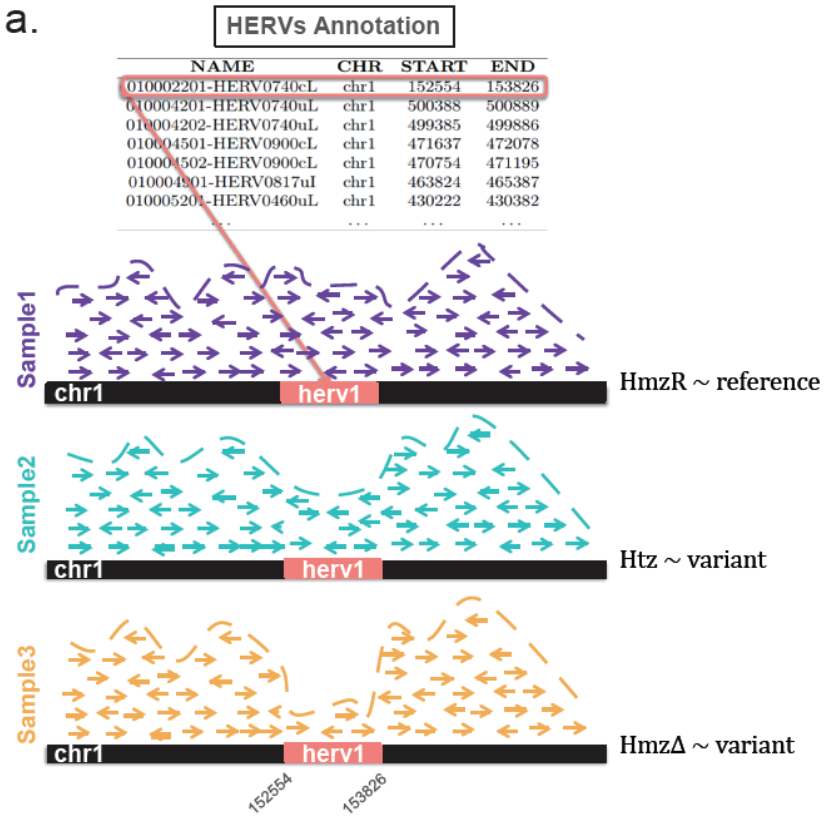
VAF computed using HervDel with genotyping thresholds of 0.107 and 0.275. Scatter plot with hexagonal binning of inferred loci VAF as a function of its GC rate (density from blue to yellow). Curve of VAF mean in 1000 bins GC rate (blue). Loess curve fitted on 1000 bins mean VAF (orange). Points where loess curve derivative is equal to -1 (GC rate = 0.31) or 1 (GC rate = 0.64) are in white. **(b.) Coverage variation along chromosome 7 in sample HG01707.** x-axis represents chromosome 7 position and y-axis represents the coverage for this sample. **(c.) Cleaned loci by chromosome.** In blue, count of loci with a GC rate lower than 0.31 or greater than 0.64. In orange, count of loci within highly variable 50kb-regions or in two low coverage 50kb-regions. In green, count of loci with a size lower than 90 bps. In total, 214,888 loci where biased annotated, among 1,601,905 initial loci (13%).

Supplementary figure 3. Locus-size distribution according to annotation database. Kernel density distributions of locus size, by annotation database. Blue curve corresponds to RepBase (median = 358), red curve to Hgdb4 (median = 155).

Supplementary figure 4. VAF distribution in the 26 populations. VAF violin plots according to population. Violin plots are color-coded by super-population. The populations are sorted by decreasing VAF.

Supplementary figure 5. Overlaps between super populations. a. Venn diagramm of loci with A. VAF higher than 1% in a super-population, B. VAF higher than 5% in a super-population, C. VAF higher than 10% in a super-population, D. VAF higher than 50% in a super-population.

Supplementary figure 6. Shiny web application overview. (a.) General view. On top left panel, the loci search functionality, on bottom, a quick summary table. On the top right panel, parameters option button, a heatmap matrix download button and the possibility to switch from a flat heatmap to an interactive heatmap. On the bottom right panel, the heatmap displayed, according to selected parameters. **(b.) Parameter view.** On the left column, the observed chromosome with possibility to focus only on a region. On the middle column, the possibility to only display loci belonging to one of the three loci sources (repeatMasker track or hgdb4 high quality of annotation -- Proto_HGDB4 -- or hgdb4 low quality -- DFAM_HGDB4 --), the possibility to change the displayed loci id and the possibility to restrain loci to ones close to genes according to a distance threshold. On the right column, the possibility to clusterize columns and to separate row dendrogram in n groups. On the bottom, the possibility to display top standard deviation loci (most polymorphic loci) or the top most deleted loci within the whole population.


a.
HERVs Annotation
NAME CHR START END
010002201-HERV0740cL chr1 152554 153826
010004201-HERV0740uL chr1 500388 500889
010004202-HERV0740uL chr1 499385 499886
010004501-HERV0900cL chr1 471637 472078
010004502-HERV0900cL chr1 470754 471195
010004901-HERV0817uI chr1 463824 465387
010005201-HERV0460uL chr1 430222 430382
Sample1
chr1
herv1
HmzR ~ reference
Sample2
chr1
herv1
Htz ~ variant
Sample3
chr1
herv1
HmzΔ ~ variant
152554
153826

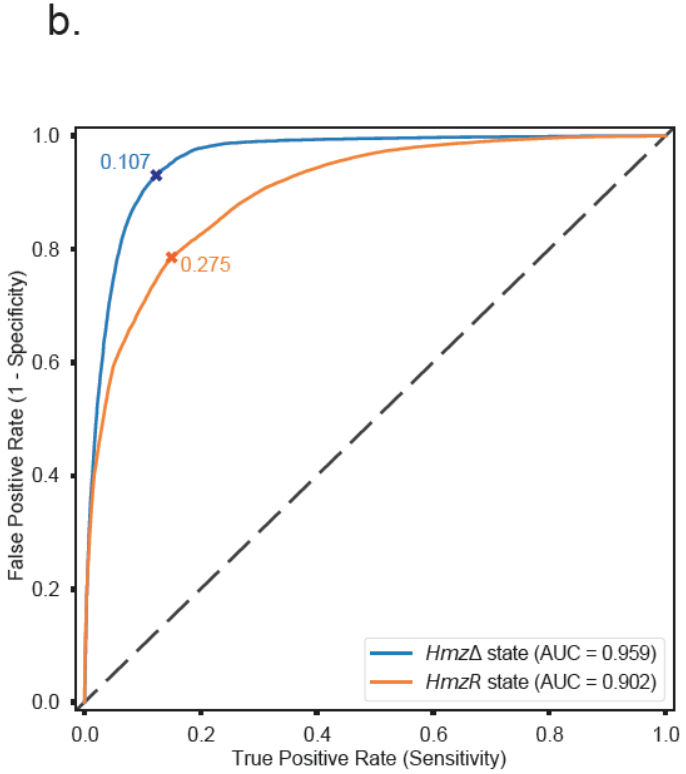

b.
1.0
0.8
0.6
0.4
0.2
0.0
False Positive Rate (1 - Specificity)
0.107
0.275
HmzΔ state (AUC = 0.959)
HmzR state (AUC = 0.902)
0.0
0.2
0.4
0.6
0.8
1.0
True Positive Rate (Sensitivity)

Figure 1

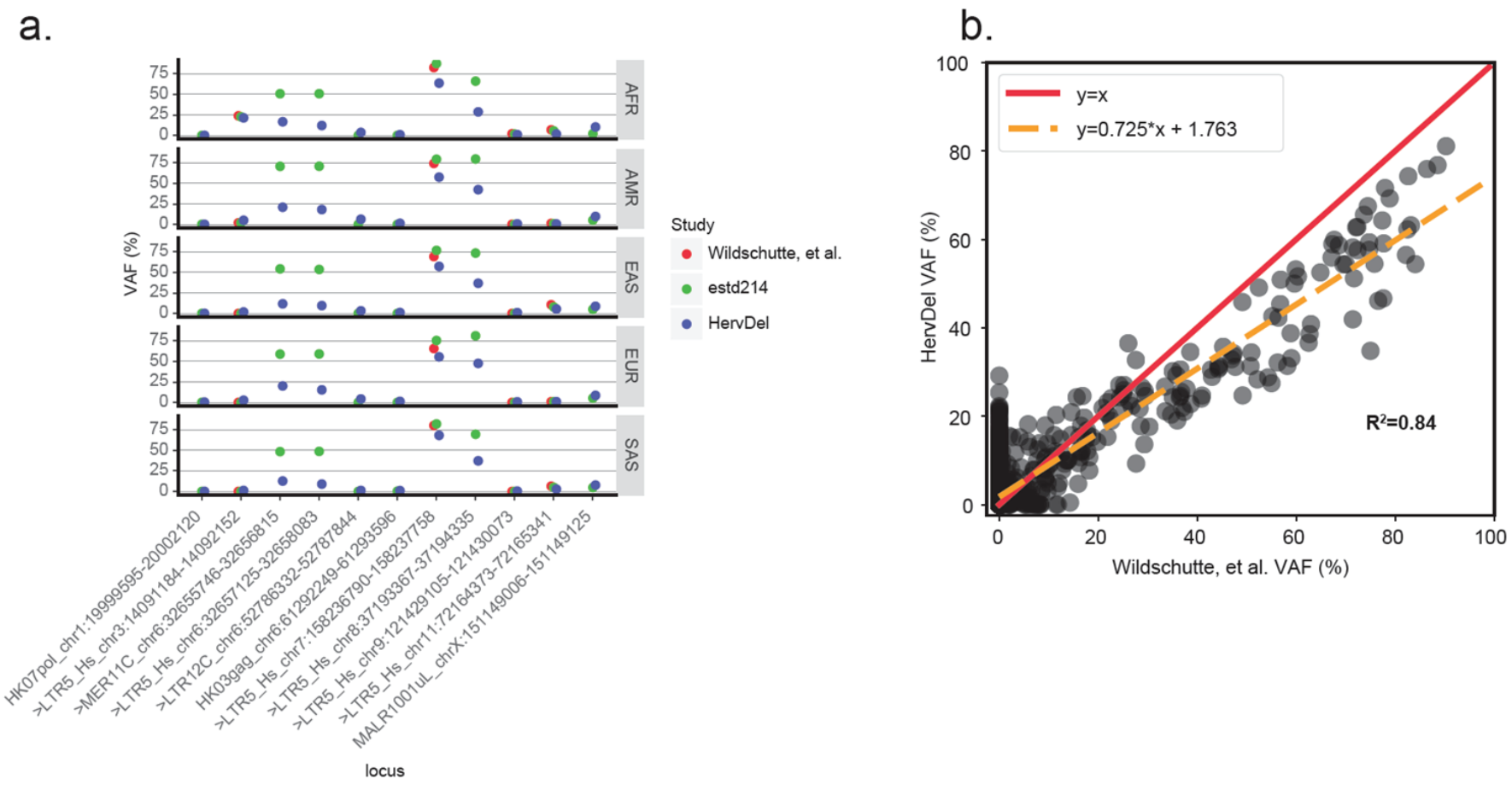

a.
b.
AFR
AMR
EAS
EUR
SAS
VAF (%)
75
50
25
0
Study
Wildschutte, et al.
estd214
HervDel
HK07pol_chr1:19999595-20002120
>LTR5_Hs_chr3:140911184-14092152
>MER11C_chr6:32655746-32656815
>LTR5_Hs_chr6:32657125-32658083
>LTR12C_chr6:52786332-52787844
HK03gag_chr6:61292249-61293596
>LTR5_Hs_chr7:158236790-158237758
>LTR5_Hs_chr8:37193367-37194335
>LTR5_Hs_chr9:121429105-121430073
>LTR5_Hs_chr11:72164373-72165341
MALR1001uL_chrX:151149006-151149125
locus
y=x
y=0.725*x + 1.763
R²=0.84
HervDel VAF (%)
Wildschutte, et al. VAF (%)
0
20
40
60
80
100

Figure 2

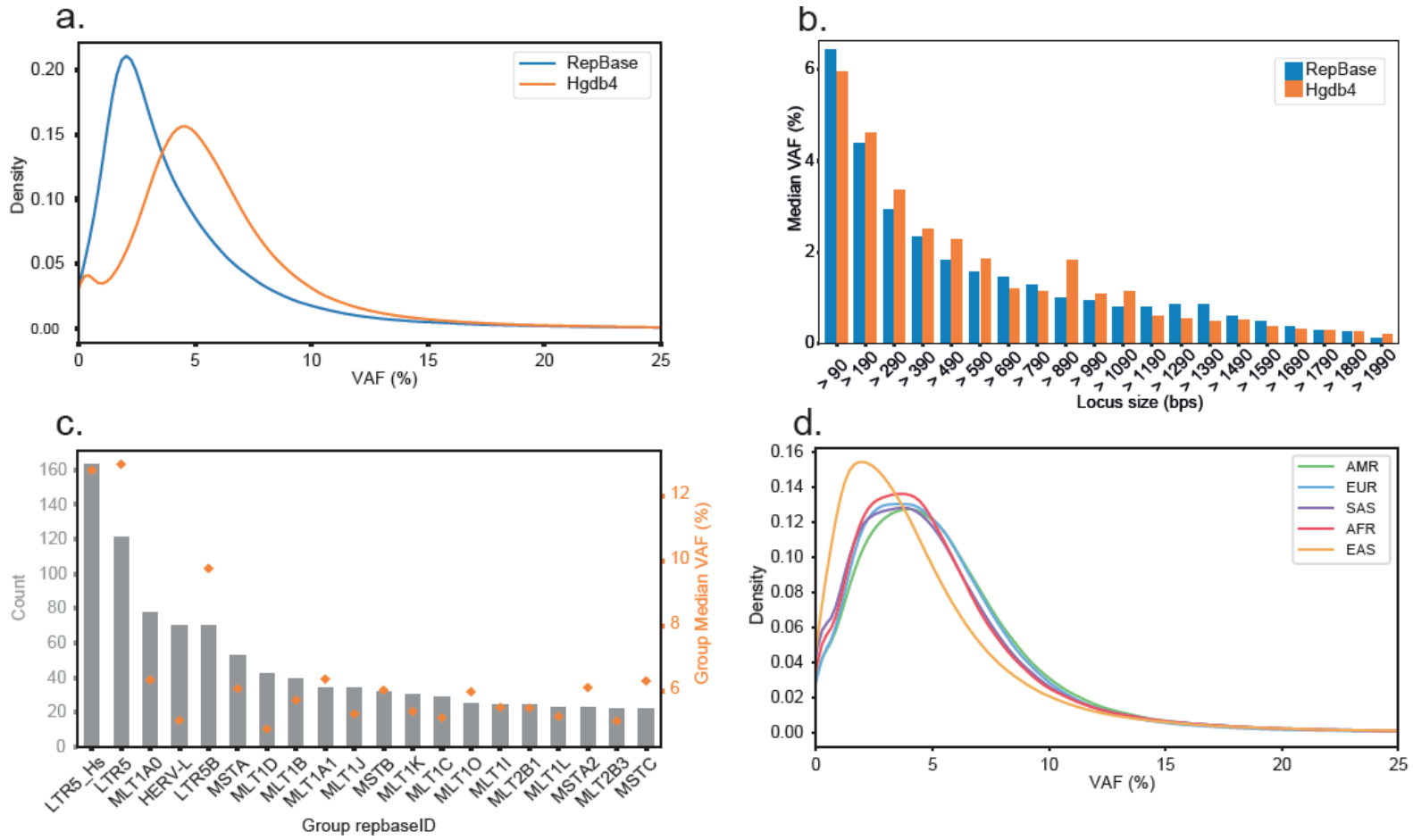

a.
RepBase
Hgdb4
Density
VAF (%)
b.
RepBase
Hgdb4
Median VAF (%)
Locus size (bps)
c.
Count
Group Median VAF (%)
Group repbaseID
LTR5_Hs
LTR5
MLT1A0
HERV-L
LTR5B
MSTA
MLT1D
MLT1B
MLT1A1
MLT1J
MSTB
MLT1K
MLT1C
MLT1O
MLT1I
MLT2B1
MLT1L
MSTA2
MLT2B3
MSTC
d.
AMR
EUR
SAS
AFR
EAS
Density
VAF (%)

Figure 3

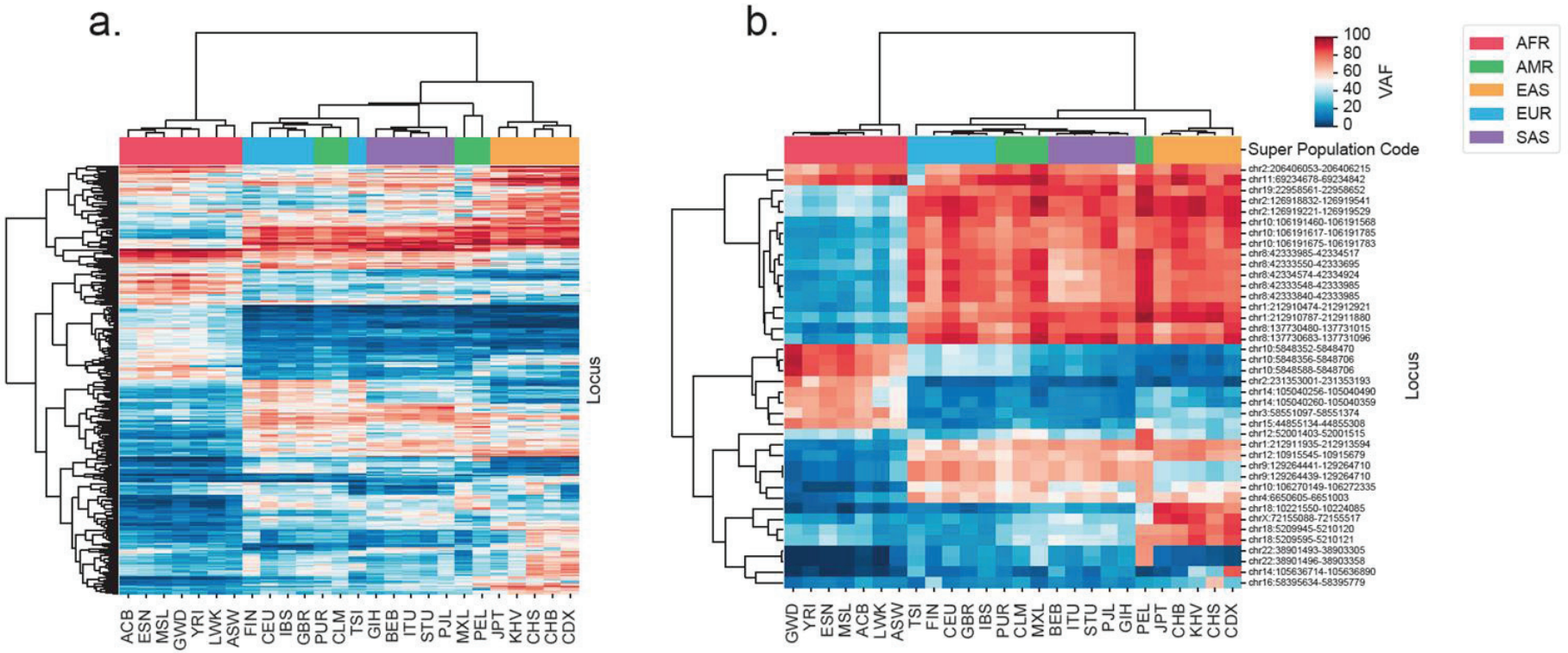
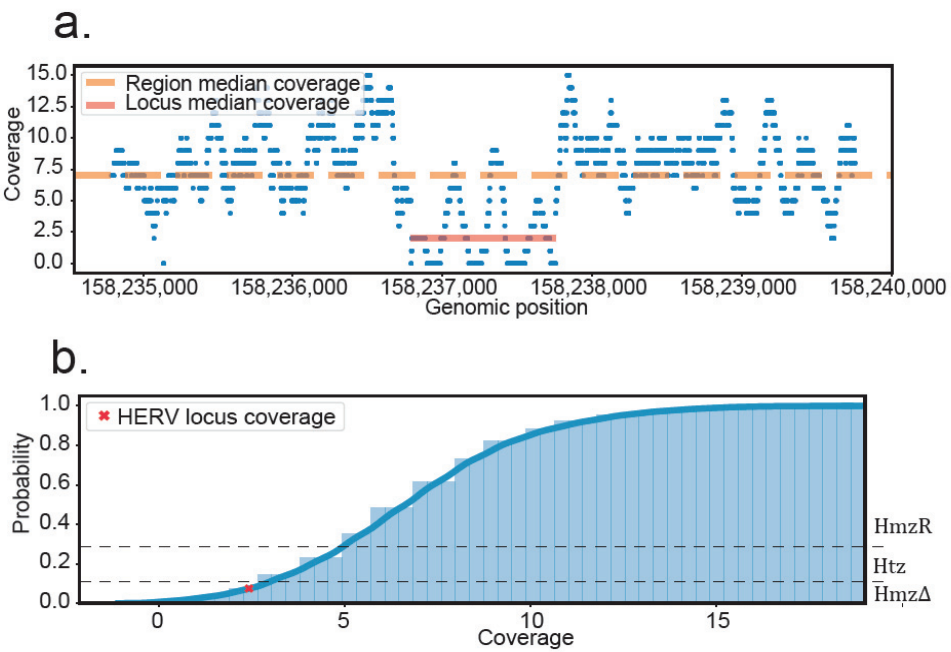
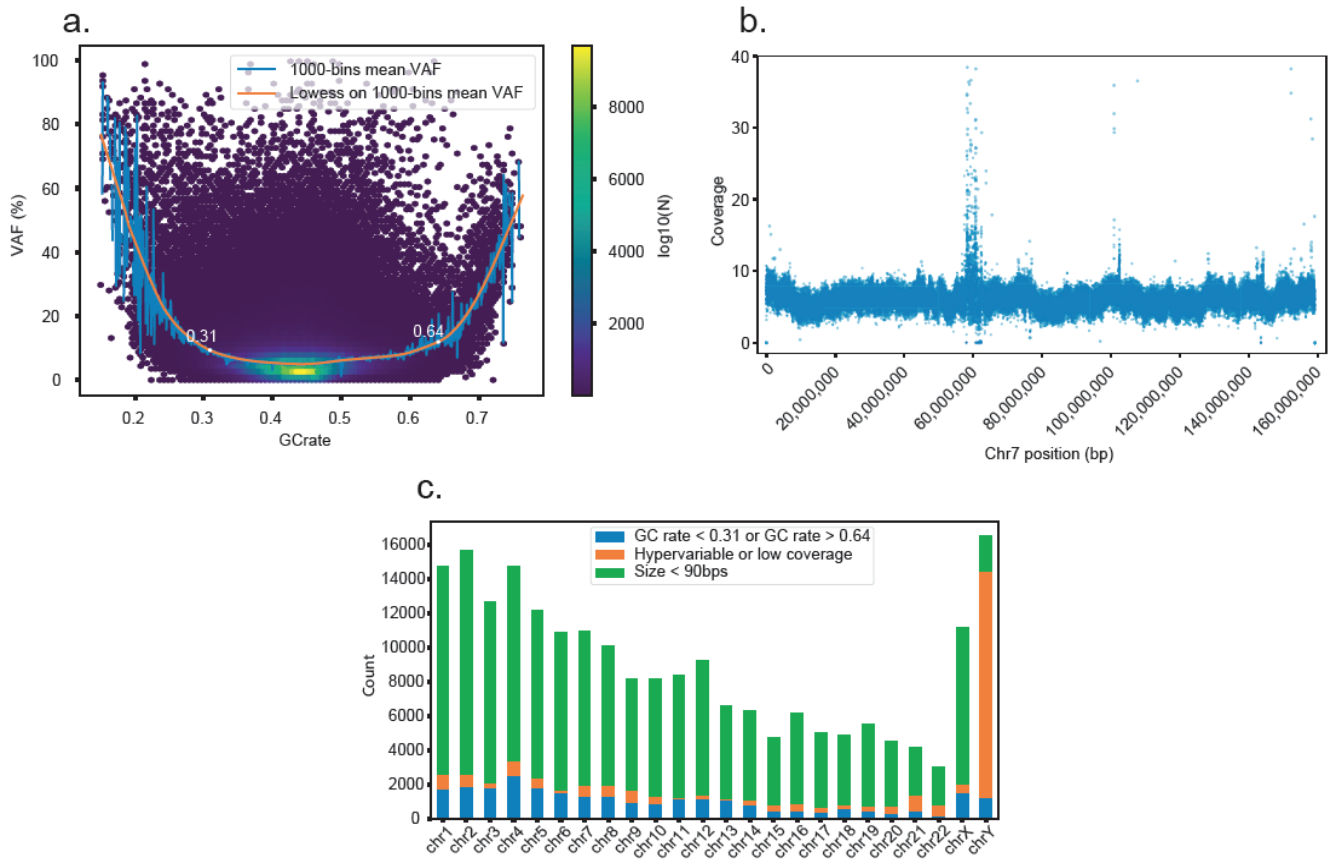

a.
Locus
ACB
ESN
MSL
GWD
YRI
LWK
ASW
FIN
CEU
IBS
GBR
PUR
CLM
TSI
GIH
BEB
ITU
STU
PJL
MXL
PEL
JPT
KHV
CHS
CHB
CDX
b.
100
80
60
40
20
0
VAF
AFR
AMR
EAS
EUR
SAS
Super Population Code
Locus
chr2:206406053-206406215
chr11:69234678-69234842
chr19:22958561-22958652
chr2:126918832-126919541
chr2:126919221-126919529
chr10:106191460-106191568
chr10:106191617-106191785
chr10:106191675-106191783
chr8:42333985-42334517
chr8:42333550-42333695
chr8:42334574-42334924
chr8:42333548-42333985
chr8:42333840-42333985
chr1:212910474-212912921
chr1:212910787-212911880
chr8:137730480-137731015
chr8:137730683-137731096
chr10:5848352-5848470
chr10:5848356-5848706
chr10:5848588-5848706
chr2:231353001-231353193
chr14:105040256-105040490
chr14:105040260-105040359
chr3:58551097-58551374
chr15:44855134-44855308
chr12:52001403-52001515
chr1:212911935-212913594
chr12:10915545-10915679
chr9:129264441-129264710
chr9:129264439-129264710
chr10:106270149-106272335
chr4:6650605-6651003
chr18:10221550-10224085
chrX:72155088-72155517
chr18:5209945-5210120
chr18:5209595-5210121
chr22:38901493-38903305
chr22:38901496-38903358
chr14:105636714-105636890
chr16:58395634-58395779
GWD
YRI
ESN
MSL
ACB
LWK
ASW
TSI
FIN
CEU
GBR
IBS
PUR
CLM
MXL
BEB
ITU
STU
PJL
GIH
PEL
JPT
CHB
KHV
CHS
CDX

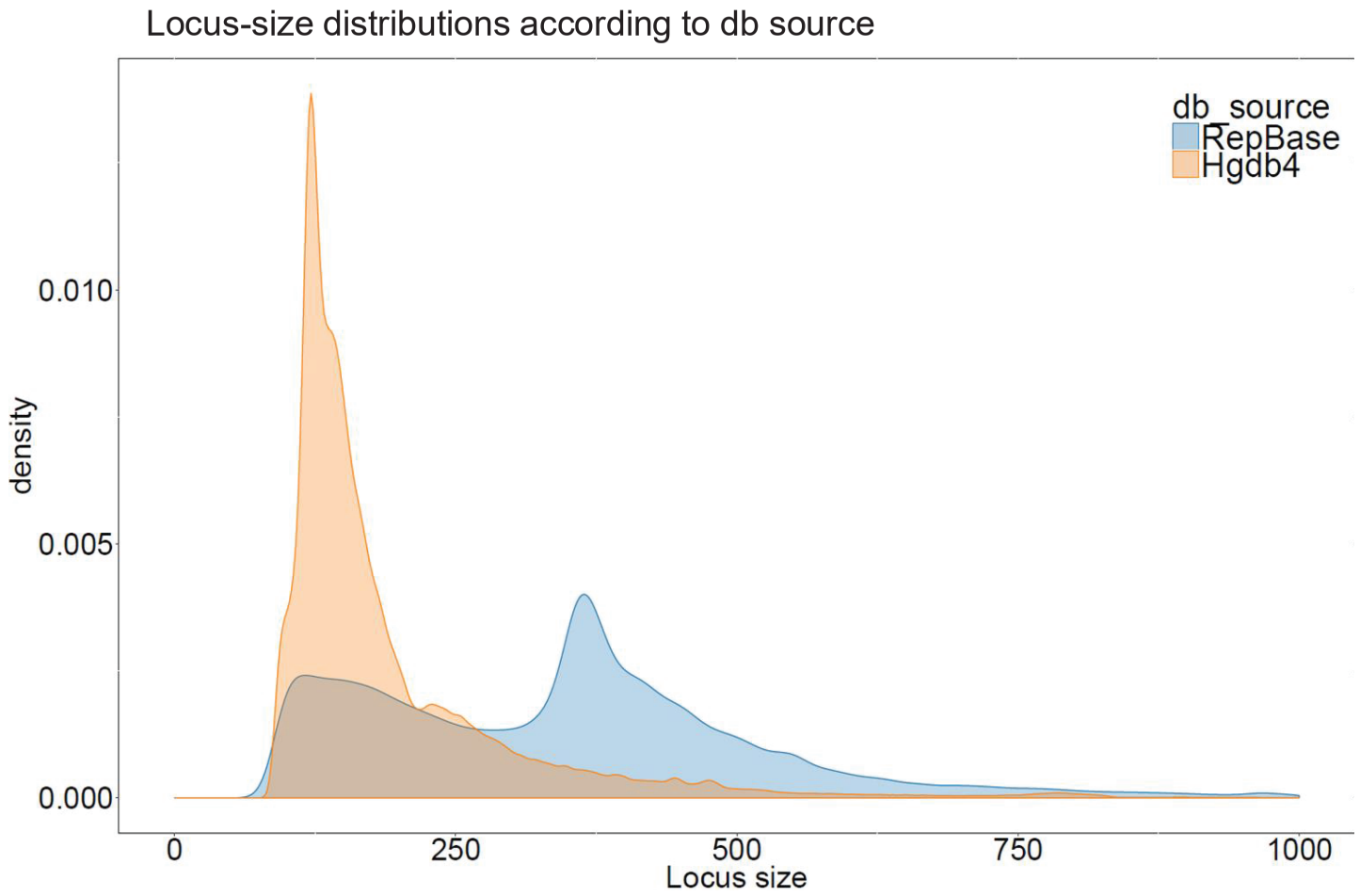
Figure 4



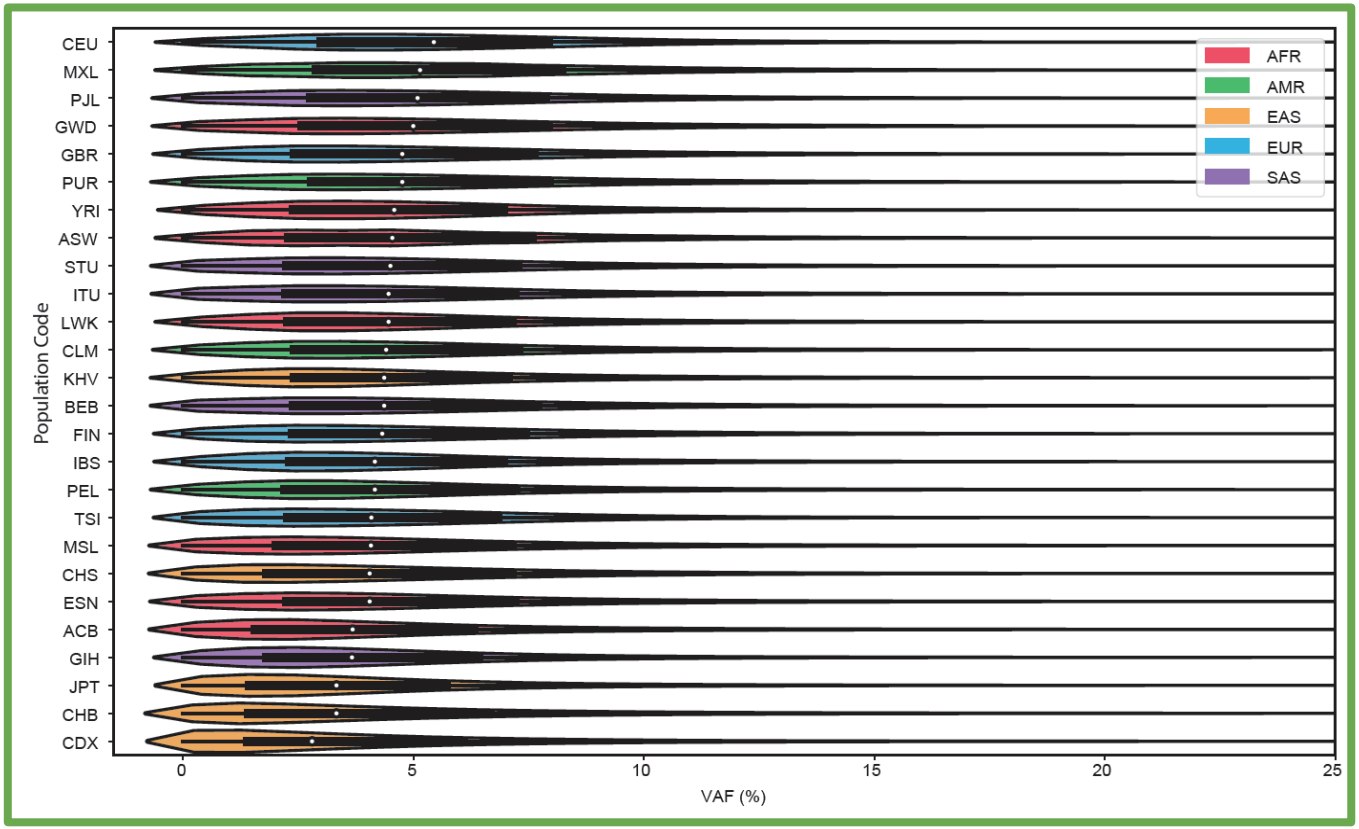
Supplementary figure 1



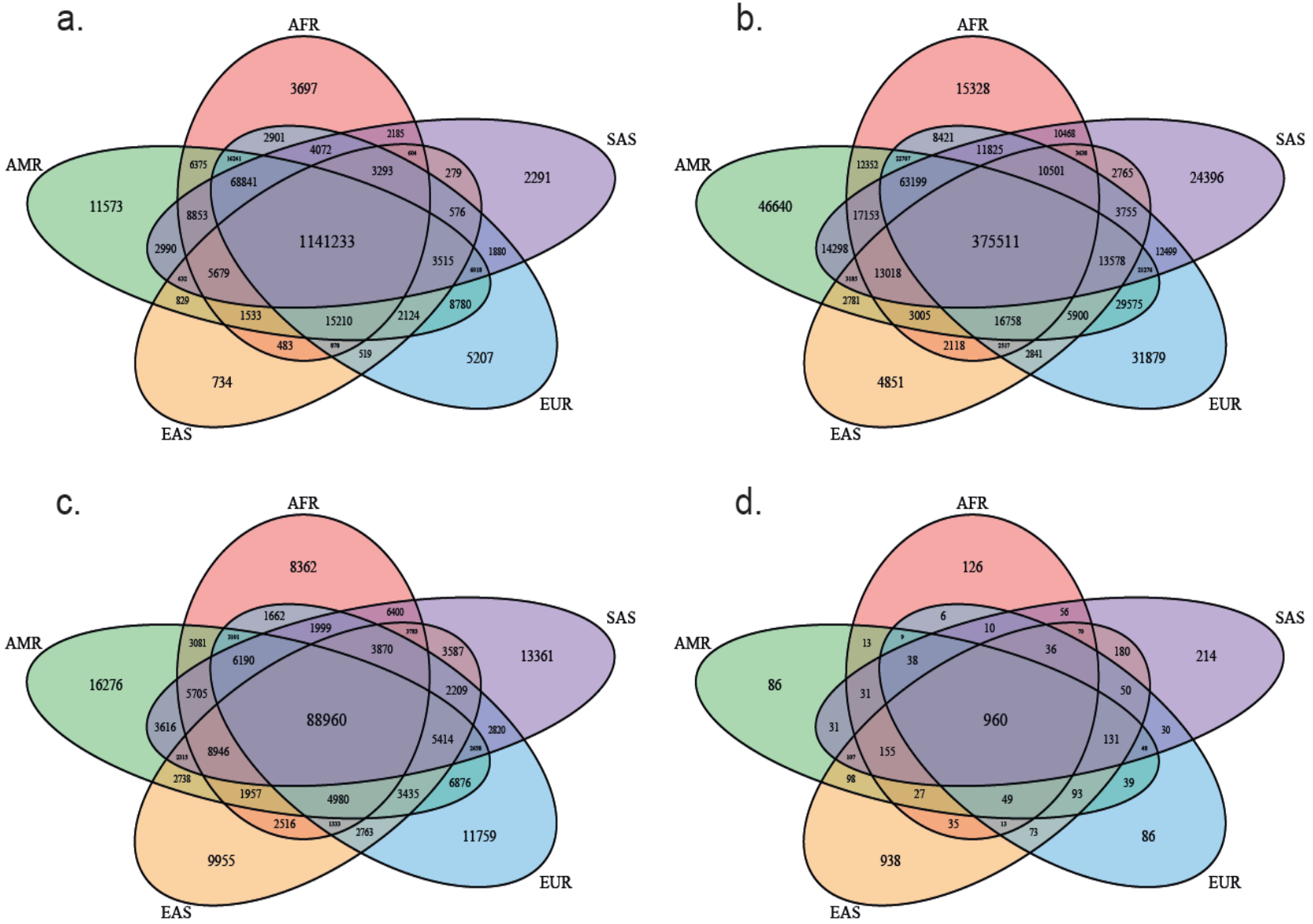
Supplementary figure 2



Supplementary figure 3.

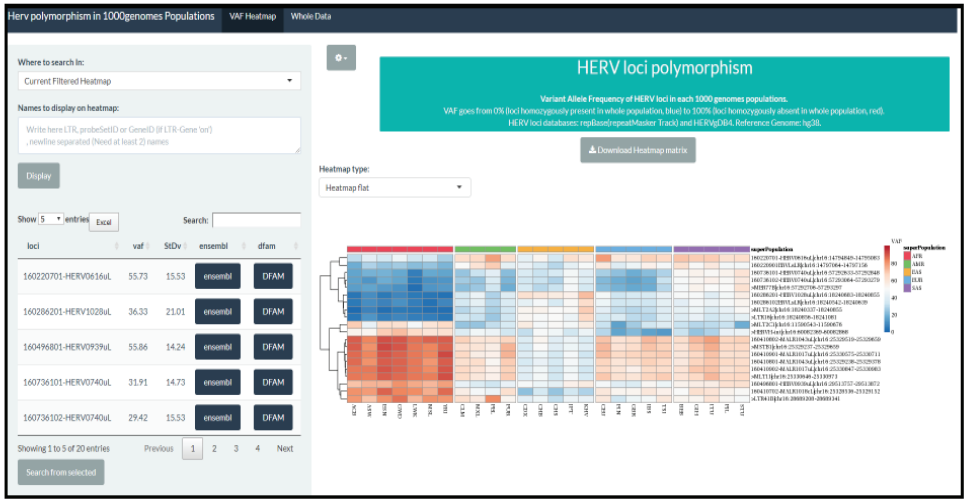


Supplementary figure 4



Supplementary figure 5

a.



b.

Supplementary figure 6

3.4 IMPACT DES HERV SUR LE TRANSCRIPTOME

3.4.1 IMPACT FONCTIONNEL DU POLYMORPHISME DE PRESENCE DES HERV SUR LEUR PROPRE EXPRESSION

Dans la section précédente (section 3.3), nous avons développé une méthode (HervDel) permettant d'évaluer les niveaux de présence ou d'absence des HERV dans la population humaine. Pour comprendre l'impact de ce polymorphisme sur le transcriptome HERV, nous étudions ce polymorphisme à la lumière de données d'expression.

Nous avons émis l'hypothèse que la variabilité importante observée dans l'expression des HERV pouvait être due à une présence non systématique de ces HERV dans les génomes des individus (Figure 3-16). A partir des résultats de fréquence allélique générés par la méthode HERVdel, et de données d'expression, nous souhaitons vérifier si ces informations sont corrélées, l'absence fréquente d'un HERV chez les individus d'une population devrait augmenter sa variabilité d'expression, voire réduire globalement son niveau d'expression.

3.4.1.1 ASSOCIATION ENTRE POLYMORPHISME ET DONNEES D'EXPRESSION SUR DES PATIENTS EN CHOC SEPTIQUE

Pour savoir s'il existe un lien entre la variabilité d'expression chez les HERV et leur fréquence allélique dans les populations humaines, nous avons cherché dans un premier temps à vérifier l'existence de corrélation entre la fréquence allélique d'absence des HERV (VAF) et leur expression dans le jeu de données de patients en choc septique MIP, généré avec la puce HERV-V3. Nous avons choisi cette cohorte car elle contient le plus d'individus, et donc une variabilité plus importante liée aux différences d'expression entre les individus. La Figure 3-17 représente la médiane d'expression (A) ou le coefficient de variation (B) des HERV en fonction de la VAF dans la population humaine. On peut conclure qu'il n'existe pas de corrélation entre VAF et variabilité d'expression après un choc septique sur cette cohorte. On peut également observer graphiquement sur la Figure 3-17 A que plus la fréquence de présence est important (VAF faible), plus le niveau expression est important. Cependant, la grande majorité des HERV n'est pas exprimée, et en regardant la densité du nuage de point,

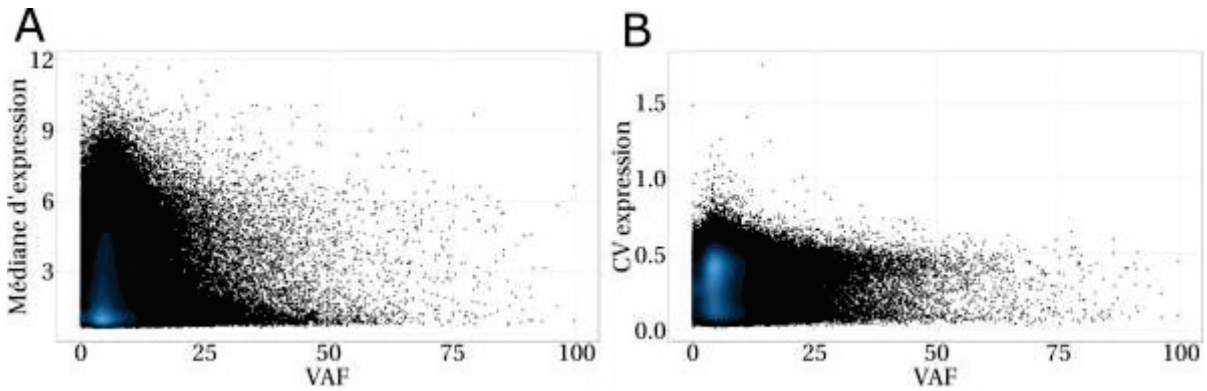


Figure 3-17 : Expression des HERV sur des patients en choc septique en fonction de la fréquence d'absence des HERV. A. Médiane d'expression des HERV en fonction de leur VAF. B. Coefficient de variation d'expression (CV) des HERV en fonction de leur VAF. Le CV est égal à l'écart-type divisé par la moyenne d'expression. Les données d'expression ont été obtenues à partir des échantillons de la cohorte MIP à J3 générée par la puce HERV-V3. Les nuances de bleus indiquent le niveau de densité de points, le bleu le plus clair représente la plus grande densité.

il suit une loi normale. Les patients d'où proviennent les données d'expression étant pour la plupart européens, nous avons également testé cette corrélation en prenant en compte la VAF de la super-population européenne, mais les courbes sont quasiment identiques (données non montrées). Nous avons également vérifié une possible corrélation entre expression et présence au niveau des groupes de HERV, mais là encore, sans association apparente.

Globalement, nous ne pouvons pas conclure à une corrélation entre fréquence allélique et expression de HERV dans le sang après un choc septique. Dans cette comparaison, les individus étudiés ne sont pas les mêmes et les conditions sont également différentes (volontaires sains dans les 1000 génomes contre des patients en choc septique dans MIP). Nous voulons donc maintenant vérifier ces corrélations avec des données d'expression provenant également des 1000 génomes (RNAseq), à partir des mêmes échantillons que pour l'étude du polymorphisme.

3.4.1.2 ASSOCIATIONS ENTRE PRESENCE ET DONNEES D'EXPRESSION A PARTIR DES MEMES INDIVIDUS SAINS PROVENANT DES 1000 GENOMES

J'ai récupéré un jeu de données publique de RNAseq réalisé à partir de 462 échantillons de volontaires provenant de 4 populations européennes et 1 population Africaine des 1000 génomes (CEU, FIN, GBR, TSI, YRI ; Détails dans la table S1 de l'article, section 3.3.2).

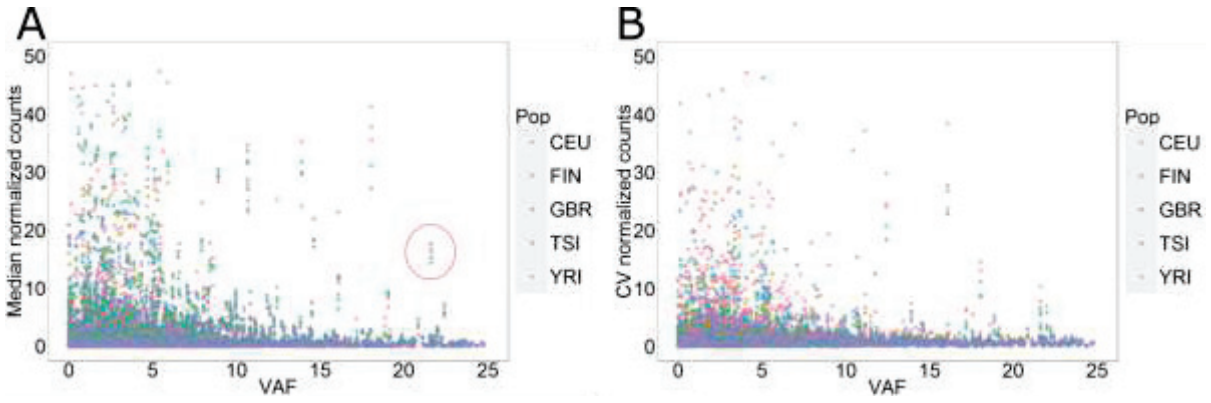


Figure 3-18 : Expression des HERV sur des échantillons des 1000 génomes en fonction de leur fréquence d'absence par population. A. Expression médiane en fonction de la VAF. B. Coefficient de Variation (CV) de l'expression en fonction de la VAF. Le cercle rouge représente 1 locus HERV intéressant en terme de valeurs de VAF et d'expression, ainsi que de localisation. Les 5 points dans le cercle correspondent aux valeurs médianes d'expression dans les 5 populations représentées, pour le même locus.

Le jeu de données est disponible sur https://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-3/.

Dans l'étude utilisant ces données (Lappalainen et al., 2013) les auteurs s'intéressent, entre autre, à l'expression des éléments répétés. Les quelques 7000 rétrotransposons à LTR (HERV et MaLR) parmi les éléments répétés qui sont exprimés dans le jeu de données ont été sélectionnés. On cherche à corréler l'expression et le polymorphisme de ces HERV. Dans un premier temps en comparant les niveaux d'expression avec les fréquences alléliques d'absence calculées avec HERVdel, pour chaque population ; dans un second temps en comparant l'expression et le niveau de couverture des HERV (percentile de la distribution de couverture à +/-25kb autour du HERV), individu par individu.

La Figure 3-18 représente les comptes normalisés des 7000 HERV exprimés dans le jeu de données RNAseq en fonction de leur fréquence allélique dans la population humaine. Nous avons en plus séparé chaque locus selon son niveau d'expression médian dans les populations correspondant aux échantillons (principalement européennes). Globalement, on ne voit pas clairement d'associations entre la fréquence allélique et le niveau d'expression, quelle que soit la population. Encore une fois, la majorité des loci se retrouvent avec une VAF et un niveau d'expression faibles, ce qui ne nous permet pas de conclure à une association sur l'ensemble des données. Bien que ce ne semble pas être un phénomène

global, il n'est pas exclu qu'une telle association entre génome et expression existe, sur plusieurs loci HERV.

Les loci présentant des valeurs élevées de VAF et d'expression peuvent se révéler intéressants. Par exemple, le locus entouré en rouge sur la Figure 3-18 A, possède un compte normalisé médian à 18, et une VAF à 21%. Ce locus est donc à la fois exprimé et est absent dans 20% des individus de la population européenne. De plus, ce locus a également été retrouvé absent dans l'étude du consortium 1000 génomes (Sudmant et al., 2015), qui avait pour but de détecter les variants structuraux du génome. Il appartient au groupe LTR5_Hs (apparenté aux HML-2) et localisé sur le chromosome 6, à 2kb en 3' du gène HLA-DQB1. Ce gène fait partie du système HLA, et est un complexe majeur d'histocompatibilité de type 2 (comme le HLA-DR), important dans la présentation des antigènes aux lymphocytes. Cette région est donc connue pour son rôle dans la réponse de l'hôte aux pathogènes et aussi pour son fort polymorphisme.

Pour identifier de manière reproductible des loci tels que décrit plus haut, nous avons testé les associations entre le niveau d'expression de loci HERV et leur génotype, au niveau de chaque individu, par deux approches. Pour rappel (papier section 3.3.2), la présence ou l'absence d'un locus HERV pour un individu est déterminée en comparant le niveau de couverture en read de séquençage ADN du locus avec son environnement (+/- 25kb). On détermine ainsi le percentile de la couverture du HERV dans la distribution de couverture de la région de 50kb. Si le niveau de couverture du locus HERV correspond à un percentile inférieur à 0,107, il est alors classé comme absent sur les deux allèles (HmzD), si le percentile est compris entre 0,107 et 0,275 le HERV est classé comme présent sur un seul des 2 allèles (Htz), sinon il est classé comme présent (HmzR). Nous avons ainsi récupéré les valeurs de percentile et le génotype associé des 7000 loci HERV exprimés dans le jeu de données RNAseq, pour chacun des 462 échantillons (qui sont en commun entre données génomiques et données transcriptomiques). Nous avons ensuite testé les corrélations ou associations entre le niveau d'expression de chaque locus, pour chaque échantillon en fonction du percentile (variable continue), ou du génotype (variable discrète), respectivement. Sur l'ensemble des loci, aucune corrélation (percentile) ou association (génotype) n'étaient visibles (données non montrées). Nous voulons identifier les HERV exprimés dans les

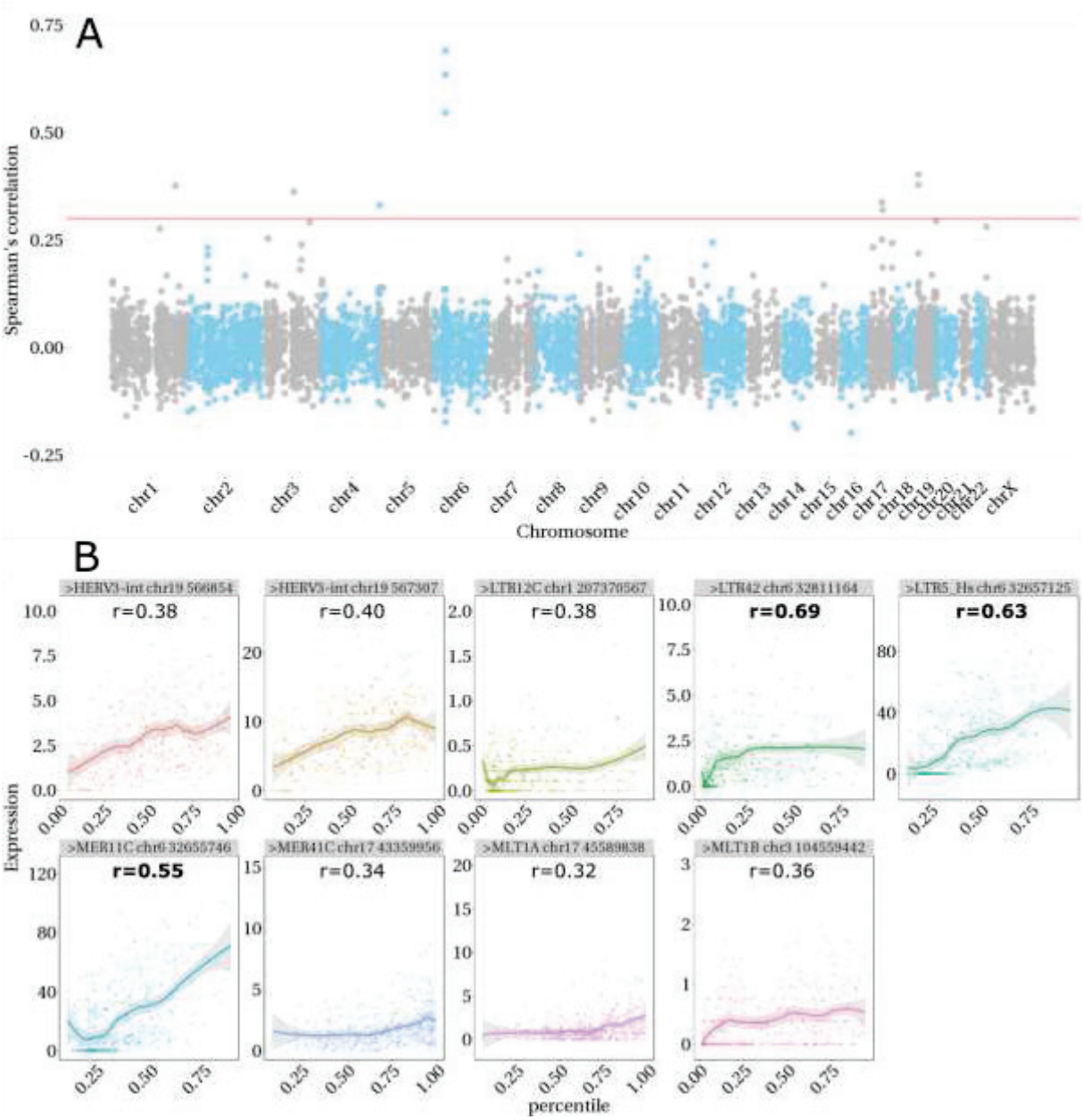


Figure 3-19: Corrélation entre expression et présence des HERV. A. Graphique de Manhattan. Chaque point représente un des 7000 HERV exprimé dans le jeu de données RNAseq. L'axe des x représente la position de chaque HERV dans le génome. L'axe des Y représente le coefficient de corrélation de Spearman entre le niveau d'expression et le percentile du HERV dans la distribution de couverture (des reads provenant du séquençage ADN) de la région génomique environnante (+/- 25kb). Plus le percentile du HERV est bas, plus il a de chances d'être absent dans le génome de l'individu. **B. Nuages de point entre niveau d'expression et probabilité de présence des 9 HERV ayant un coefficient de Spearman supérieur à 0,3.** L'axe des x représente le percentile de couverture des reads de séquençage ADN correspondant au HERV par rapport à la distribution de couverture de son environnement génomique proche (+/-25kb). Plus le percentile est petit, plus le HERV a de chances d'être absent. Pour rappel, les seuils appliqués pour l'absence sur les 2 allèles et sur 1 seul allèle est de 0,107 et 0,275 respectivement.

conditions physiologiques et qui ne sont pas présents chez tous les individus. Nous identifions les loci dont l'expression est impactée selon leur présence ou absence. Nous sélectionnons d'une part les loci qui ont le plus grand coefficient de corrélation de Spearman entre niveau d'expression et percentile de la distribution de couverture pour chaque HERV. D'autre part, nous sélectionnons ceux ayant l'association la plus significative entre les niveaux d'expression et le génotype des individus (HmzD, Htz ou HmzR) en réalisant des tests ANOVA.

La Figure 3-19 représente les résultats de l'étude de corrélation entre percentiles et niveaux d'expression. Les percentiles représentent la valeur du niveau de couverture en reads du HERV par rapport à la région environnante de +/-25kb. Si le percentile est bas, le locus a de fortes chances d'être absent. A l'inverse s'il est haut, il est présent. En ne considérant que les coefficients de Spearman entre expression du HERV et valeur du percentile supérieurs à 0,3, on voit que seuls 9 loci ont un coefficient de corrélation supérieur à cette limite, dont 3 avec un coefficient supérieur à 0,5 (Figure 3-19 A). Cela montre que la très large majorité des HERV exprimés n'est pas influencée par leur polymorphisme de présence. Les HERV avec le meilleur coefficient montrent qu'il existe sur certains loci, une corrélation entre expression et présence, notamment pour les éléments avec fort taux d'absence. Pour chacun des 9 loci au-dessus du seuil, on peut observer que, entre 0,1 et 0,5 de percentile, le niveau d'expression et le percentile corrèlent bien (Figure 3-19 B). Les 3 HERV avec un coefficient de Spearman supérieur à 0,5 sont tous localisés sur le chromosome 6, au niveau du système HLA. Le locus, avec un coefficient égal à 0,63, est le même locus identifié sur la Figure 3-18, un HERV du groupe LTR5_Hs, situé proche du gène HLA_DQB1. De manière intéressante, on observe aussi sur la plupart des loci, un nombre important d'échantillons qui ont une expression nulle quand le percentile est inférieur à 0,25. Cette observation conforte le fait que le HERV, avec un percentile faible, est absent.

La Figure 3-20 représente les résultats du test d'association entre génotypes et niveaux d'expression (ANOVA). Contrairement à la figure précédente, nous avons appliqué les seuils de génotype définis dans l'article de la section 3.3.2 (HmzD, Htz ou HmzR). En prenant un seuil de p-value (corrigée pour le multi-testing) à 0,05, nous observons seulement 6 HERV dont

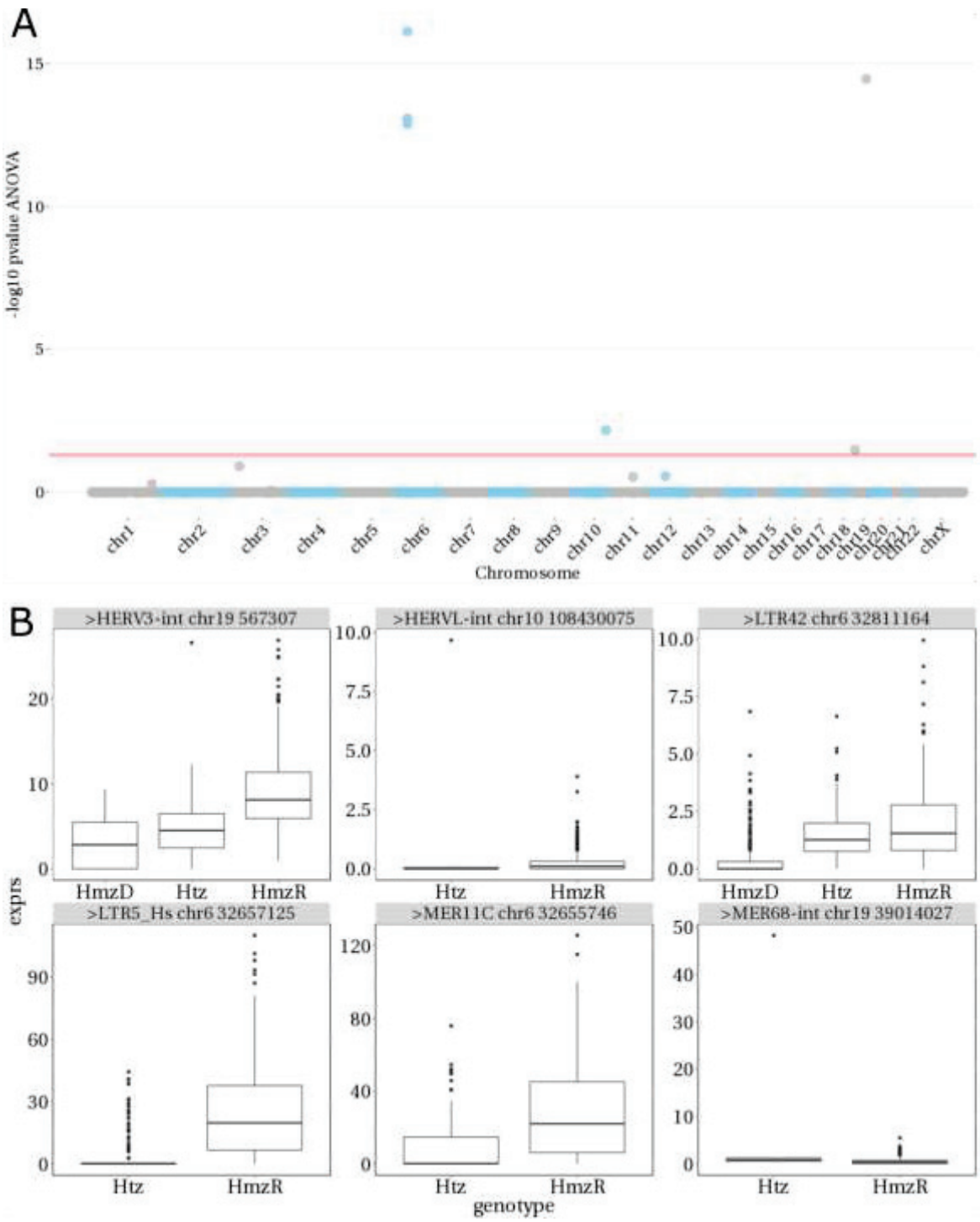


Figure 3-20 : Association entre génotype HERV et expression des HERV. A. Graphique de Manhattan. Chaque point représente un des 7000 HERV exprimé dans le jeu de données RNAseq. L'axe des x représente la position de chaque HERV dans le génome. L'axe des Y représente la p-value en –log10 de l'ANOVA testant l'association entre Génotype HERV et expression. La ligne horizontale rouge correspond au seuil de significativité de p-value de 0,05. L'ensemble des p-values ont été ajustées pour le multi-test par la méthode de Benjamini-Hochberg. **B. Boxplot des 6 HERV dont l'expression est associée au génotype.** L'axe des x représente les génotypes de chaque individus (HmzD, Htz, HmzR), l'axe des y le niveau d'expression. Le titre de chaque graphique correspond dans l'ordre au groupe de l'élément, son chromosome et sa position de départ. HmzD correspond à Homozygote Déleté, (le HERV est absent sur les 2 allèles), Htz correspond à Hétérozygote, et HmzR correspond à Homozygote référence (le HERV est présent sur les 2 allèles).

l'expression est significativement associée au génotype des individus pour ce HERV (Figure 3-20 A). Ces 6 HERV font partie des 9 HERV précédemment trouvés. La Figure 3-20 B représente leur profils d'expression en fonction du génotype.

Pour conclure sur cette partie, seule une minorité des HERV voient leur expression impactée par leur présence ou non, dans ces contextes. On rappelle tout de même que le jeu de données des 1000 génomes contient des échantillons d'individus sains, rien n'exclut que dans d'autres contextes, et notamment après un choc septique où plus de HERV doivent être exprimés, on ne trouve pas plus de telles associations. Reste que sur le peu d'associations détectées, la quasi-totalité des HERV se trouvent dans la région HLA. Ceci peut laisser penser qu'il existe des liens étroits entre réponse immunitaire et HERV. Il sera intéressant d'étudier plus en détail ces loci.

3.4.1.3 ANALYSE D'ENRICHISSEMENT FONCTIONNEL SUR LES 1000 GENOMES

Dans l'article de Lappalainen et al. (Lappalainen et al., 2013) à partir des 1000 génomes, les auteurs ont fait des analyses d'eQTL (expression Quantitative Trait Loci). Le but est de trouver des associations entre SNP à une position donnée du génome et expression d'un gène (pouvant être proche en cis, ou plus éloignée, en trans). Afin de savoir si il y a un enrichissement en SNP localisés dans un HERV, nous récupérons tous les SNP associés à l'expression d'un gène en cis à partir de ce jeu de données. On filtre ensuite en ne prenant que les SNP situés dans des HERV. Au total, 1481 SNP impliqués dans un eQTL sont localisés dans un HERV (6,7% du nombre total de SNP impliqués dans un eQTL). Comparé aux 8% de HERV composant le génome, il ne semble pas y avoir globalement un enrichissement en HERV impliqués dans des eQTL.

Pour savoir ensuite quels types de gènes (d'un point de vue fonctionnel) sont associés aux SNP impliqués dans les HERV, nous réalisons une analyse d'enrichissement en réseaux fonctionnels (avec Ingenuity Pathway Analysis). Le réseau de gènes qui sort le plus enrichi est la voie de présentation des antigènes (comportant des gènes du système HLA et PSMB8). Cette région est déjà connue pour son fort taux de polymorphisme, donc le résultat n'est pas étonnant. Malgré tout, ce résultat nous apporte un nouvel élément indiquant que les HERV situés dans cette région sont également polymorphiques et pourraient jouer un rôle

important dans la réponse de l'hôte aux pathogènes, en influençant l'expression des gènes dans la région HLA.

3.4.2 IMPACT DU POLYMORPHISME DE PRESENCE DES HERV SUR L'EXPRESSION DES GENES, SUR DES INDIVIDUS SAINS

Nous savons que les HERV peuvent influencer l'expression de gènes situés à proximité dans le génome, soit en agissant comme promoteur, soit comme fournisseur de sites de polyAdénylation. Mais les HERV peuvent également impacter l'expression de gènes d'autres manières, notamment en fournissant des sites de fixation de facteurs de transcription (ou de restriction) augmentant ou réduisant l'expression du gène voisin. On cherche ainsi à savoir si la présence ou l'absence de HERV modifie l'expression de gènes voisins. Pour cela, nous avons mis en place une approche de type cis-eQTL, permettant de chercher les associations entre génotype HERV et niveau d'expression du gène proche.

Le graphique de Manhattan représente le niveau de significativité de l'association entre génotypes HERV et expression du gène (Figure 3-21 A). Plus de 2400 paires réparties sur tous les chromosomes sont significativement associées (p-value <0,01). De manière étonnante, sur 4 paires parmi les 6 paires HERV / gène avec la plus faible p-value, on remarque que l'absence du HERV est associée avec une plus forte expression du gène (Figure 3-21 B, transcrits HLA-DRB6, GSTM1, UGT2B17 et MAPK8P1P1). Autrement dit, quand le HERV est absent (HmzD), ou qu'il est présent à l'état hétérozygote (Htz) chez un individu, le gène situé proche est plus fortement exprimé que lorsque le HERV est présent à l'état homozygote (HmzR). Les 2 autres associations semblent mettre en évidence des associations ponctuelles entre génotype et expression du gène (transcrits CAVIN2 et LINC02460), et ne reflètent pas un effet systématique. Parmi les 4, on retrouve une paire située dans la région HLA. Le HERV appartient au groupe HML-3 (super-famille des HERV-K) et est associé à l'expression du gène HLA_DRB6, situé à 65pb en amont du HERV. Le HERV n'est jamais totalement absent chez les individus, mais l'état hétérozygote de présence du HERV est associé à une plus forte expression du gène que lorsque le HERV est présent à l'état homozygote. Plus généralement, parmi les paires significativement associées, 13 se situent dans la région HLA, impactant l'expression des gènes HLA-DRB6, HLA-DQB1, HLA_DPA3 et HLA-K (1 gène peut être associé à plusieurs HERV). On retrouve le HERV, du groupe LTR5_Hs (super-famille HERV-K) et

proche de HLA-DQB1, que nous avons trouvé exprimé dans le sang quand il est lui-même présent à l'état homozygote (Figure 3-20 B, en bas à gauche). La présence homozygote de ce HERV est également associée à une expression plus importante du gène HLA-DQB1 par rapport à sa présence hétérozygote. Juste en amont de ce HERV, on observe un autre HERV, du groupe MER11C (super-famille HERV-K), dont la présence à l'état homozygote est fortement associée à une diminution de l'expression du gène HLA-DQB1. La Figure 3-22 illustre l'annotation génomique des loci HERV et des transcrits et résume les résultats d'associations dans la région du gène HLA-DQB1. De manière intéressante, la présence à l'état homozygote du locus HML-2 est associé à une augmentation à la fois de l'expression du locus HML-2 lui-même (Figure 3-20, Figure 3-22) et du gène HLA_DQB1 (données non montrées). Nous observons également que les 2 loci HERV sont déjà détectés comme absents chez certains individus dans l'étude des variants structuraux des 1000 génomes (fréquence allélique d'environ 50% dans les super-populations européenne, africaine, d'Asie du sud et de l'est (Consortium, 2010)). De plus, un site de fixation du facteur de transcription CTCF est présent dans le HERV MER11C. CTCF est impliqué, en autres, dans le repliement de l'ADN. Enfin, une séquence EST a été détectée jusqu'au locus HML-2, suggérant qu'un des transcrits du gène HLA-DQB1 possède une 3'UTR plus longue et se termine au niveau de la LTR, qui pourrait contenir un signal de polyAdénylation (Un motif AATAAA se trouve à 80pb de la fin de la séquence EST détectée).

Nous mettons donc en évidence un nombre important d'associations entre génotype HERV et expression de gènes proches. Les associations les plus significatives montrent pour la plupart, que la présence du HERV entraîne une diminution de l'expression du gène. Nous avons porté plus spécifiquement notre attention sur des paires de la région HLA, et notamment proche du gène HLA-DQB1, montrant un potentiel rôle des HERV situés en 3'. Ces résultats, bien que révélés à partir d'individus sains, suggèrent que la présence de ces HERV modifient le contexte génomique et empêchent l'expression du gène situé à proximité. Il sera intéressant pour la suite du projet de confirmer ces hypothèses et de chercher à mieux comprendre les mécanismes qui entrent en jeu (site de fixation de facteur de répression de l'expression, repliement ou compaction de l'ADN, …) et des éventuels impact sur la réponse immunitaire.

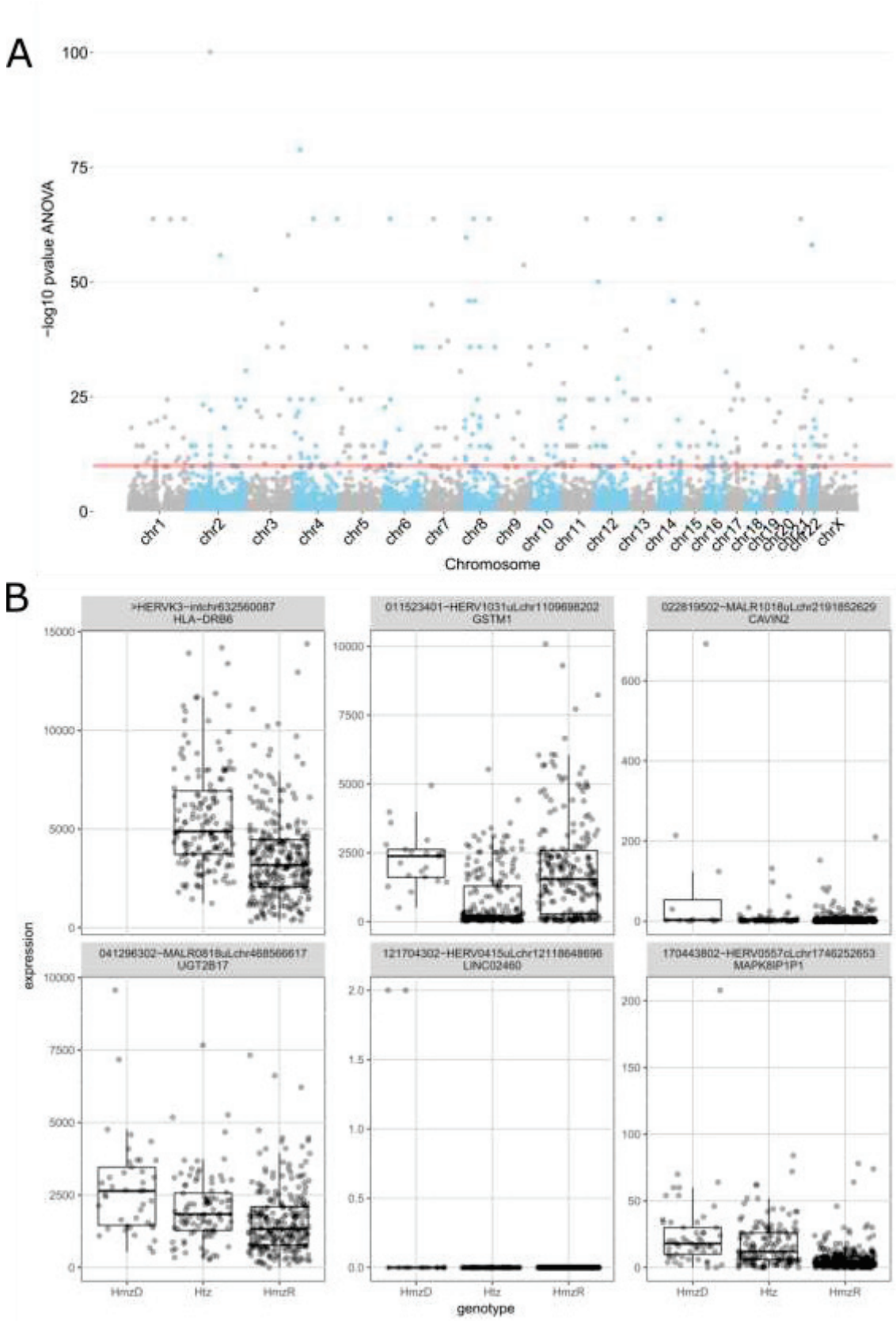


Figure 3-21 : Associations en cis entre présence de HERV et expression des gènes. A. Graphique de Manhattan. Chaque point représente une paire LTR / gène situés à une distance inférieure à 10kb. L'axe des x représente la position de chaque paire dans le génome. L'axe des Y représente le niveau de significativité statistique de l'ANOVA testant l'association entre Génotype HERV et expression du

gène (test ANOVA, p-values ajustées pour le multi-testing par la méthode de Benjamini Hochberg, exprimées en –log10). La ligne horizontale rouge correspond au seuil de significativité de p-value ajustée de $1{,}10^{-10}$. **B. Top 6 des associations.** L'axe des x représente les génotypes de chaque individus (HmzD, Htz, HmzR), l'axe des y le niveau d'expression du gène. Le titre de chaque graphique correspond en haut à l'identifiant du locus HERV (groupe de l'élément, chromosome et position de départ), en bas au nom du gène.

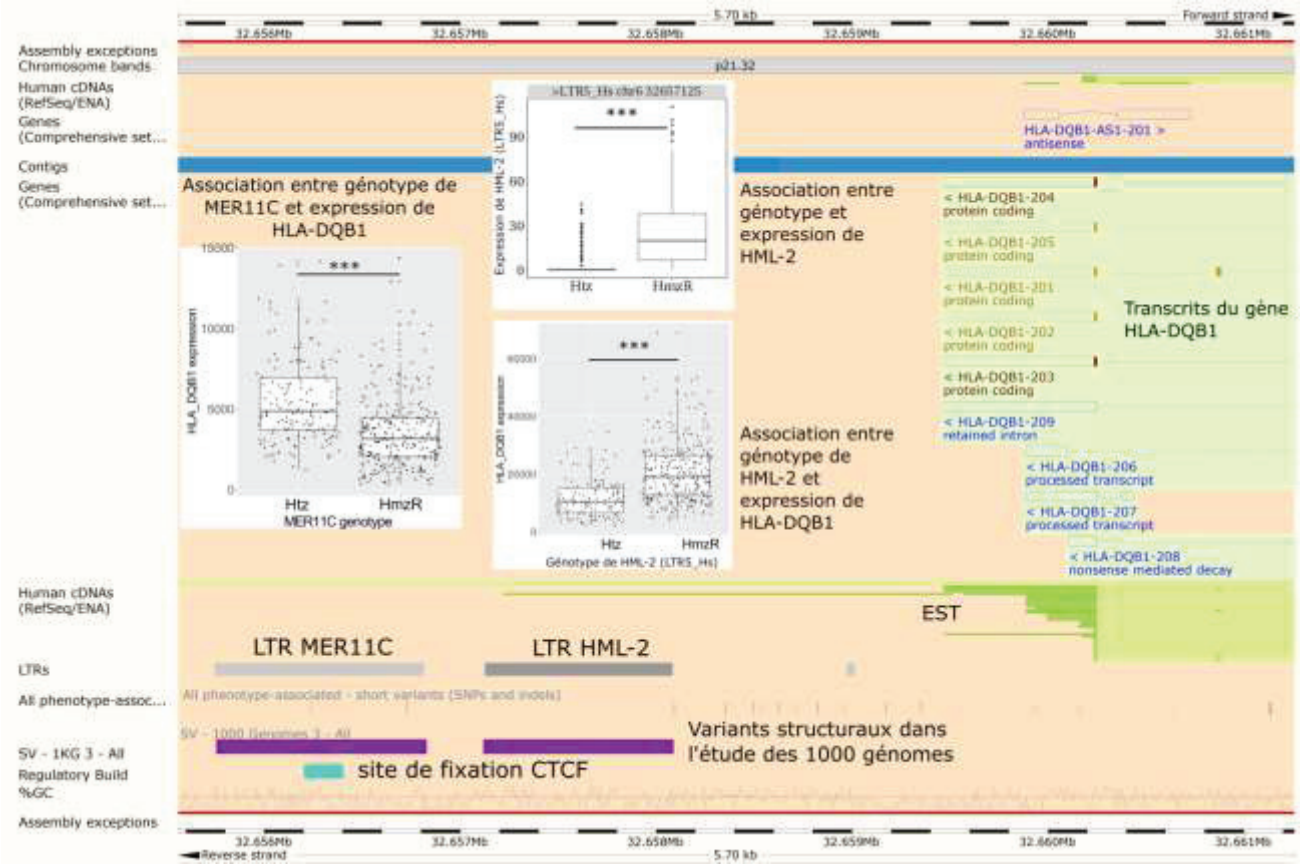


Figure 3-22: Région 3' de HLA_DQB1. Région génomique (chr6:32655560-32661255) représentant la 3' UTR du gène HLA_DQB1 (brin reverse, à droite) et sa région 3'. Les rectangles rouge et orange (codant pour une protéine) et bleus (non codant) représentent les 3'UTR de différents transcrits du gène. Les rectangles gris représentent la position des deux LTR, MER11C à gauche et HML-2 à droite. Les 2 rectangles violets indiquent que ces deux LTR ont été annotées délétées sur certains individus des 1000 génomes. Le rectangle turquoise représente le site de fixation du facteur de transcription CTCF. Les traits verts représentent des séquences EST mappées sur cette région. Image tirée de Ensembl (ensembl.org). Le boxplot à gauche montre la différence significative d'expression du gène HLA-DQB1 en fonction de la présence homozygote (HmzR) ou hétérozygote (Htz) de la LTR MER11C. Celui à droite en haut présente l'association trouvée précédemment entre génotype du HERV HML-2 (ou LTR5_Hs) et sa propre expression. Le boxplot à droite en bas représente la différence significative d'expression du gène HLA-DQB1 en fonction de la présence HmzR ou Htz de la LTR HML-2

3.4.3 IMPACT DE L'EXPRESSION DES LTR SUR L'EXPRESSION DES GENES DE LA REPONSE IMMUNITAIRE

Nous venons de mettre en évidence que des HERV, par leur présence ou absence, peuvent influencer l'expression de gènes situés proches dans le génome. Nous avons également vu dans la section 3.2, que des HERV sont exprimés dans des contextes liés à la réponse immunitaire (modèle mimant ou cohortes de choc septiques). D'autre part, la majorité des HERV annotés dans le génome sont des LTR solo. Les LTR sont les centres de contrôle de l'expression des gènes viraux chez les rétrovirus, et des preuves existent sur le rôle de certaines LTR dans l'expression de gènes conventionnels, notamment dans le cancer, grâce à un rôle promoteur ou de terminaison des transcrits (polyAdénylation). Dans cette partie, nous souhaitons savoir s'il existe des LTR, dont l'expression peut jouer un rôle promoteur ou de terminaison de transcrits, sur l'expression de gène de l'immunité dans le sang, suite à un choc septique.

3.4.3.1 FONCTIONS POTENTIELLES DES LTR DANS LES DIFFERENTS JEUX DE DONNEES

3.4.3.1.1 FONCTIONS DANS L'ENSEMBLE DES ECHANTILLONS

Pour rappel, l'attribution d'une fonction potentielle pour une LTR se fait sur la dichotomie des signaux observés à partir de la puce HERV-V3 dans les sous-régions de la LTR (méthodes des papiers sections 3.2.2.4 et 3.2.4.2). Pour résumer, une LTR se divise en sous-régions U3, R et U5 (Figure 1-6). Pour les LTR avec les sous-région U3 et U5 ciblées par la puce, si un signal d'expression est détecté sur la région U5 mais pas U3, on pourra attribuer un rôle promoteur du transcrit (Pr, le transcrit démarre à la sous-région R), au contraire si le signal est observé uniquement sur la sous-région U3, on attribuera un rôle de terminaison du transcrit (pA, le transcrit s'arrête à la sous-région R). Si le signal est identique et non nul dans les 2 sous-régions, on attribue un état « Read Through » à la LTR, l'expression passe seulement par la LTR (RdT), sinon la LTR est silencieuse (No). Les analyses sur les fonctions des LTR prototypes (bien annotées) se trouvent dans les 2 papiers précédemment cités. Pour résumer, nous avons montré que la majorité des LTR ne sont pas transcrites, qu'environ un quart des LTR ont une fonction potentielle soit promotrice, soit de polyadénylation et que 10% peuvent passer d'un état silencieux à un état actif entre différents échantillons. Ici, je

présente les analyses des fonctions sur l'ensemble des LTR de la puce (HERV Dfam, MaLR Dfam et HERV prototypes), à partir des 3 jeux de données générés avec la puce HERV-V3, ET , IS et MIP. La Figure 3-23 illustre les répartitions des fonctions des LTR dans chacun des jeux de données. On voit tout d'abord que les proportions relatives sont très similaires entre les jeux de données. On voit que la grande majorité des LTR sont silencieuses (entre 80 et 85%). On observe également des proportions similaires de fonctions pA et Pr dans chaque jeu de données, avec des proportions légèrement supérieures de pA. Certaines de ces LTR ont été validées par qRT-PCR par Marine Mommert (Figures S6 des articles sections 3.2.2.4 et 3.3.2), validant ces observations.

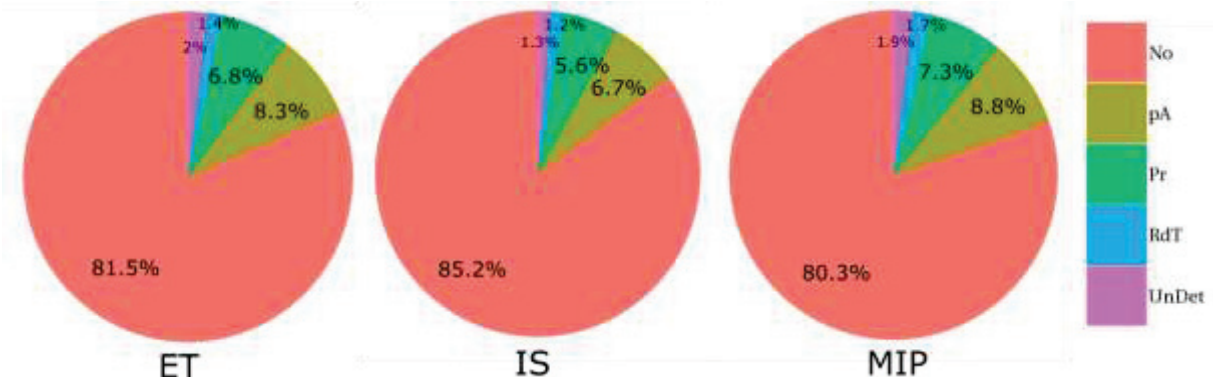


Figure 3-23 : Répartition des états des LTR dans les 3 jeux de données de la puce HERV-V3. Les paramètres utilisés pour l'attribution de fonctions aux LTR sont identiques pour les 3 jeux de données.

3.4.3.1.2 FONCTIONS MULTIPLES

Nous avons également assigné des fonctions multiples aux LTR. C'est-à-dire que pour chaque LTR dans un jeu de données, nous lui attribuons un état selon les fonctions rencontrées dans chaque échantillon. Par exemple, si une LTR possède 44 fois l'état « No » et 1 fois l'état « Pr », nous lui attribuerons l'état « No/Pr ». De cette manière, cela permet d'avoir une idée des fonctions que peut prendre une LTR. Les figures 2C (article section 3.2.2.4) et S2 (article section 3.2.4.2) résument ces états pour les LTR prototypes. Les figures S2 de ces mêmes papiers résument ces états pour l'ensemble des LTR complètes de la puce. Les résultats sont très similaires entre les jeux de données, avec une grande majorité de LTR silencieuse dans 100% des échantillons, environ 10% de LTR sont No/Pr, et entre 11 et 13% des LTR sont No/pA. De manière intéressante, dans chaque jeu de données, aucune des LTR n'est Pr/pA.

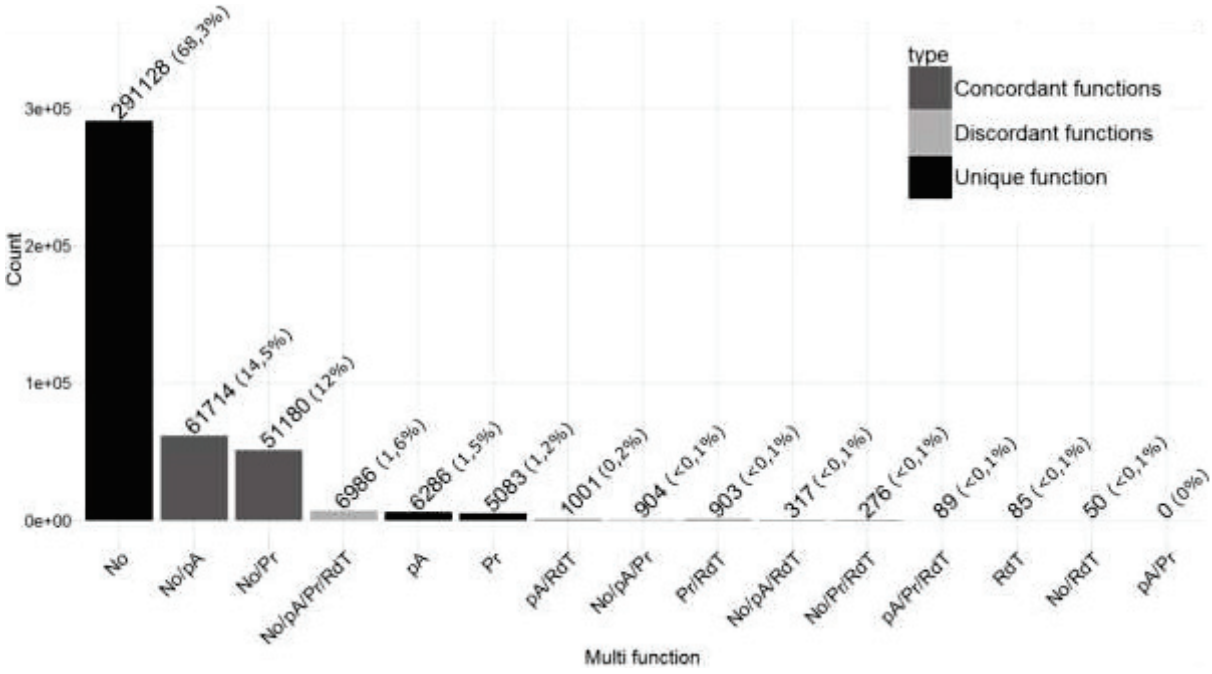


Figure 3-24 : Fonctions multiples dans ET et IS groupés. Un état a été assigné à chaque LTR complète (pour laquelle on peut attribuer une fonction). Un état se définit par chaque fonction qu'une LTR a rencontré dans les échantillons des jeux de données ET et IS regroupés. L'axe des x représente les états multiples possibles des LTR, l'axe des x représente les comptes de chaque état. Les barres sont colorées en fonction du type de fonction multiple de la LTR. On considère qu'une LTR qui regroupe les fonctions Pr et pA (entre autres) a un état discordant.

Cela suggère que la fonction d'une LTR est prédéterminée, c'est ce qu'on appelle le déterminisme opérationnel. Pour confirmer ce déterminisme, j'ai regroupé les 2 jeux de données et réalisé la même figure sur l'ensemble des LTR (Figure 3-24). On peut observer que les résultats sont sensiblement les mêmes quand on regroupe les 2 jeux de données ou quand ils sont pris séparément. Cela montre qu'une LTR ne peut pas prendre n'importe quel état. Aucune LTR ne possède d'état pA/Pr, ce qui confirme le déterminisme opérationnel dans le compartiment sanguin. Seul 2% des LTR possèdent des fonctions discordantes, principalement représentées par des LTR pour lesquelles on retrouve tous les états possibles (No/pA/Pr/RdT), suggérant une expression. On retrouve des pourcentages très intéressants de LTR ayant des fonctions soit No/Pr (12%), soit No/pA (14,5%). Ces LTR sont tout particulièrement intéressantes car ces changements d'états entre silencieuses et actives pour se passer selon les conditions de stimulation, ou conditions pathologiques des patients, mettant en évidence une modulation fonctionnelle des LTR lié à l'état immunitaire. Pour tester cela, nous avons repris le jeu de données ET seul, et récupéré l'ensemble des LTR No/Pr et No/pA. Nous avons ensuite testé l'association entre la fonction des LTR et la

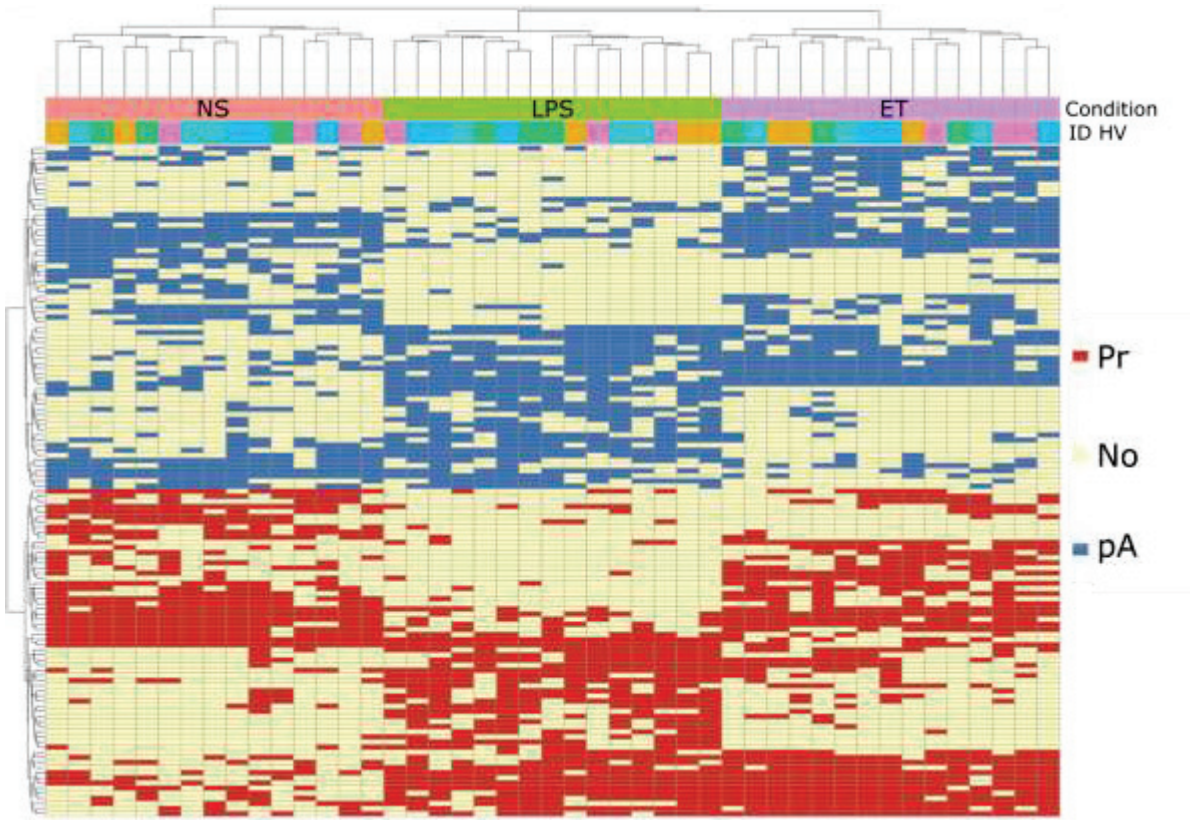


Figure 3-25 : Heatmap représentant les fonctions de LTR dans chaque échantillon du jeu de données ET. Chaque ligne représente une LTR, chaque colonne un échantillon. Une LTR peut avoir soit une fonction de polyAdénylation (pA, bleu), soit une fonction promotrice (Pr, rouge), soit être silencieuse (No, jaune). Les LTR sélectionnées sont soit No/Pr, soit No/pA dans le jeu ET, et sont celles ayant une p-value inférieure à 0,005 pour le test exact de Fisher testant l'indépendance des fonctions entre les conditions.

condition de stimulation, par un test exact de Fisher. Après correction pour le multi-testing, nous avons choisi de représenter les 131 LTR avec une p-value inférieure à 0,005 (Figure 3-25, un seuil plus élevé ne permettait pas de distinguer clairement des LTR dont la fonction est modulée entre les conditions). Un nombre similaire de LTR est présent dans chaque type, 64 LTR sont No/Pr, et 67 sont No/pA. Visuellement, on remarque bien que ces LTR ont une fonction différente selon la condition de stimulation NS, LPS ou ET. Plusieurs profils de fonction selon la conditions sont visibles, mais le plus souvent, c'est la condition LPS qui est la plus différente des 2 autres, c'est-à-dire qu'on a le plus souvent des LTR qui sont actives (pA ou Pr) dans LPS et pas dans NS et ET, ou qui sont silencieuses (No) dans LPS et pas dans NS et ET. D'autres profils sont également représentés. La LTR avec la plus petite p-value correspond à une LTR No/pA située sur le chromosome 19, juste à la fin d'un transcrit du gène CD209 (ou DC-SIGN), impliqué dans la reconnaissance de divers pathogènes. Sur la Figure 3-26 sont représentés quatre profils d'expression de LTR tirées de la heatmap, avec

des profils différents. On voit bien dans ces exemples, que seule une sous-région de la LTR est modulée entre les conditions, l'autre restant nettement en dessous du seuil de bruit de fond. On observe aussi, sur le profil 3, une grande variabilité d'expression d'une sous-région montrant une forte variabilité inter-individuelle d'expression de cette LTR. Ce locus à une VAF moyenne dans la population Européenne à 5% environ. Il n'est pas impossible que cette variabilité soit liée à l'absence de cette LTR chez certains des individus. Pour résumer, cette analyse a permis de montrer qu'il existe des LTR dont la fonction est modulée par la condition. Nous cherchons maintenant à savoir si ces fonctions putatives des LTR peuvent être associées à une expression différentielle de gènes conventionnels.

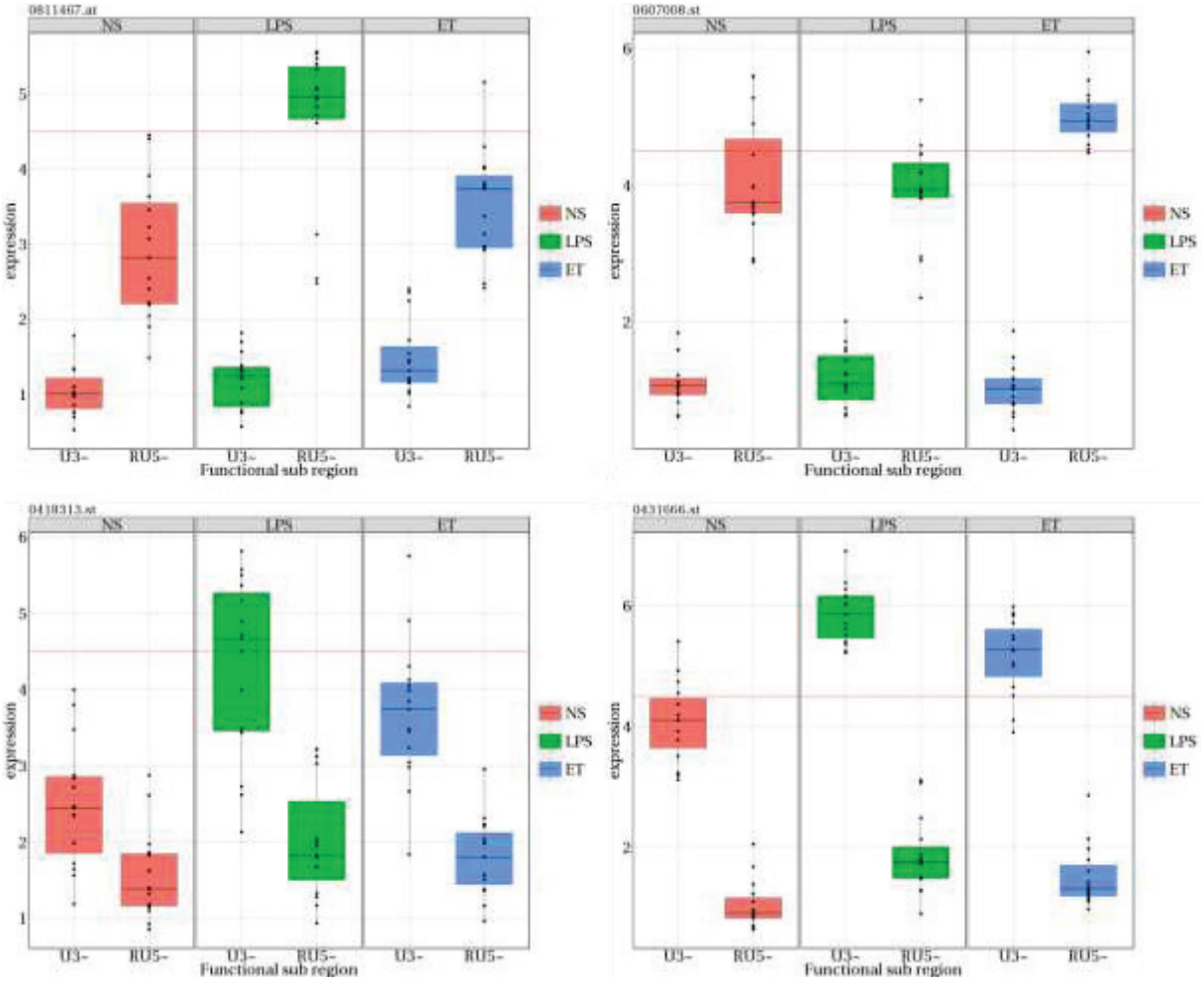


Figure 3-26 : Profils d'expression de LTR dont la fonction varie selon les conditions. De gauche à droite et de haut en bas : LTR No/Pr, avec la sous région U5 plus exprimée que la région U5 et au dessus du seuil de bruit de fons, donc une fonction promotrice dans la condition LPS (vert), LTR No/Pr avec fonctions promotrices dans la condition ET (bleu) principalement mais aussi quelques échantillons dans les conditions NS et LPS, LTR No/pA avec une fonction de polyadénylation dans la condition LPS principalement, et LTR No/pA avec fonction de polyadénylation dans les conditions LPS et ET principalement.

3.4.3.2 LTR REGULATRICES EN CIS DE L'EXPRESSION DES GENES DE L'IMMUNITE

On souhaite tester l'association entre la fonction des LTR et l'expression d'un gène situé proche. Pour cela, j'ai employé 2 modèles différents, un modèle inspiré des eQTL (eQTL-like), et un modèle basé sur la co-expression entre LTR et gène. Les étapes sont résumées sur la Figure 3-27.

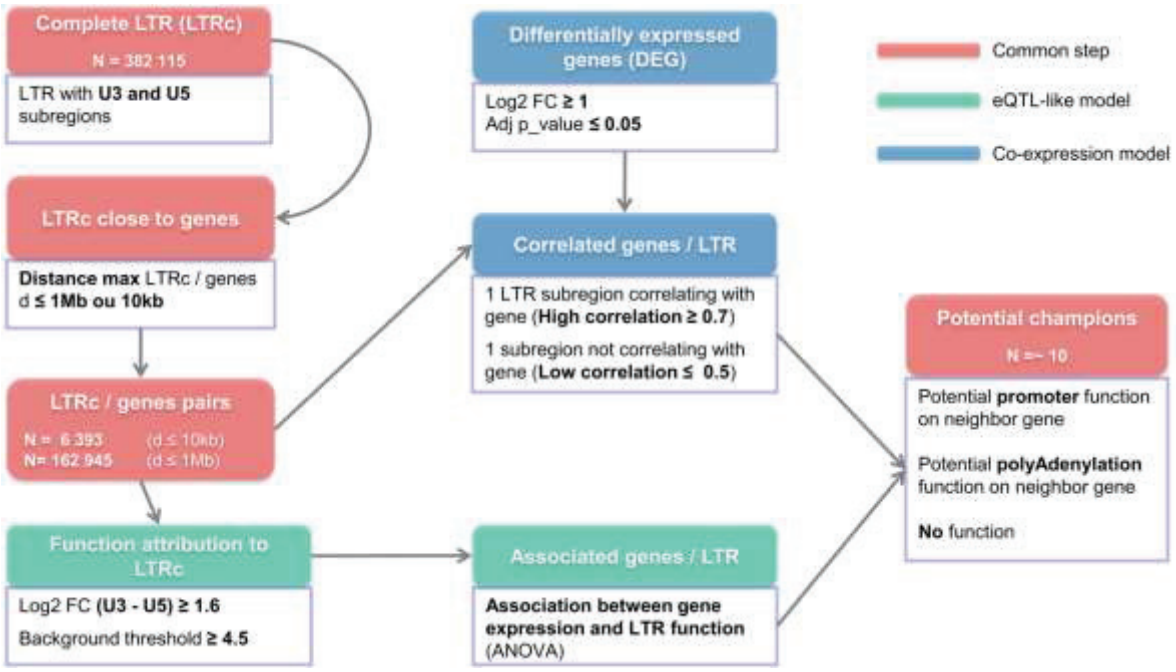


Figure 3-27 : Procédure pour sélectionner des LTR potentiellement régulatrices en cis de l'expression d'un gène voisin. Deux modèles différents sont employés, un modèle inspiré des eQTL associant fonction des LTR et expression des gènes, et un modèle basé sur la co-expression entre LTR et gène. Les première étapes sont communes aux 2 méthodes (rouge). Premièrement, L'ensemble des LTR complètes (LTRc) de la puce HERV-V3 sont sélectionnées (c'est-à-dire les LTR pour lesquelles la puce cible à la fois les sous-régions U3 et U5). Ensuite nous formons des paires LTRc / gène en sélectionnant le gène le plus proche de chaque LTRc, quand il est à une distance inférieure à un seuil de distance (les seuils sont de 10kb ou 1Mb de distance). Ensuite, selon le modèle qu'on va utiliser, les étapes qui suivent divergent. Pour la méthode eQTL-like model (vert), toutes les LTRc ont une fonction attribuée, de la même manière que précédemment (seuil de bruit de fond égal à 4,5 et différence minimale d'expression entre les région U3 et U5 de 1,6 en log2). Un modèle type eQTL est ensuite utilisé (basé sur des tests multiples d'ANOVA testant l'association entre fonction de chaque LTR et expression du gène. Pour le modèle de co-expression (bleu), le principe est de sélectionner les sous-régions de LTR qui corrèlent le mieux avec des gènes différentiellement exprimés. Premièrement, tous les gènes différentiellement exprimés de la puce sont sélectionnés (log2FC supérieur ou égal à 1 et p-value ajustée par Benjamini-Hochberg inférieure à 0,05). Ensuite, on récupère toutes les LTRc qui forment une paire avec les gènes différentiellement exprimés. Puis nous sélectionnons les sous-régions d'une même LTR (U3 ou U5) qui ont à la fois un coefficient de corrélation de Pearson supérieur à 0,7 pour une sous-région et un coefficient inférieur à 0,5 pour l'autre sous-région. Ces deux méthodes permettent d'identifier un certain nombre de loci, potentiellement promoteur de l'expression d'un gène ou potentiellement terminateur de l'expression par un signal de polyAdénylation.

Pour le modèle eQTL-like, nous obtenons donc des paires LTR / gène, pour lesquelles les fonctions des LTR sont associées au niveau d'expression de gènes situés proches. Pour le jeu de données ET, pour une distance maximum entre LTR et gène de 1Mb, nous avons obtenu 42 paires significativement associées (FDR <0,1), dont 8 avec un FDR <0,01. Parmi ces paires, on a identifié celles qui semblaient montrer une fonction de LTR préférentielle par condition, et un différentiel d'expression du gène. On a identifié 2 profil intéressants. Une paire composée d'une LTR45 avec le gène CD209 sur le chromosome 19. La LTR est située à cheval sur la fin d'un transcrit existant pour CD209. Cette LTR a déjà été identifiée précédemment, comme étant No/pA et modulées selon les conditions de stimulation. De plus, sa position par rapport à un transcrit du gène nous laisse fortement penser que cette LTR joue un rôle de polyAdénylation sur le transcrit CD209-203. Son profil d'expression en fonction du statut de la LTR est représenté sur la Figure 3-28 en bas. On voit que la LTR a exclusivement un rôle pA dans la condition NS, associé à une expression élevée du gène, puis la LTR devient silencieuse dans la moitié des échantillons dans la condition LPS, et la LTR est exclusivement silencieuse dans la condition ET, associée à une expression globalement plus faible du gène CD209. Cette paire serait intéressante à valider en paillasse. L'autre profil représenté sur la Figure 3-28 montre celui d'une paire avec fonction No/Pr et du gène DMBT1, située sur le chromosome 10. Cependant la distance entre les 2 loci est importante, à plus de 500kb, et suggère que cette association n'est pas directe.

Mise à part cette paire identifiée, il semble difficile d'en identifier d'autres avec à la fois un profil et une distance intéressants. Pour le jeu de données MIP, nous avons obtenu respectivement 153 et 394 paires significatives avec FDR <0,01 et 0,1. Pour IS, 377 et 760 paires significatives avec FDR<0,01 et 0,1, tous les paramètres étant identiques par ailleurs. Cet écart de nombre entre ET et IS, MIP est étonnant à première vue. Pour MIP, qui comporte plus de 300 échantillons à partir de 102 patients, on comprend aisément le gain de puissance pour ces modèles statistiques. ET et IS contiennent quasiment le même nombre d'échantillons (45 et 44 respectivement). Cependant, les échantillons de ET proviennent de 5 volontaires sains différents seulement, et pourrait expliquer le faible nombre de paires LTRc/gènes significativement associées. Il sera important par la suite d'affiner les listes trouvées séparément dans IS et MIP, afin d'essayer d'isoler de nouvelles LTR jouant potentiellement un rôle sur un gène voisin.

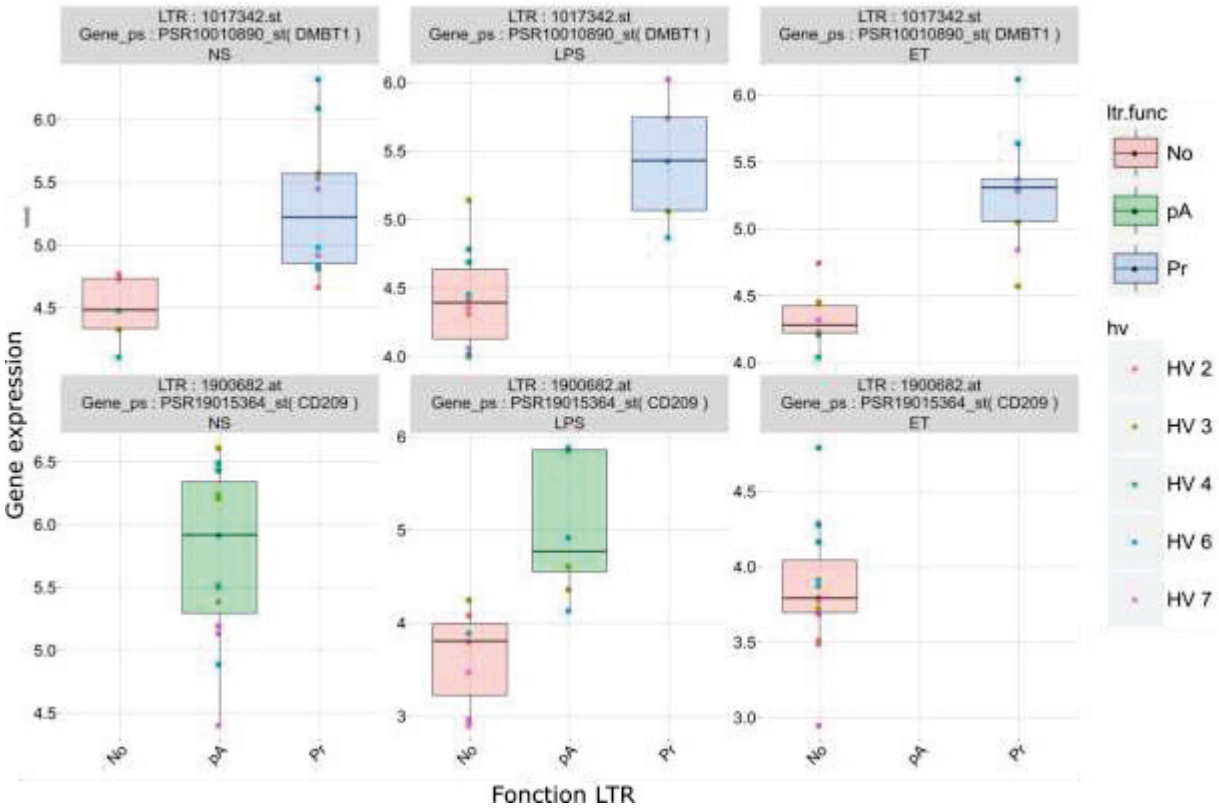


Figure 3-28: Profils de 2 paires LTR/Gene associées par l'approche eQTL-like. Représentation des profils d'expression de 2 paires LTR / gène identifiées, de haut en bas, paire sur chromosome 10 entre un MaLR et le gène DMBT1, paire sur chromosome 19 entre un HERV et le gène CD209. Chaque graphique sur une ligne, représente une condition différente (NS, LPS ou ET). Sur chaque graphique, le titre, de haut en bas correspond à l'identifiant de la LTR, au probeset et alias du gène et la condition expérimentale. En abscisses sont représentées les fonctions retrouvées pour la LTR, en ordonnées le niveau d'expression du gène. Les boxplots sont colorés en fonction de la fonction de la LTR. Les points sont colorés en fonction de l'individu de départ.

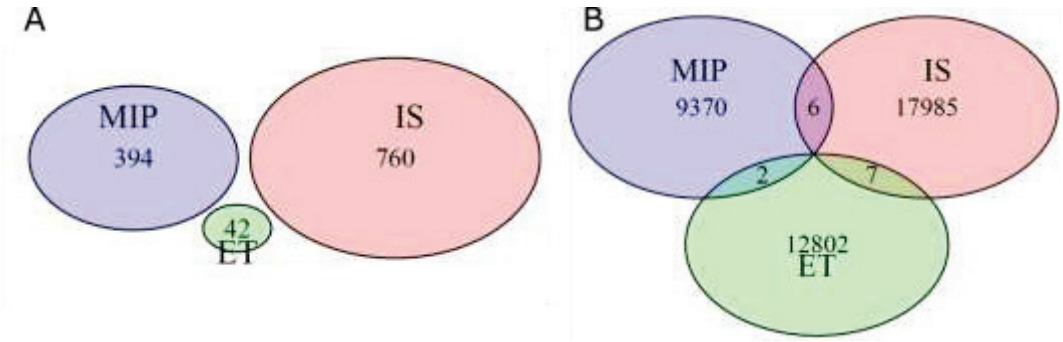


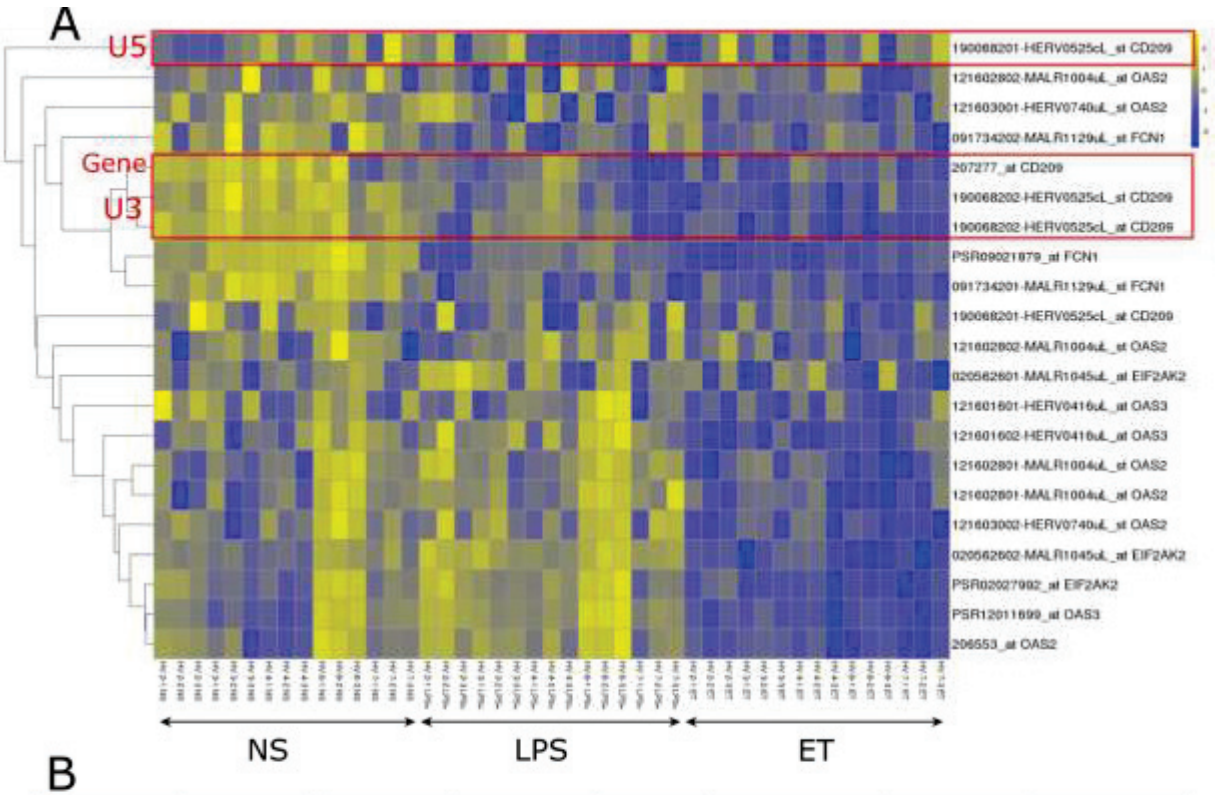
Figure 3-29: : Intersections à partir du modèle eQTL-like entre les 3 jeux de données. A. Intersection des paires LTR/gènes avec un contrôle du FDR à 10%. B. Intersection des paires LTR/gènes sans contrôle du FDR. Les paramètres utilisés pour générer ces résultats sont identiques entre les 3 jeux de données. Distance maximale entre LTR et gène de 1Mb. Les tests utilisés pour modéliser les associations entre fonction de la LTR et gène sont des ANOVA. Avec un contrôle du FDR inférieur à 10%.

Avant cela, nous avons voulu faire des intersections entre les résultats des 3 jeux de données. Sur la Figure 3-29A, on voit qu'il n'existe pas de paire LTR/ gène significativement associée en commun entre les 3 jeux de données. Même quand on ne contrôle pas le taux de fausse découverte, en prenant donc beaucoup plus de paires, on n'obtient que très peu d'intersections, 2 à 2.

Cependant, en faisant des intersections entre les probesets des gènes impliqués dans une association, quelle que soit la LTR, on obtient 230 probesets de gènes à l'intersection des 3 jeux de données (données non montrées). De même en faisant les intersections entre les LTR impliquées dans une association, quelle que soit la LTR, on obtient 9 LTR à l'intersection des 3 jeux de données (données non montrées). Cependant, ces LTR ne sont pas associées au même gène dans les 3 jeux.

Pour le modèle de co-expression, nous obtenons des paires LTR / gène, pour lesquelles le profil d'expression d'une sous-région de la LTR (U3 ou U5) est corrélée au profil d'expression du gène, l'autre sous-région n'étant pas corrélée (Figure 3-27). Sélectionner les LTR avec seulement une sous-région corrélée est important, étant donnée qu'on veut sélectionner des LTR ayant potentiellement un rôle sur l'expression du gène. Les 2 sous-régions corrélées et exprimées marqueraient probablement un embarquement de la LTR dans l'unité de transcription du gène. Cette méthode a été utilisée sur les jeux ET et IS. Pour la figure provenant d'ET, plusieurs paramètres ont été testés, et les résultats pour une distance maximale entre LTR et gène de 10kb et un coefficient de corrélation supérieur à 0,7 sont représentés sur la Figure 3-30. On trouve 5 paires LTR / gènes qui ont passé tous les filtres. La table de la Figure 3-30 B décrit les paires trouvées. Parmi ces paires on retrouve la paire 1900682 / CD209, déjà identifiée avec la méthode eQTL-like. On trouve également 3 paires dans la région des gènes OAS2 et OAS3, gènes impliqués dans la réponse immunitaire à l'infection virale. Pour la figure IS, plusieurs paramètres ont été testés, et les résultats pour une distance maximale entre LTR et gène de 10kb et un coefficient de corrélation supérieur à 0,75 sont représentés sur la Figure 3-31. On trouve 10 paires LTR / gènes qui respectent les critères fixés. La table de la Figure 3-31 B décrit les paires trouvées. Parmi ces paires, on retrouve la paire 1216028 / OAS2, déjà retrouvée pour ET, avec un rôle promoteur potentiel. Cependant, la LTR est placée plutôt à la fin des transcrits d'OAS2.

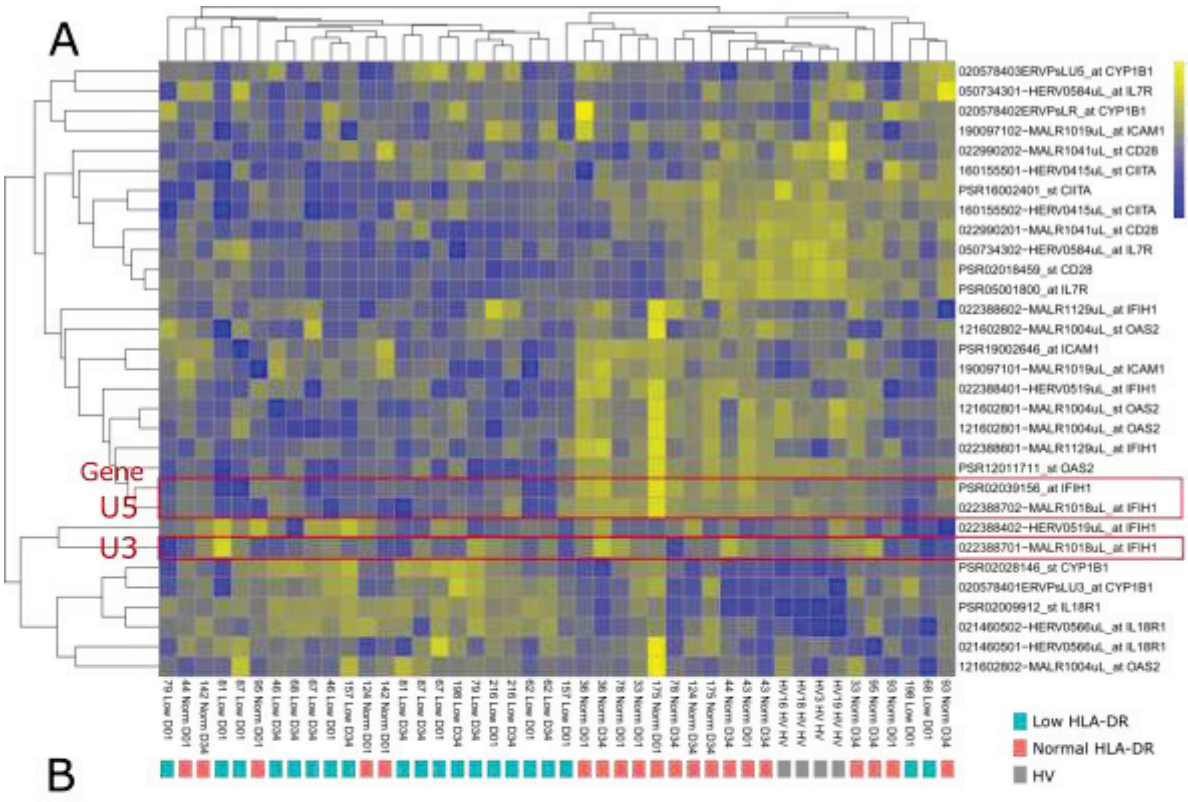
A



B

LTR id	U3 probeset	U5 probeset	Fonction potentielle	Gène alias	Gène probeset	Corrélation haute	Distance (kb)
0205626	**01**	02	**Pr**	**EIF2AK2**	**PSR02027992**	**0.86**	**0**
0917342	01	**02**	**pA**	**FCN1**	**PSR09021879**	**0.78**	**0**
1216016	**01**	02	**Pr**	**OAS3**	**PSR12011699**	**0.79**	**0**
1216028	**02**	01	**Pr**	**OAS2**	**206553_at**	**0.76**	**0**
1216030	**01**	02	**Pr**	**OAS2**	**206553_at**	**0.77**	**0**
1900682	02	**01**	**pA**	**CD209**	**207277_at**	**0.86**	**0**

Figure 3-30: Modèle de co-expression sur ET. A. Co-expression entre LTR et gène. Heatmap représentant les profils d'expression des 10 paires LTR / gènes identifiées par la méthode de co-expression sur la cohorte ET (haute corrélation supérieure à 0,7 et distance LTR – gène inférieure à 10kb). Pour chaque locus, le profil d'expression du gène et des 2 sous-régions U3 et U5 de la LTR sont représentés. Chaque ligne représente un locus sous-région de la LTR ou gène, chaque colonne représente un échantillons. Les loci et les échantillons sont clusterisés selon la méthode complète et la distance de corrélation de Pearson. Sur chaque ligne est noté le nom de probeset du locus (HERV ou gène) ainsi que l'alias du gène associé à la paire. Les cadres rouges montrent l'exemple d'une paire identifiée. On remarque la bonne corrélation entre probeset du gène (207277_at) et probesets de la sous-région 02 (U3), de la LTR 1900682, et la mauvaise corrélation avec les probesets de la sous-région 01 (U5), de la LTR 1900682. **B. Table décrivant les 5 paires identifiées par la méthode.** La colonne *LTR_id* décrit les identifiants des LTR et correspond aux 7 premiers chiffres des noms de probesets notés sur les lignes de la heatmap. Les colonnes *U3 et U5 probesets* décrivent les identifiants des sous-régions de la LTR (U3 ou U5), et correspondent au 8 et 9ème chiffres du nom de probeset complet. La colonne *Fonction Potentielle* décrit la fonction potentielle de la LTR sur l'expression du gène, selon que la sous-région U3 est corrélée avec le gène (fonction pA), ou que la région U5 est corrélée avec le gène (fonction Pr).



LTR id	U3 probeset	U5 probeset	Fonction potentielle	Gène alias	Gène probeset	Corrélation haute	Distance (kb)
0205784	U3	**U5**	**pA**	**CYP1B1**	**PSR02028146**	**0.76**	**1.97**
0214605	02	**01**	**pA**	**IL18R1**	**PSR02009912**	**0.76**	**1.86**
0223884	**02**	01	**Pr**	**IFIH1**	**PSR02039156**	**0.76**	**5.53**
0223886	01	**02**	**pA**	**IFIH1**	**PSR02039156**	**0.75**	**0.62**
0223887	**01**	02	**Pr**	**IFIH1**	**PSR02039156**	**0.82**	**2.29**
0229902	01	**02**	**pA**	**CD28**	**PSR02018459**	**0.79**	**5.64**
0507343	02	**01**	**pA**	**IL7R**	**PSR05001800**	**0.82**	**0.67**
1216028	**02**	01	**Pr**	**OAS2**	**PSR12011711**	**0.76**	**8.31**
1601555	02	**01**	**pA**	**CIITA**	**PSR16002401**	**0.77**	**6.6**
1900971	01	**02**	**pA**	**ICAM1**	**PSR19002646**	**0.76**	**0.020**

Figure 3-31: Modèle de co-expression sur IS. A. Co-expression entre LTR et gène. Heatmap représentant les profils d'expression des 10 paires LTR / gènes identifiées par la méthode de co-expression sur la cohorte IS (haute corrélation supérieure à 0,75 et distance LTR – gène inférieure à 10kb). Pour chaque locus, le profil d'expression du gène et des 2 sous-régions U3 et U5 de la LTR sont représentés. Chaque ligne représente un locus sous-région de la LTR ou gène, chaque colonne représente un échantillons. Les loci et les échantillons sont clusterisés selon la méthode complète et la distance de corrélation de Pearson. Sur chaque ligne est noté le nom de probeset du locus (HERV ou gène) ainsi que l'alias du gène associé à la paire. Les cadres rouges montrent l'exemple d'une paire identifiée. On remarque la bonne corrélation entre probeset du gène (PSR02039156_at) et probesets de la sous-région 02 (U5), de la LTR 0223887, et la mauvaise corrélation avec les probesets de la sous-région 01 (U3), de la LTR 0223887. **B. Table décrivant les 10 paires identifiées par la méthode.** La colonne *LTR_id* décrit les identifiants des LTR et correspond aux 7 premiers chiffres des noms de probesets notés sur les lignes de la heatmap. Les colonnes *U3 et U5 probesets* décrivent les identifiants des sous-régions de la LTR (U3 ou U5), et correspondent au 8 et 9ème chiffres du nom de

probeset complet. La colonne *Fonction Potentielle* décrit la fonction potentielle de la LTR sur l'expression du gène, selon que la sous-région U3 est corrélée avec le gène (fonction pA), ou que la région U5 est corrélée avec le gène (fonction Pr).

Pour conclure, à partir de différentes approches, nous avons identifié dans cette partie, une dizaine de LTR potentiellement régulatrices de l'expression de gènes impliqués dans la réponse immunitaire. Cela met en évidence des liens possibles entre HERV et réponse immunitaire. Il sera nécessaire par la suite de développer des stratégies expérimentales humides permettant de valider ou d'invalider les rôles potentielles de ces LTR sur ces gènes.

4 DISCUSSION

Dans ce projet, nous cherchons à décrire et à mieux comprendre la contribution des HERV au sein de la réponse immunitaire de l'hôte en conditions d'agression inflammatoire. Pour apporter des éléments de réponse, nous avons développé des méthodes et outils spécifiquement dédiés à la description du HERVome, que ce soit au niveau génomique ou transcriptomique. Ces outils sont également la base pour émettre des hypothèses fonctionnelles sur le rôle des HERV dans la réponse immunitaire de l'hôte. On décrit pour la première fois l'expression de l'ensemble du transcriptome HERV dans le sang, suite à une agression inflammatoire. On observe également une modulation commune de loci spécifiques entre des brulés, des traumatisés et des chocs septiques, comparées à des volontaires sains, avec la puce commerciale U133plus2. A partir de la puce HERV-V3, spécifiquement dédiée à l'étude de l'expression des HERV, on montre une expression et une modulation dans le sang de ces éléments, selon différents marqueurs du statut immunitaire des patients après un choc septique. Et de manière intéressante, un certain nombre de ces loci se trouvent proches de gènes impliqués dans la réponse immunitaire. Nous cherchons alors à identifier s'il existe des loci pouvant jouer un rôle sur l'expression de gènes situés proches dans le génome. Par une approche rendue possible grâce au niveau fin d'annotation de notre base de données *hervgdb4*, nous avons assigné un état et un rôle potentiel en cis sur l'expression de gène de chaque LTR ciblée par la puce. Puis nous avons cherché à associer l'expression de LTR à celle de gènes situés à proximité, par deux approches différentes. Bien qu'il ne semble pas que ce soit un effet global et massif, nous identifions un certain nombre de LTR candidates pouvant jouer, soit un rôle promoteur, soit un rôle terminateur de transcrits de gènes. Dans ce projet, nous cherchons également à comprendre la plus grande variabilité d'expression observée chez les HERV par rapport aux gènes et nous évaluons leur fréquence d'absence dans le génome. Nous avons développé une méthode évaluant spécifiquement le niveau de polymorphisme de présence des HERV annotés sur plus de 2000 individus regroupés en populations (Projet 1000 génomes). On montre ici une fréquence d'absence importante et suivant des motifs liés à la population d'appartenance des individus étudiés. Le groupe HML-2 est plus polymorphique entre individus que les autres groupes. De manière intéressante, nous mettons en évidence un certain nombre de loci situés proches des gènes de la région HLA et dont la présence ou l'absence est associée à

l'expression de gènes à proximité. Cette région est connue pour son fort polymorphisme et son rôle important dans la reconnaissance des pathogènes, suggérant des liens entre immunité et HERV. Au final, ce projet nous permet de poser des hypothèses sur le rôle de certains HERV dans la réponse immunitaire de l'hôte.

4.1 DEVELOPPEMENTS METHODOLOGIQUES

Les HERV représentent une grande partie de notre génome et pourtant c'est un répertoire encore relativement peu exploré. Pour cela, que ce soit au niveau génomique ou transcriptomique, nous avons mis en place un certain nombre d'outils.

La base de données *hervgdb4* au service des analyses du transcriptome

La base de données *hervgdb4*, comporte plus d'entrées que les bases d'annotation publiques, générées à partir de *RepeatMasker* (Smit et al., 2013). Au-delà de cette différence numérique, l'intérêt de *hervgdb4* par rapport à ce qui existe est qu'elle contient un niveau d'annotation plus fin, et ce, à deux niveaux : au niveau locus, dans *hervgdb4*, deux entrées situées proches dans le génome et provenant du même groupe sont regroupées entre elles et sont considérées comme étant un seul et même locus – au niveau sous-région fonctionnelle, pour une LTR annotée dans les tracks publiques, *hervgdb4* contient deux ou trois entrées (U3/U5 ou U3/R/U5), selon le niveau d'annotation (répertoires Dfam ou Prototype). Cette annotation plus fine dans *hervgdb4* nous a permis de développer des stratégies pour mettre en évidence d'éventuels rôles fonctionnels des LTR dans des situations d'agression inflammatoire dans le sang. Cette différence au niveau de l'annotation entre les deux bases de données implique également des entrées de taille plus petite dans *hervgdb4*. On peut penser que cette différence de taille de chaque entrée dans les deux bases puisse influencer les résultats des analyses. C'est probablement le cas pour l'étude des variations en nombre de copies des HERV dans le génome humain où on a observé une fréquence d'absence plus élevée des entrées de la base *hervgdb4* que celles de *RepeatMasker* (article section 3.3.2, figure 3A).

De plus, *hervgdb4* a servi de base pour le développement des sondes de la puce HERV-V3. Cette puce, qui cible spécifiquement les HERV et MaLR du génome, nécessitait une validation et le développement d'un pipeline d'analyse robuste pour la suite. La technologie des puces à ADN a souvent été critiquée, notamment à cause de son incapacité à identifier des informations directement transférables en clinique (Webb et al., 2007), malgré des études montrant de bonnes performances en terme de reproductibilité (Shi et al., 2006) et de spécificité (Canales et al., 2006). Il était donc important d'avoir des performances au moins similaires sur la puce HERV-V3 sur ces critères. Il était également nécessaire de développer un pipeline de pré-processing et d'analyse adapté à cette technologie. Il reste toutefois très similaire à celui des puces commerciales (http://arrayanalysis.org/main.html). Pour le RNAseq, nous avons choisi en première approche d'utiliser des pipelines classiques existants, avec des outils et des critères assez stricts afin de s'affranchir le plus possibles des problèmes de multi-mapping sur les éléments répétés. Grâce à la longueur correcte des reads et au fait qu'ils soient appariés, le pourcentage de reads mappant correctement et de manière unique est élevé (98.4%). Cependant ce bon pourcentage ne nous garantit pas, sur des éléments répétés, de l'exactitude du mapping de tous les reads. En effet, sur des éléments répétés pouvant varier d'une seule paire de base, on peut imaginer qu'une erreur dans le séquençage ou une mutation ponctuelle puisse entraîner le mapping du read sur le mauvais élément (Treangen and Salzberg, 2011). Pour s'assurer de résultats robustes de RNAseq et permettre des comparaisons avec la puce HERV-V3, il sera probablement intéressant d'enrichir le pipeline RNAseq en utilisant les séquences des sondes de la puce HERV-V3 qui ciblent de manière unique les HERV. En effet, en ne mappant que les reads qui contiennent une séquence de probe provenant de la puce HERV-V3 on s'assurerait de l'unicité du mapping (à une erreur de séquençage près). Ensuite on pourrait reconstruire les transcrits HERV dans leur totalité via un assemblage *de novo* de proche en proche à partir des reads mappés sur les séquences de probes.

Par rapport aux puces, le RNAseq pourrait également avoir un avantage à l'avenir. Dans ce projet, nous avons assigné des fonctions potentielles aux LTR, notamment promotrices ou terminatrices, en utilisant les signaux des probesets ciblant spécifiquement les sous-régions fonctionnelles des LTR. Ensuite, pour valider les états promoteurs putatifs des LTR sur un gène proche, il a été nécessaire d'avoir une approche *in silico* pour les apparier avec des

gènes suivant le même profil d'expression. Mais de cette manière on ne peut pas savoir si c'est la paire LTR-gène qui forme bien un seul et même transcrit et non pas deux transcrits co-régulés mais distincts. On pourrait utiliser le RNAseq pour construire un pipeline permettant d'identifier des transcrits hybrides LTR-gènes, qui commenceraient, ou se termineraient par la LTR, suggérant que la LTR est soit promotrice, soit terminatrice du transcrit. Une telle méthode a déjà été utilisée avec succès dans les ovocytes et a permis d'identifier un certain nombre de LTR promotrices dans ce contexte (Babaian et al., 2017; Brind'Amour et al., 2018).

Le choix des seuils influence les interprétations biologiques

Plus globalement, tout au long de ce projet et en bioinformatique en général, les méthodes développées ont nécessité la définition de seuils. Que ce soit pour définir la présence ou l'absence de HERV à partir du niveau de couverture des reads de séquençage ADN, ou bien pour définir le seuil de bruit de fond des puces HERV-V3, ou encore pour définir le niveau de significativité des différents tests statistiques utilisés, nous avons dû faire des choix. Ces choix ont un impact sur le niveau de polymorphisme, le nombre de HERV exprimés ou modulés dans les différents jeux de données et donc sur les interprétations biologiques. Quand c'était possible, bien que discutable, nous avons utilisé des valeurs couramment utilisées et conseillées, comme pour les seuils de fold change et de significativité des études d'expression différentielle (McCarthy and Smyth, 2009; Shi et al., 2006). Mais dans d'autres cas, l'approche est nouvelle et il a fallu faire des choix. Par exemple, pour définir une fonction promotrice (ou terminatrice) à une LTR, nous avons défini qu'il devait y avoir une intensité minimum sur la sous-région U5 (ou U3) et qu'elle devait avoir au moins 3 logs d'intensité en plus que la région U3 (ou U5). Après plusieurs changements, ces seuils ont été définis de manière empirique, où nous avons assigné des seuils qui nous paraissait « stricts », au regard des résultats obtenus. Faut-il alors toujours privilégier des seuils stricts, favorisant la spécificité au détriment de la sensibilité ? Ou à l'inverse privilégier, dans un contexte de découverte, la sensibilité, au risque de faire des fausses découvertes ? Il n'y a probablement pas de bonne réponse, chaque cas doit être traité indépendamment, mais l'interprétation doit toujours être réalisée en gardant en tête les choix réalisés tout au long du pipeline. Une solution pourrait être de ne pas faire de choix, et classifier les éléments selon un critère quantitatif et continu. Cependant, ce genre

d'approches peut rendre difficile les interprétations. Pour le polymorphisme par exemple, nous avons choisi deux approches différentes pour associer le niveau de polymorphisme des HERV avec leur expression. D'une part, nous avons défini une approche sans assigner de seuil de présence ou absence du HERV dans chacun des génomes, en corrélant la valeur du percentile de la distribution de couverture des reads de séquençage avec le niveau d'expression des HERV. D'autre part, une approche où l'on a défini un génotype (défini par un seuil) pour chaque individu et chaque HERV (homozygote absent, hétérozygote et homozygote présent), que l'on a associé au niveau d'expression du HERV. Nous avons trouvé relativement peu de HERV dont l'expression est associée à leur présence, mettant en évidence qu'il n'existe pas de manière globale, de lien significatif entre le degré de polymorphisme et l'expression des HERV. Autrement dit, ce n'est pas parce qu'un élément est hautement polymorphique que son expression est plus variable. De manière intéressante, la majorité des HERV identifiés étaient trouvés par chacune des deux méthodes, mais les deux listes ne sont pas identiques. On pourrait ainsi se demander quel est l'impact des méthodes et seuils utilisés sur les interprétations biologiques. Pour y répondre, il faudrait soit se baser et comparer sur ce qui existe déjà dans la littérature, soit mettre en place des études comparatives entre plusieurs seuils et plusieurs méthodes pour analyser leur influence. En pratique il est difficile de réaliser ce genre d'études. Compte-tenu du nombre important de nouvelles approches employées et du manque de connaissance sur les HERV dans les contextes étudiés dans ce projet, nous avons choisi, quand cela était possible, des seuils élevés, minimisant le nombre de fausses découvertes.

Le développement de ces approches nous a permis de décrire le polymorphisme des HERV dans le génome humain ainsi que le transcriptome HERV dans le sang, sur des modèles de choc septique et des patients suivant une agression inflammatoire.

4.2 TRANSCRIPTOME HERV ET ROLE SUR LA REPONSE IMMUNITAIRE DE L'HOTE A L'AGRESSION INFLAMMATOIRE

Les HERV sont exprimés et modulés dans le sang en condition inflammatoire, indépendamment de leur groupe d'appartenance

A travers les différents jeux de données étudiés et les différents outils utilisés, nous avons montré qu'entre 5 et 9% des HERV annotés sont transcrits dans le sang (représente environ 85 000 probesets et 40 000 loci distincts). L'étude réalisée à partir des données de la puce commerciale U133 a montré une proportion importante de HERV exprimés par rapport aux nombre ciblé (plus de 20%). Cependant, le nombre de HERV ciblés est très faible et ces HERV se trouvent pour la plupart proches de gènes, dans des régions transcrites, augmentant ainsi le pourcentage de loci actifs par rapport au total ciblés. Ainsi globalement on trouve que la grande majorité des HERV ne sont pas exprimés, ces résultats sont en accord avec les études précédentes montrant que les HERV sont globalement peu exprimés, dans des tissus cancéreux (Pérot et al., 2012), la peau lors de psoriasis (Lättekivi et al., 2018), ou encore le cerveau lors de sclérose latérale amyotrophique (Prudencio et al., 2017). Chacune de ces études s'intéresse à la comparaison entre échantillons sains et pathologiques et met en évidence que globalement, les pourcentages de HERV exprimés dans les échantillons sains et pathologiques ne sont pas différents. Dans le jeu de données de puce HERV-V3 comportant à la fois des échantillons sains et de patients en choc septique (IS), la majorité des HERV sont exprimés dans les deux conditions. Ainsi, la majorité des HERV exprimés ne sont pas modulés entre états sains et pathologiques (plus de 90% des HERV exprimés ne sont pas modulés entre volontaires sains et choc septiques dans IS). L'approche par famille (ou groupe) réalisée dans les jeux de données HERV-V3 et RNAseq, n'a pas montré d'enrichissement ou de déplétion de l'expression d'une famille par rapport aux autres. Chacun des 42 groupes annotés dans *Hervgdb4* comporte des éléments exprimés dans le sang, quel que soit le jeu de données utilisé. On remarque un nombre important de loci des groupes HERV-H et HML-2 (HERV-K), dont des éléments ont déjà été retrouvés exprimés dans les PBMC (Balestrieri et al., 2015) et dans de nombreux autres tissus et contextes (Brudek et al., 2009; Johnston et al., 2001; Pérot et al., 2015; Wallace et al., 2014; Young et al., 2018).

Bien que la majorité des HERV ne soit pas différentiellement exprimée entre les conditions, un nombre non négligeable est modulé. Ce nombre varie en fonction de la question posée et de la puissance statistique. De manière étonnante, l'intersection des éléments différentiellement exprimés entre les jeux de données ET, IS et MIP est faible. On peut expliquer cela par plusieurs raisons : les conditions entre les trois jeux de données ne sont pas tout à fait les mêmes, dans ET nous sommes dans un modèle in vitro de PBMC sains stimulés ou non par du LPS, dans IS et MIP nous sommes sur du sang total directement prélevé de patients. Le nombre d'éléments HERV modulés en commun dans IS et dans MIP n'est pourtant pas plus important. Pour rappel, les chocs septiques dans IS ont été sélectionnés par rapport à leur statut immunitaire (expression du mHLA-DR à la surface des monocytes), et forment deux groupes distincts et de même taille. Nous avons cherché à identifier des HERV modulés entre ces deux groupes de patients. Les chocs septiques dans MIP n'ont pas été sélectionnés sur ce critère, et la classification des patients en deux groupes selon le ratio de CD74 entre le jour 3 et le jour 1 après l'entrée en réanimation donne une information différente, étant un proxy de la survenue d'HAI (Peronnet et al., 2017). Cela explique très probablement la faible intersection entre les jeux de données. Une autre explication est, comme dans la plupart des jeux de données de patient, la très forte hétérogénéité entre les patients, spécialement sur l'expression des HERV (par rapport à celle des gènes) qui est importante, réduisant la capacité des tests à détecter des différences quand il y en a vraiment une. Ces résultats nous ont également poussé à procéder autrement et à utiliser la cohorte MIP comme jeu de validation des HERV trouvés modulés dans IS. De cette manière, nous avons montré que les HERV, modulés selon le statut immunitaire de patients en choc (HLA-DR haut ou bas), permettent de discriminer parmi des patients graves en choc septique, ceux les plus graves avec un statut immunitaire plus détérioré que les autres, dans une cohorte où les patients ne sont pas sélectionnés selon leur statut immunitaire.

Les HERV modulés dans ces contextes inflammatoires sont proches de gènes de l'immunité

Nous avons observé plusieurs régions spécifiques du génome, transcriptionnellement actives, qui contiennent également un grand nombre de HERV exprimés. Par exemple, toute la région des récepteurs à l'IL1 est modulée selon le statut immunitaire des patients dans IS

et dans MIP (Figure 23 rapport MIP Annexe 3 : Rapport d'analyses de la cohorte MIP). Dans cette région, on retrouve modulés, non seulement les gènes de la famille de récepteurs (IL18R1, IL1R1, IL1R2, IL18RAP) mais aussi une dizaine de HERV situés dans cette région. Cette observation nous laisse penser que beaucoup de HERV retrouvés modulés, sont embarqués dans les unités transcriptionnelles des gènes (ou groupes de gènes). Il n'est tout de même pas à exclure que dans certains cas, certains HERV ont un rôle sur l'expression du gène situé proche dans le génome. D'ailleurs, si on s'intéresse plus précisément à cette région, on se rend compte que tous les HERV retrouvés modulés sont situés aux extrémités de la région. Il pourrait donc y avoir une influence de la position génomique de ces éléments sur leur expression. Cette répartition, qui semble non aléatoire, nous laisser penser à un possible impact de ces HERV sur l'activation ou l'inactivation de l'expression des gènes de cette région. Avec l'étude réalisée sur la puce U133, au-delà de nous montrer une preuve de concept de l'existence d'expression et de modulation HERV dans des contextes d'agression inflammatoire grave, nous avons identifié 5 HERV co-modulés entre volontaires sains et chocs septique, traumas et brulés. Ces éléments se situent à proximité, au début ou à la fin de gènes, tous impliqués dans la réponse de l'hôte. Les HERV localisés en 3' de NFE4 et CD300LF, possèdent par exemple des séquences pouvant marquer des signaux de polyAdénylation. Ces observations nous ont permis de poser les hypothèses quant à un rôle des LTR sur l'expression de gènes en cis dans des contextes d'agression inflammatoire.

Des LTR peuvent agir en cis sur l'expression de gènes de l'immunité

La puce HERV-V3 permet d'attribuer des états putatifs aux LTR, dont celui de promoteur et de polyAdénylation, via la présence de probesets ciblant spécifiquement les sous régions fonctionnelles U3 et U5 des LTR. Sur les 3 jeux de données de la puce HERV-V3, nous avons étudié les LTR pouvant avoir des fonctions en cis sur des gènes de l'immunité. Les approches de corrélation d'expression entre LTR et gène ou d'association entre état de la LTR et expression du gène, nous a permis d'identifier une dizaine de LTR pouvant avoir un rôle soit promoteur, soit terminateur de l'expression d'un gène de la réponse de l'hôte. Parmi ceux-ci, on retrouve une majorité de gènes de la réponse à l'infection virale, OAS2, OAS3, IFIH1, EIF2AK2, CD209 (Donovan et al., 2015; Kang et al., 2009; Lin et al., 2003; Sarkar et al., 1999; Züst et al., 2011). Plusieurs exemples dans la littératures décrivent une modulation de HERV à la suite d'infection virale par HIV (Vincendeau et al., 2015), CMV (Assinger et al., 2013) ou

encore EBV (Mameli et al., 2012). Ces différentes observations, additionnées à l'association de certaines LTR sur l'expression de gènes de la réponse virale, mettent en évidence des liens étroits entre réponse de l'hôte à un virus et expression de HERV. Dans les cas évoqués, il est difficile de déterminer si le HERV joue effectivement un rôle dans la transcription du gène à la réponse virale, ou bien si le HERV, situé par hasard proche d'un gène de la réponse virale, se retrouve embarqué dans l'unité de transcription du gène. Des cas de rôle promoteur de LTR sont avérés dans la littérature, notamment dans le contexte du cancer (Babaian and Mager, 2016; Babaian et al., 2016; Lamprecht et al., 2010; Wiesner et al., 2015). Plusieurs parallèles peuvent être fait entre la réponse immunitaire à un pathogène et la réponse immunitaire à la tumeur (Finn, 2012; Goldszmid et al., 2014). Ainsi, bien que nous n'ayons pas pu aller valider in silico la fonction des LTR, ces dernières font de bonnes candidates pour une validation en laboratoire humide. Parmi les candidats, nous avons sélectionné la LTR du groupe LTR16A1, qui a une fonction putative promotrice dans le jeu de données ET sous la condition LPS et est localisée juste en amont du transcrit OAS3-204. Nous avons ainsi initié un projet consistant à dépléter la LTR sur des PBMC stimulés au LPS via la méthode CRISPR-cas9 (Hsu et al., 2014) et regardé l'impact sur l'expression du gène. Pour le moment, ce projet n'est pas encore pu être mené à bien. Le passage à la validation en paillasse reste une perspective importante pour la suite.

Plus globalement, à travers les analyses du transcriptome HERV, il est apparu que les HERV ont globalement une expression plus basse que les gènes, et avec une plus grande variabilité interindividuelle. Cette expression plus basse et plus variable des HERV par rapport aux gène peut signifier une fuite de transcription dans les régions de gènes modulés et ayant un effet sur la réponse de l'hôte, ou peut signifier une contribution des HERV indirecte dans la réponse de l'hôte via la fixation de facteurs de transcription. Cette dernière hypothèse reste difficile à vérifier sans l'apport de données épigénétiques. Par exemple, des études récentes mettent en évidence l'apport de sites de fixation du facteur de transcription STAT1 par la famille des HERV MER41, modulant l'expression de gènes inductibles par l'IFN γ (Chuong et al., 2016; Schmid and Bucher, 2010). De notre côté, nous avons mis en évidence qu'un grand nombre de HERV sont co-modulés et probablement co-régulés avec des gènes de l'immunité appartenant à un même réseau fonctionnel, après stimulation au LPS. Nous

avons également révélé un léger enrichissement en sites de fixation du facteur de transcription AP-1 (régule l'expression de gènes de la réponse de l'hôte) dans les LTR potentiellement promotrices par rapport aux LTR silencieuses, dans le modèle ET. Il sera intéressant de confirmer ces résultats sur un plus grand nombre et plus précisément dans une étude dédiée. Toutes ces observations démontrent en tous cas l'intérêt d'explorer l'apport des HERV en sites de fixation de facteur de transcription dans les réseaux de gènes impliqués dans l'immunité.

Au-delà des potentiels rôles promoteurs ou terminateurs de transcrits, voire de l'apport en sites de fixation de facteurs de transcription des HERV, leur expression pourrait influencer d'autres manières la réponse transcriptionnelle de l'hôte à l'agression inflammatoire. Nous n'avons par exemple pas cherché à identifier des LTR pouvant initier des transcrits anti-sens de gènes et ainsi empêcher la transcription normale du gène de l'hôte (Gosenca et al., 2012; Kim and Hahn, 2010). C'est une approche qui pourrait se mettre en place dans un premier temps *in silico,* en identifiant des LTR intra-géniques dont l'expression est anti-corrélée à celle du gène proche duquel (ou dans lequel) elle est située et représente une perspective intéressante dans l'étude de l'impact des LTR sur l'expression des gènes lors de la réponse immunitaire de l'hôte.

Une partie de l'importante variabilité interindividuelle observée à travers les études du transcriptome, HERV ou gène, peut s'expliquer par le choix d'étudier les échantillons de sang ou de PBMC, regroupant plusieurs types cellulaires. Ce choix est dicté par le cadre de recherche du laboratoire, translationnelle à l'hôpital, dont le but est de rester le plus proche possible des conditions *in vivo*. D'un point de vue mécanistique, afin de mieux comprendre quelle type cellulaire exprime tel ARN dans telle condition, il serait pertinent de mettre en place des études du transcriptome dans un seul type cellulaire. Pour les HERV plus spécifiquement, cette variabilité interindividuelle encore plus importante que celle des gènes, a mené à l'hypothèse que ces HERV, annotés dans le génome de référence, ne sont pas toujours présents, chez tous les individus de la population humaine.

4.3 VARIATIONS EN NOMBRE DE COPIES DES HERV DANS LE GENOME ET IMPACTS SUR LA REPONSE DE L'HOTE

HERVdel, une méthode sensible

La méthode originale (HERVdel) que nous avons développée pour étudier la variation en nombre de copies des HERV est la première à s'intéresser à l'ensemble des HERV annotés dans le génome humain de référence, et ce sur 2691 individus (Consortium, 2010). Elle nous a permis de mettre en évidence une fréquence d'absence des HERV relativement importante. De manière intéressante, plusieurs centaines d'éléments sont absents dans la majorité des individus des différentes populations humaines. D'autre part, les niveaux d'absence de nombreux HERV sont influencés par l'appartenance à une super-population, suggérant que ces HERV sont soit apparus, soit disparus récemment dans une population, probablement après les phénomènes migratoires de ces populations. De plus, nous avons vérifié si les HERV trouvés modulés dans les différentes études étaient polymorphiques, et globalement, nous n'avons pas trouvé de différences avec les fréquences sur l'ensemble des HERV (données non montrées). Cependant pour qu'un locus soit statistiquement différentiellement exprimé, sa variabilité d'expression ne doit pas être trop importante. Il est ainsi probable que les HERV que nous avons trouvé modulés dans les analyses de transcriptome, et donc avec peu de variabilité au sein d'une même condition, soient présents dans le génome de la grande majorité des individus étudiés. Par la suite, il serait intéressant de prendre en compte le niveau de polymorphisme pour les études de transcriptome et retirer ceux qui sont le plus souvent absents dans la population étudiée. L'idéal serait d'avoir, pour les mêmes échantillons, à la fois les données du génome et du transcriptome. Ceci permettrait de pondérer l'étude transcriptomique par l'information de présence ou d'absence de chaque HERV dans chaque échantillon et ainsi gagner en puissance statistique. Dans ce travail, nous sommes restés au mêmes niveaux d'annotations que les bases utilisées, c'est-à-dire au niveau région structurelle (LTR ou interne) pour RepBase et au niveau région fonctionnelle pour hervgdb4 (U3, U5, pol,…). La taille de chaque entrée est en moyenne de 400 pb pour RepBase et de 150 pb pour hervgdb4. C'est aussi en cela que notre approche diffère des approches classiques d'études des variations en nombre de copie d'éléments du génome (Sudmant et al., 2015; Zarrei et al., 2015). Notre approche HERV-centrée est adaptée pour

détecter des délétions de quelques centaines de pb, alors que les variant structuraux ont pour la plupart une taille plus importante (taille médiane de 2500 pb pour les délétions dans l'étude de Sudmant et al.). Notre méthode nous permet probablement d'être plus sensible, et de détecter plus de HERV que les autres études (seuls 11 délétions détectées par l'étude de Zarrei et al. correspondent exactement à un locus HERV, et seuls 16% des HERV sont localisés dans une région délétée, (article section 3.3.2, figure 2a et table S3).

Il sera également intéressant de réaliser le même type d'analyse au niveau des loci entiers. Nous avons observé des fréquences d'absence similaires entre régions d'un même locus (par exemple les sous-région U3 et U5 de la LTR solo d'un MaLR situé sur le chromosome 10 à la position 106191460 suivent le même profil de polymorphisme entre populations). Mais mis à part sur quelques exemples, nous n'avons pas regardé à quel point le polymorphisme qu'on observe est lié à l'ensemble d'un locus HERV ou bien à ses régions qui le composent.

Plusieurs études ont montré plusieurs nouvelles insertions du groupe HML-2, plus récent groupe du génome humain et le seul à ce jour identifié comme étant encore capable de se dupliquer (Bannert and Kurth, 2006; Marchi et al., 2014; Wildschutte et al., 2016). Bien que nous ne mettons pas en évidence de nouvelles insertions, nous retrouvons tout de même les LTR de ce groupe (LTR5) comme étant les plus polymorphiques entre les populations. Plusieurs de ces loci se retrouvent dans la région HLA du chromosome 6, connue depuis longtemps pour son polymorphisme très important (Apanius et al., 1997; Weitzman, 2000; Williams, 2001), et son rôle dans la reconnaissance des pathogènes (Doherty and Zinkernagel, 1975).

La présence de loci HERV impacte l'expression de gènes de la réponse de l'hôte

A partir des données publiques des 1000 génomes, nous avons cherché les associations entre présence ou absence des HERV dans les génomes étudiés et expression de ces mêmes HERV sur plus de 400 individus, pour lesquels étaient disponibles à la fois les données génomiques et transcriptomiques (Lappalainen et al., 2013). De manière étonnante, nous avons mis en évidence un faible nombre de loci dont l'absence dans le génome était associée à une absence d'expression. Cela met en lumière le fait que les HERV soient majoritairement non exprimés, spécialement dans les tissus sains. Sur la dizaine de HERV identifiés, trois se trouvent dans la région HLA, dont une LTR du groupe HML-2. Cette LTR est

directement située en 3' du gène HLA-DQB1, et juste en amont du transcrit anti-sens HLA-DQB1. De plus, la présence d'une LTR du groupe MER11C (HERV-K11), située dans la même région, entraîne une baisse du niveau d'expression du gène HLA-DQB1. Cette LTR contient un site de fixation du facteur de transcription CTCF, comme celle trouvée proche des gènes IFI44 et IFI44L. Ce facteur a pour rôle , entre beaucoup d'autres, de réguler la structure dimensionnelle de la chromatine, et définit également les limites entre chromatine active et hétérochromatine (Filippova et al., 1996; Kim et al., 2007; Rubio et al., 2008). Il a également été montré que CTCF contrôle des gènes de la région HLA, notamment HLA-DRB1 et HLA-DQA1 (Majumder et al., 2008). On peut ainsi émettre l'hypothèse que la LTR, quand elle est présente, amène un site de fixation à CTCF et entraine un repliement de l'ADN empêchant la transcription du gène HLA_DQB1. Il sera nécessaire de mettre en place une étude pour valider ou non cette hypothèse, en associant par exemple nos données avec les zones de repliement de l'ADN et les marques histones dans les cellules du sang (Rao et al., 2014).

De manière intéressante, un HERV du groupe HML-3 (ou THE1A) a été suspecté d'avoir un rôle promoteur sur ce même gène HLA-DQB1 et dont la présence serait liée à un facteur de risque plus important dans la maladie d'Addison et le diabète de type I (Donner et al., 1999; Krach et al., 2003; Pani et al., 2002). Nous avons également trouvé un provirus HERV du groupe HML-3, dont la présence est associée à une diminution de l'expression du gène non codant HLA_DRB6. Le HERV est situé en amont du gène, et la LTR chevauche avec la 5'UTR d'un transcrit alternatif du gène. Il n'y a pas de site CTCF dans ce HERV, mais on pourrait penser que la présence de ce HERV empêche l'initiation de l'expression d'un des transcrits de HLA_DRB6 réduisant ainsi son expression. Au-delà du super-groupe des HERV-K, nous avons trouvé parmi les HERV les plus variables entre populations, un nombre important de HERV du groupe MER41, connus pour comporter des sites de fixation au facteur de transcription STAT1 (Chuong et al., 2016). En recherchant des sites potentiels de fixation des facteurs de transcription STAT1 et IRF1, 60% des séquences des HERV apparentés au groupe MER41 comportent des sites de fixation potentiels (données non montrées). Aucun lien n'a été fait avec l'expression de gènes des voies STAT1 et IRF1, cependant ces éléments peuvent représenter une base intéressante pour étudier l'impact du polymorphisme du groupe MER41 sur l'expression de gènes. Dans leur étude, Chuong et al. suggèrent que l'apport de site de fixation aux facteurs de transcription par les LTR des HERV, en se dupliquant dans

tout le génome, a permis de construire ou moduler plus rapidement des réseaux de régulation de gènes permettant à l'hôte de mieux s'adapter face à de nouveaux pathogènes. Ils décrivent le rôle du groupe MER41 dans la mise en place du réseau de réponse à l'interféron gamma, et ont validé un locus proche du gène AIM2. De la même manière, le polymorphisme de présence HERV aux abords des gènes du système HLA, entraînerait une diversité de régulation et d'expression de la réponse de l'hôte plus importante entre les individus et pourrait ainsi participer à une plus grande adaptabilité de l'espèce face à des agressions immunitaires. Cela va dans le sens d'une des théories expliquant le polymorphisme HLA dirigé par les pathogènes. Les individus hétérozygotes aux loci HLA répondent à une plus grande diversité d'antigènes due à la co-expression des gènes HLA (Hughes and Nei, 1989; Lau et al., 2015; Spurgin and Richardson, 2010). Ces exemples sont peut-être le reflet d'une synergie évolutive entre des virus qui ont pu s'endogénéiser dans le génome de leur hôte et se transmettre par d'autres moyens que l'infection, et l'hôte qui se servirait de la capacité de rétrotransposition des rétrovirus pour obtenir un avantage sélectif dans la réponse aux pathogènes.

L'ensemble de ces exemples sont en faveur d'un rôle des HERV, notamment par leur présence ou leur absence dans le génome, sur la fonctionnalité des gènes dans la région HLA, et pourraient ainsi impacter la capacité des cellules immunitaires de l'hôte (macrophages, lymphocytes B) à présenter les antigènes aux lymphocytes T. Nous n'avons cependant pas eu l'occasion de réaliser des approches similaires sur des patients suivant une agression inflammatoire, afin de mieux connaître le possible impact et les mécanismes du polymorphisme HERV sur la réponse de l'hôte.

4.4 CONCLUSIONS ET PERSPECTIVES

Ces travaux ont permis de montrer que des HERV sont exprimés et modulés suivant une agression inflammatoire, notamment après le choc septique. Ils ont également permis de mettre en évidence une régulation de l'expression commune entre des HERV et des gènes de l'immunité. Parmi ceux étant modulés entre des patients en choc septique avec un statut immunitaire plus ou moins grave, certains semblent pertinents pour être sélectionnés

comme potentiels marqueurs de l'état d'immunodépression pour une utilisation en clinique et vont être validés dans de futures études. L'étude REALISM (Rol et al., 2017) a inclus plus de 100 volontaires sains et 378 patients ayant subi des agression inflammatoires graves (chocs septiques, brulés, traumatisés, chirurgie lourde). Des prélèvements sanguins sur ces patients ont été collectés régulièrement pendant les deux mois suivant l'agression. Dans ce cadre, plus d'une dizaine de HERV vont être évalué par RTqPCR afin de connaître plus précisément leur niveau d'expression et de modulation.

Par différentes méthodes et différents outils spécifiquement développés pour ce projet, nous avons également mis en évidence des associations entre HERV et réponse de l'hôte, notamment suivant l'agression inflammatoire. Nous avons identifié plusieurs LTR pouvant agir en cis sur l'expression de gènes du système immunitaire, tels que CD209, OAS3, ou encore la famille des récepteurs à l'IL-1 suivant le choc septique ou d'autres agressions inflammatoires. Nous avons également vu que les HERV pouvaient être porteurs de sites de fixations de facteurs de transcription tels que STAT1 ou CTCF, et pourraient ainsi participer à la régulation de gènes, notamment ceux situés dans la région HLA. Enfin, nous avons associé données de séquençage du génome et du transcriptome dans le sang de volontaires sains, pour montrer que la présence de HERV peut moduler l'expression de gènes du système HLA et probablement impacter la réponse immunitaire de l'hôte. Par la suite, l'ensemble de ces candidats devront être validés expérimentalement et des études mécanistiques devront être mise en place afin de mieux comprendre leur modalités d'action et leur impact sur la réponse immunitaire de l'hôte.

Au-delà de représenter un nouveau vivier de biomarqueurs potentiels du statut immunitaire des patients suivant l'agression inflammatoire, ces résultats pourraient inciter les chercheurs à recentrer les HERV au cœur des analyses omiques, au même titre que les gènes. A terme, la multiplication des données et outils disponibles sur les HERV devrait permettre une meilleure compréhension de leur apport dans la modulation et la capacité d'adaptation du système immunitaire en réponse à de nouveaux pathogènes.

5 REFERENCES BIBLIOGRAPHIQUES

Agnese, D.M., Calvano, J.E., Hahm, S.J., Coyle, S.M., Corbett, S.A., Calvano, S.E., and Lowry, S.F. (2002). Human Toll-Like Receptor 4 Mutations but Not CD14 Polymorphisms Are Associated with an Increased Risk of Gram-Negative Infections. J. Infect. Dis. *186*, 1522–1525.

Aken, B.L., Achuthan, P., Akanni, W., Amode, M.R., Bernsdorff, F., Bhai, J., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., et al. (2017). Ensembl 2017. Nucleic Acids Res. *45*, D635–D642.

Allantaz-Frager, F., Turrel-Davin, F., Venet, F., Monnin, C., Jean, A.D.S., Barbalat, V., Cerrato, E., Pachot, A., Lepape, A., and Monneret, G. (2013). Identification of Biomarkers of Response to IFNg during Endotoxin Tolerance: Application to Septic Shock. PLOS ONE *8*, e68218.

Annane, D., Renault, A., Brun-Buisson, C., Megarbane, B., Quenot, J.-P., Siami, S., Cariou, A., Forceville, X., Schwebel, C., Martin, C., et al. (2018). Hydrocortisone plus Fludrocortisone for Adults with Septic Shock. N. Engl. J. Med. *378*, 809–818.

Antony, J.M., van Marle, G., Opii, W., Butterfield, D.A., Mallet, F., Yong, V.W., Wallace, J.L., Deacon, R.M., Warren, K., and Power, C. (2004). Human endogenous retrovirus glycoprotein-mediated induction of redox reactants causes oligodendrocyte death and demyelination. Nat. Neurosci. *7*, 1088–1095.

Apanius, V., Penn, D., Slev, P.R., Ruff, L.R., and Potts, W.K. (1997). The nature of selection on the major histocompatibility complex. Crit. Rev. Immunol. *17*, 179–224.

Assinger, A., Yaiw, K.-C., Göttesdorfer, I., Leib-Mösch, C., and Söderberg-Nauclér, C. (2013). Human Cytomegalovirus (HCMV) induces Human Endogenous Retrovirus (HERV) transcription. Retrovirology *10*, 132.

Babaian, A., and Mager, D.L. (2016). Endogenous retroviral promoter exaptation in human cancer. Mob. DNA *7*, 24.

Babaian, A., Romanish, M.T., Gagnier, L., Kuo, L.Y., Karimi, M.M., Steidl, C., and Mager, D.L. (2016). Onco-exaptation of an endogenous retroviral LTR drives *IRF5* expression in Hodgkin lymphoma. Oncogene *35*, 2542–2546.

Babaian, A., Lever, J., Gagnier, L., and Mager, D.L. (2017). LIONS: Analysis Suite for Detecting and Quantifying Transposable Element Initiated Transcription from RNA-seq. BioRxiv 149864.

Balestrieri, E., Pica, F., Matteucci, C., Zenobi, R., Sorrentino, R., Argaw-Denboba, A., Cipriani, C., Bucci, I., and Sinibaldi-Vallebona, P. (2015). Transcriptional Activity of Human Endogenous Retroviruses in Human Peripheral Blood Mononuclear Cells.

Bannert, N., and Kurth, R. (2004). Retroelements and the human genome: New perspectives on an old relation. Proc. Natl. Acad. Sci. *101*, 14572–14579.

Bannert, N., and Kurth, R. (2006). The Evolutionary Dynamics of Human Endogenous Retroviral Families. Annu. Rev. Genomics Hum. Genet. *7*, 149–173.

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA *6*, 11.

Becker, M., De Bastiani, M.A., Parisi, M.M., Guma, F.T.C.R., Markoski, M.M., Castro, M.A.A., Kaplan, M.H., Barbé-Tuana, F.M., and Klamt, F. (2015). Integrated Transcriptomics Establish Macrophage Polarization Signatures and have Potential Applications for Clinical Health and Disease. Sci. Rep. *5*, 13351.

Bengtsson, A., Blomberg, J., Nived, O., Pipkorn, R., Toth, L., and 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*, 1654–1663.

Blond, J.-L., Besème, F., Duret, L., Bouton, O., Bedin, F., Perron, H., Mandrand, B., and Mallet, F. (1999). Molecular Characterization and Placental Expression of HERV-W, a New Human Endogenous Retrovirus Family. J. Virol. *73*, 1175–1185.

Bolze, P.-A., Mommert, M., and Mallet, F. (2017). Contribution of Syncytins and Other Endogenous Retroviral Envelopes to Human Placenta Pathologies. Prog. Mol. Biol. Transl. Sci. *145*, 111–162.

Boomer, J.S., To, K., Chang, K.C., Takasu, O., Osborne, D.F., Walton, A.H., Bricker, T.L., Jarman, S.D., Kreisel, D., Krupnick, A.S., et al. (2011). Immunosuppression in patients who die of sepsis and multiple organ failure. JAMA *306*, 2594–2605.

Brind'Amour, J., Kobayashi, H., Albert, J.R., Shirane, K., Sakashita, A., Kamio, A., Bogutz, A., Koike, T., Karimi, M.M., Lefebvre, L., et al. (2018). LTR retrotransposons transcribed in oocytes drive species-specific and heritable changes in DNA methylation. Nat. Commun. *9*, 3331.

Brodsky, I., Foley, B., and Gillespie, D. (1993). Expression of human endogenous retrovirus (HERV-K) in chronic myeloid leukemia. Leuk. Lymphoma *11 Suppl 1*, 119–123.

Brudek, T., Christensen, T., Aagaard, L., Petersen, T., Hansen, H.J., and Møller-Larsen, A. (2009). B cells and monocytes from patients with active multiple sclerosis exhibit increased surface expression of both HERV-H Env and HERV-W Env, accompanied by increased seroreactivity. Retrovirology *6*, 104.

Brun-Buisson, C., Roudot-Thoraval, F., Girou, E., Grenier-Sennelier, C., and Durand-Zaleski, I. (2003). The costs of septic syndromes in the intensive care unit and influence of hospital-acquired sepsis. Intensive Care Med. *29*, 1464–1471.

Campbell, I.M., Gambin, T., Dittwald, P., Beck, C.R., Shuvarikov, A., Hixson, P., Patel, A., Gambin, A., Shaw, C.A., Rosenfeld, J.A., et al. (2014). Human endogenous retroviral elements promote genome instability via non-allelic homologous recombination. BMC Biol. *12*, 74.

Canales, R.D., Luo, Y., Willey, J.C., Austermiller, B., Barbacioru, C.C., Boysen, C., Hunkapiller, K., Jensen, R.V., Knight, C.R., Lee, K.Y., et al. (2006). Evaluation of DNA microarray results with quantitative gene expression platforms. Nat. Biotechnol. *24*, 1115–1122.

Cavaillon, J.-M., and Adib-Conquy, M. (2006). Bench-to-bedside review: Endotoxin tolerance as a model of leukocyte reprogramming in sepsis. Crit. Care *10*, 233.

Christensen, T., Jensen, A.W., Munch, M., Haahr, S., Sørensen, P.D., Riemann, H., Hansen, H.J., and Møller-Larsen, A. (1997). Characterization of retroviruses from patients with multiple sclerosis. Acta Neurol. Scand. Suppl. *169*, 49–58.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science *351*, 1083–1087.

Cohen, C.J., Lock, W.M., and Mager, D.L. (2009). Endogenous retroviral LTRs as promoters for human genes: A critical assessment. Gene *448*, 105–114.

Consortium, T. 1000 G.P. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061.

Contreras-Galindo, R., Kaplan, M.H., Markovitz, D.M., Lorenzo, E., and Yamamura, Y. (2006). Detection of HERV-K(HML-2) viral RNA in plasma of HIV type 1-infected individuals. AIDS Res. Hum. Retroviruses *22*, 979–984.

Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. *10*, 691–703.

Cordaux, R., Hedges, D.J., Herke, S.W., and Batzer, M.A. (2006). Estimating the retrotransposition rate of human Alu elements. Gene *373*, 134–137.

Cotton, J. (2001). Retroviruses from retrotransposons. Genome Biol. *2*, reports0006.

Danai, P.A., Moss, M., Mannino, D.M., and Martin, G.S. (2006). The epidemiology of sepsis in patients with malignancy. Chest *129*, 1432–1440.

Davenport, E.E., Burnham, K.L., Radhakrishnan, J., Humburg, P., Hutton, P., Mills, T.C., Rautanen, A., Gordon, A.C., Garrard, C., Hill, A.V.S., et al. (2016). Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study. Lancet Respir. Med. *4*, 259–271.

Demaret, J., Venet, F., Friggeri, A., Cazalis, M.-A., Plassais, J., Jallades, L., Malcus, C., Poitevin-Later, F., Textoris, J., Lepape, A., et al. (2015). Marked alterations of neutrophil functions during sepsis-induced immunosuppression. J. Leukoc. Biol. *98*, 1081–1090.

Deutschman, C.S., and Tracey, K.J. (2014). Sepsis: Current Dogma and New Perspectives. Immunity *40*, 463–475.

Dewannieux, M., Harper, F., Richaud, A., Letzelter, C., Ribet, D., Pierron, G., and Heidmann, T. (2006). Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. Genome Res. *16*, 1548–1556.

van Dissel, J.T., van Langevelde, P., Westendorp, R.G., Kwappenberg, K., and Frölich, M. (1998). Anti-inflammatory cytokine profile and mortality in febrile patients. Lancet Lond. Engl. *351*, 950–953.

Doherty, P.C., and Zinkernagel, R.M. (1975). A biological role for the major histocompatibility antigens. Lancet Lond. Engl. *1*, 1406–1409.

Donadi, E.A., Castelli, E.C., Arnaiz-Villena, A., Roger, M., Rey, D., and Moreau, P. (2011). Implications of the polymorphism of HLA-G on its function, regulation, evolution and disease association. Cell. Mol. Life Sci. *68*, 369–395.

Donner, H., Tönjes, R.R., Bontrop, R.E., Kurth, R., Usadel, K.H., and 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.

Donovan, J., Whitney, G., Rath, S., and Korennykh, A. (2015). Structural mechanism of sensing long dsRNA via a noncatalytic domain in human oligoadenylate synthetase 3. Proc. Natl. Acad. Sci. U. S. A. *112*, 3949–3954.

Drewry, A.M., Samra, N., Skrupky, L.P., Fuller, B.M., Compton, S.M., and Hotchkiss, R.S. (2014). Persistent lymphopenia after diagnosis of sepsis predicts mortality. Shock Augusta Ga *42*, 383–391.

Eichacker, P.Q., Parent, C., Kalil, A., Esposito, C., Cui, X., Banks, S.M., Gerstenberger, E.P., Fitz, Y., Danner, R.L., and Natanson, C. (2002). Risk and the efficacy of antiinflammatory agents: retrospective and confirmatory studies of sepsis. Am. J. Respir. Crit. Care Med. *166*, 1197–1205.

Ellison, C.E., and Bachtrog, D. (2013). Dosage Compensation via Transposable Element Mediated Rewiring of a Regulatory Network. Science *342*, 846–850.

Escalona, M., Rocha, S., and Posada, D. (2016). A comparison of tools for the simulation of genomic next-generation sequencing data. Nat. Rev. Genet. *17*, 459–469.

Esper, A.M., Moss, M., and Martin, G.S. (2009). The effect of diabetes mellitus on organ dysfunction with sepsis: an epidemiological study. Crit. Care Lond. Engl. *13*, R18.

Ewing, A.D. (2015). Transposable element detection from whole genome sequence data. Mob. DNA *6*.

Fairfax, B.P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., et al. (2014). Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. Science *343*, 1246949.

Ferrer, R., Martin-Loeches, I., Phillips, G., Osborn, T.M., Townsend, S., Dellinger, R.P., Artigas, A., Schorr, C., and Levy, M.M. (2014). Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. Crit. Care Med. *42*, 1749–1755.

Feschotte, C., and Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. Nat. Rev. Genet. *13*, 283–296.

Filippova, G.N., Fagerlie, S., Klenova, E.M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P.E., Collins, S.J., and Lobanenkov, V.V. (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. Mol. Cell. Biol. *16*, 2802–2813.

Finn, O.J. (2012). Host response in tumor diagnosis and prognosis: Importance of immunologists and pathologists alliance. Exp. Mol. Pathol. *93*, 315–318.

Fleischmann, C., Scherag, A., Adhikari, N.K.J., Hartog, C.S., Tsaganos, T., Schlattmann, P., Angus, D.C., Reinhart, K., and International Forum of Acute Care Trialists (2016). Assessment of Global Incidence and Mortality of Hospital-treated Sepsis. Current Estimates and Limitations. Am. J. Respir. Crit. Care Med. *193*, 259–272.

Flockerzi, A., Ruggieri, A., Frank, O., Sauter, M., Maldener, E., Kopper, B., Wullich, B., Seifarth, W., Müller-Lantzsch, N., Leib-Mösch, C., 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.

Forsman, A., Yun, Z., Hu, L., Uzhameckis, D., Jern, P., and 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.

Friedman, G., Silva, E., and Vincent, J.L. (1998). Has the mortality of septic shock changed with time. Crit. Care Med. *26*, 2078–2086.

Galbois, A., Aegerter, P., Martel-Samb, P., Housset, C., Thabut, D., Offenstadt, G., Ait-Oufella, H., Maury, E., Guidet, B., and Collège des Utilisateurs des Bases des données en Réanimation (CUB-Réa) Group (2014). Improved prognosis of septic shock in patients with cirrhosis: a multicenter study*. Crit. Care Med. *42*, 1666–1675.

Gifford, R.J., Blomberg, J., Coffin, J.M., Fan, H., Heidmann, T., Mayer, J., Stoye, J., Tristem, M., and Johnson, W.E. (2018). Nomenclature for endogenous retrovirus (ERV) loci. Retrovirology *15*, 59.

Gimenez, J., Montgiraud, C., Oriol, G., Pichon, J.-P., Ruel, K., Tsatsaris, V., Gerbaud, P., Frendo, J.-L., Evain-Brion, D., and Mallet, F. (2009). Comparative Methylation of ERVWE1/Syncytin-1 and Other Human Endogenous Retrovirus LTRs in Placenta Tissues. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes *16*, 195–211.

Gimenez, J., Montgiraud, C., Pichon, J.-P., Bonnaud, B., Arsac, M., Ruel, K., Bouton, O., and 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.

Gogvadze, E., Stukacheva, E., Buzdin, A., and Sverdlov, E. (2009). Human-specific modulation of transcriptional activity provided by endogenous retroviral insertions. J. Virol. *83*, 6098–6105.

Goldszmid, R.S., Dzutsev, A., and Trinchieri, G. (2014). Host Immune Response to Infection and Cancer: Unexpected Commonalities. Cell Host Microbe *15*, 295–305.

Gosenca, D., Gabriel, U., Steidler, A., Mayer, J., Diem, O., Erben, P., Fabarius, A., Leib-Mösch, C., Hofmann, W.-K., and Seifarth, W. (2012). HERV-E-Mediated Modulation of PLA2G4A Transcription in Urothelial Carcinoma. PLOS ONE *7*, e49341.

Gregory, T.R. (2005). Synergy between sequence and size in Large-scale genomics. Nat. Rev. Genet. *6*, 699–708.

Grimaldi, D., Llitjos, J.F., and Pène, F. (2014). Post-infectious immune suppression: a new paradigm of severe infections. Med. Mal. Infect. *44*, 455–463.

Gross, H., Barth, S., Pfuhl, T., Willnecker, V., Spurk, A., Gurtsevitch, V., Sauter, M., Hu, B., Noessner, E., Mueller-Lantzsch, N., et al. (2011). The NP9 protein encoded by the human endogenous retrovirus HERV-K(HML-2) negatively regulates gene activation of the Epstein-Barr virus nuclear antigen 2 (EBNA2). Int. J. Cancer *129*, 1105–1115.

Grow, E.J., Flynn, R.A., Chavez, S.L., Bayless, N.L., Wossidlo, M., Wesche, D.J., Martin, L., Ware, C.B., Blish, C.A., Chang, H.Y., et al. (2015). Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature *522*, 221–225.

Hancks, D.C., and Kazazian, H.H. (2012). Active human retrotransposons: variation and disease. Curr. Opin. Genet. Dev. *22*, 191–203.

Hishikawa, T., Ogasawara, H., Kaneko, H., Shirasawa, T., Matsuura, Y., Sekigawa, I., Takasaki, Y., Hashimoto, H., Hirose, S., Handa, S., et al. (1997). Detection of antibodies to a recombinant gag protein derived from human endogenous retrovirus clone 4-1 in autoimmune diseases. Viral Immunol. *10*, 137–147.

Ho, J., Chan, H., Wong, S.H., Wang, M.H.T., Yu, J., Xiao, Z., Liu, X., Choi, G., Leung, C.C.H., Wong, W.T., et al. (2016). The involvement of regulatory non-coding RNAs in sepsis: a systematic review. Crit. Care *20*, 383.

Hoogen, A. van den, Gerards, L.J., Verboon-Maciolek, M.A., Fleer, A., and Krediet, T.G. (2010). Long-Term Trends in the Epidemiology of Neonatal Sepsis and Antibiotic Susceptibility of Causative Agents. Neonatology *97*, 22–28.

Hotchkiss, R.S., Monneret, G., and Payen, D. (2013). Sepsis-induced immunosuppression: from cellular dysfunctions to immunotherapy. Nat. Rev. Immunol. *13*, 862–874.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and Applications of CRISPR-Cas9 for Genome Engineering. Cell *157*, 1262–1278.

Huang, X., Venet, F., Wang, Y.L., Lepape, A., Yuan, Z., Chen, Y., Swan, R., Kherouf, H., Monneret, G., Chung, C.-S., et al. (2009). PD-1 expression by macrophages plays a pathologic role in altering microbial clearance and the innate inflammatory response to sepsis. Proc. Natl. Acad. Sci. U. S. A. *106*, 6303–6308.

Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F.A., and Wheeler, T.J. (2016). The Dfam database of repetitive DNA families. Nucleic Acids Res. *44*, D81–D89.

Hughes, A.L., and Nei, M. (1989). Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc. Natl. Acad. Sci. *86*, 958–962.

Hughes, J.F., and Coffin, J.M. (2005). Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. Genetics *171*, 1183–1194.

Humer, J., Waltenberger, A., Grassauer, A., Kurz, M., Valencak, J., Rapberger, R., Hahn, S., Löwer, R., Wolff, K., Bergmann, M., et al. (2006). Identification of a melanoma marker derived from melanoma-associated endogenous retroviruses. Cancer Res. *66*, 1658–1663.

Hurst, T.P., and Magiorkinis, G. (2015). Activation of the innate immune response by endogenous retroviruses. J. Gen. Virol. *96*, 1207–1218.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostat. Oxf. Engl. *4*, 249–264.

Isaak, D.D., and Cerny, J. (1983). T and B lymphocyte susceptibility to murine leukemia virus moloney. Infect. Immun. *40*, 977–984.

Iwashyna, T.J., Ely, E.W., Smith, D.M., and Langa, K.M. (2010). Long-term cognitive impairment and functional disability among survivors of severe sepsis. JAMA *304*, 1787–1794.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. *36*, 338–345.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

Johnston, J.B., Silva, C., Holden, J., Warren, K.G., Clark, A.W., and Power, C. (2001). Monocyte activation and differentiation augment human endogenous retrovirus expression: implications for inflammatory brain diseases. Ann. Neurol. *50*, 434–442.

Kang, J.-I., Kwon, S.-N., Park, S.-H., Kim, Y.K., Choi, S.-Y., Kim, J.P., and Ahn, B.-Y. (2009). PKR protein kinase is activated by hepatitis C virus and inhibits viral replication through translational control. Virus Res. *142*, 51–56.

Karimi, M.M., Goyal, P., Maksakova, I.A., Bilenky, M., Leung, D., Tang, J.X., Shinkai, Y., Mager, D.L., Jones, S., Hirst, M., et al. (2011). DNA methylation and SETDB1/H3K9me3 regulate

predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. Cell Stem Cell *8*, 676–687.

Kawai, T., and Akira, S. (2010). The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. Nat. Immunol. *11*, 373–384.

Kim, D.S., and Hahn, Y. (2010). Human-specific antisense transcripts induced by the insertion of transposable element. Int. J. Mol. Med. *26*, 151–157.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231–1245.

Klebanov, L., and Yakovlev, A. (2007). How high is the level of technical noise in microarray data? Biol. Direct *2*, 9.

Kollef, K.E., Schramm, G.E., Wills, A.R., Reichley, R.M., Micek, S.T., and Kollef, M.H. (2008). Predictors of 30-day mortality and hospital costs in patients with ventilator-associated pneumonia attributed to potentially antibiotic-resistant gram-negative bacteria. Chest *134*, 281–287.

Korlach, J. (2015). Understanding Accuracy in SMRT Sequencing.

Krach, K., Badenhoop, K., and Tönjes, R.R. (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.

Kwon, D.-N., Lee, Y.-K., Greenhalgh, D.G., and Cho, K. (2011). Lipopolysaccharide stress induces cell-type specific production of murine leukemia virus type-endogenous retroviral virions in primary lymphoid cells. J. Gen. Virol. *92*, 292–300.

van de Lagemaat, L.N., Landry, J.-R., Mager, D.L., and Medstrand, P. (2003). Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet. TIG *19*, 530–536.

Lamprecht, B., Walter, K., Kreher, S., Kumar, R., Hummel, M., Lenze, D., Köchert, K., Bouhlel, M.A., Richter, J., Soler, E., et al. (2010). Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. Nat. Med. *16*, 571–579.

Lappalainen, T., Sammeth, M., Friedländer, M.R., ‘t Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Laska, M.J., Brudek, T., Nissen, K.K., Christensen, T., Møller-Larsen, A., Petersen, T., and Nexø, B.A. (2012). Expression of HERV-Fc1, a Human Endogenous Retrovirus, Is Increased in Patients with Active Multiple Sclerosis. J. Virol. *86*, 3713–3722.

Lättekivi, F., Kõks, S., Keermann, M., Reimann, E., Prans, E., Abram, K., Silm, H., Kõks, G., and Kingo, K. (2018). Transcriptional landscape of human endogenous retroviruses (HERVs) and other repetitive elements in psoriatic skin. Sci. Rep. *8*, 4358.

Lau, Q., Yasukochi, Y., and Satta, Y. (2015). A limit to the divergent allele advantage model supported by variable pathogen recognition across HLA-DRB1 allele lineages. Tissue Antigens *86*, 343–352.

Laudanski, K., Miller-Graziano, C., Xiao, W., Mindrinos, M.N., Richards, D.R., De, A., Moldawer, L.L., Maier, R.V., Bankey, P., Baker, H.V., et al. (2006). Cell-specific expression and pathway analyses reveal alterations in trauma-related human T cell and monocyte pathways. Proc. Natl. Acad. Sci. U. S. A. *103*, 15564–15569.

Lee, Y.N., and Bieniasz, P.D. (2007). Reconstitution of an Infectious Human Endogenous Retrovirus. PLOS Pathog. *3*, e10.

Lee, K.-H., Rah, H., Green, T., Lee, Y.-K., Lim, D., Nemzek, J., Wahl, W., Greenhalgh, D., and Cho, K. (2014). Divergent and dynamic activity of endogenous retroviruses in burn patients and their inflammatory potential. Exp. Mol. Pathol. *96*, 178–187.

Lee, Y.-J., Jeong, B.-H., Park, J.-B., Kwon, H.-J., Kim, Y.-S., and Kwak, I.-S. (2013). The prevalence of human endogenous retroviruses in the plasma of major burn patients. Burns *39*, 1200–1205.

Lengauer, T. (2008). Bioinformatics — From Genomes to Therapies. In Bioinformatics-From Genomes to Therapies, (John Wiley & Sons, Ltd), pp. 33–47.

Levet, S., Medina, J., Joanou, J., Demolder, A., Queruel, N., Réant, K., Normand, M., Seffals, M., Dimier, J., Germi, R., et al. (2017). An ancestral retroviral protein identified as a therapeutic target in type-1 diabetes. JCI Insight *2*.

Li, F., Nellåker, C., Sabunciyan, S., Yolken, R.H., Jones-Brando, L., Johansson, A.-S., Owe-Larsson, B., and Karlsson, H. (2014). Transcriptional derepression of the ERVWE1 locus following influenza A virus infection. J. Virol. *88*, 4328–4337.

Lin, G., Simmons, G., Pöhlmann, S., Baribaud, F., Ni, H., Leslie, G.J., Haggarty, B.S., Bates, P., Weissman, D., Hoxie, J.A., et al. (2003). Differential N-linked glycosylation of human immunodeficiency virus and Ebola virus envelope glycoproteins modulates interactions with DC-SIGN and DC-SIGNR. J. Virol. *77*, 1337–1346.

Lorenz, E., Mira, J.P., Frees, K.L., and Schwartz, D.A. (2002). Relevance of mutations in the TLR4 receptor in patients with gram-negative septic shock. Arch. Intern. Med. *162*, 1028–1032.

Lukaszewicz, A.-C., Grienay, M., Resche-Rigon, M., Pirracchio, R., Faivre, V., Boval, B., and Payen, D. (2009). Monocytic HLA-DR expression in intensive care patients: interest for prognosis and secondary infection prediction. Crit. Care Med. *37*, 2746–2752.

Macfarlan, T.S., Gifford, W.D., Agarwal, S., Driscoll, S., Lettieri, K., Wang, J., Andrews, S.E., Franco, L., Rosenfeld, M.G., Ren, B., et al. (2011). Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. Genes Dev. *25*, 594–607.

Mager, D.L., Hunter, D.G., Schertzer, M., and Freeman, J.D. (1999). Endogenous retroviruses provide the primary polyadenylation signal for two new human genes (HHLA2 and HHLA3). Genomics *59*, 255–263.

Majumder, P., Gomez, J.A., Chadwick, B.P., and Boss, J.M. (2008). The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions. J. Exp. Med. *205*, 785–798.

Malik, H.S., Henikoff, S., and Eickbush, T.H. (2000). Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res. *10*, 1307–1318.

Mameli, G., Astone, V., Arru, G., Marconi, S., Lovato, L., Serra, C., Sotgiu, S., Bonetti, B., and Dolei, A. (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.

Mameli, G., Poddighe, L., Mei, A., Uleri, E., Sotgiu, S., Serra, C., Manetti, R., and Dolei, A. (2012). Expression and Activation by Epstein Barr Virus of Human Endogenous Retroviruses-W in Blood Cells and Astrocytes: Inference for Multiple Sclerosis. PLoS ONE *7*.

MAQC Consortium, Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J.A., Baker, S.C., Collins, P.J., de Longueville, F., Kawasaki, E.S., et al. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat. Biotechnol. *24*, 1151–1161.

Marchi, E., Kanapin, A., Magiorkinis, G., and Belshaw, R. (2014). Unfixed endogenous retroviral insertions in the human population. J. Virol. *88*, 9529–9537.

Marshall, E. (2004). Getting the Noise Out of Gene Arrays. Science *306*, 630–631.

Martin, G.S., Mannino, D.M., and Moss, M. (2006). The effect of age on the development and outcome of adult sepsis. Crit. Care Med. *34*, 15–21.

McCarthy, D.J., and Smyth, G.K. (2009). Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics *25*, 765–771.

Meakins, J.L., Pietsch, J.B., Bubenick, O., Kelly, R., Rode, H., Gordon, J., and MacLean, L.D. (1977). Delayed Hypersensitivity: Indicator of Acquired Failure of Host Defenses in Sepsis and Trauma. Ann. Surg. *186*, 241–249.

Medzhitov, R. (2007). Recognition of microorganisms and activation of the immune response. Nature *449*, 819–826.

Meisler, M.H., and Ting, C.N. (1993). The remarkable evolutionary history of the human amylase genes. Crit. Rev. Oral Biol. Med. Off. Publ. Am. Assoc. Oral Biol. *4*, 503–509.

Monneret, G., Finck, M.-E., Venet, F., Debard, A.-L., Bohé, J., Bienvenu, J., and Lepape, A. (2004). The anti-inflammatory response dominates after septic shock: association of low monocyte HLA-DR expression and high interleukin-10 concentration. Immunol. Lett. *95*, 193–198.

Monneret, G., Lepape, A., Voirin, N., Bohé, J., Venet, F., Debard, A.-L., Thizy, H., Bienvenu, J., Gueyffier, F., and Vanhems, P. (2006). Persisting low monocyte human leukocyte antigen-DR expression predicts mortality in septic shock. Intensive Care Med. *32*, 1175–1183.

Morishima, Y., Kashiwase, K., Matsuo, K., Azuma, F., Morishima, S., Onizuka, M., Yabe, T., Murata, M., Doki, N., Eto, T., et al. (2015). Biological significance of HLA locus matching in unrelated donor bone marrow transplantation. Blood *125*, 1189–1197.

Mortelmans, K., Wang-Johanning, F., and Johanning, G.L. (2016). The role of human endogenous retroviruses in brain development and function. APMIS Acta Pathol. Microbiol. Immunol. Scand. *124*, 105–115.

Mrus, J.M., Braun, L., Yi, M.S., Linde-Zwirble, W.T., and Johnston, J.A. (2005). Impact of HIV/AIDS on care and outcomes of severe sepsis. Crit. Care *9*, R623–R630.

Muradrasoli, S., Forsman, A., Hu, L., Blikstad, V., and Blomberg, J. (2006). Development of real-time PCRs for detection and quantitation of human MMTV-like (HML) sequences HML expression in human tissues. J. Virol. Methods *136*, 83–92.

Nagai, M., Furihata, T., Matsumoto, S., Ishii, S., Motohashi, S., Yoshino, I., Ugajin, M., Miyajima, A., Matsumoto, S., and Chiba, K. (2012). Identification of a new organic anion transporting polypeptide 1B3 mRNA isoform primarily expressed in human cancerous tissues and cells. Biochem. Biophys. Res. Commun. *418*, 818–823.

Nakada, T., Russell, J.A., Boyd, J.H., Thair, S.A., and Walley, K.R. (2015). Identification of a nonsynonymous polymorphism in the SVEP1 gene associated with altered clinical outcomes in septic shock. Crit. Care Med. *43*, 101–108.

Novakovic, B., Habibi, E., Wang, S.-Y., Arts, R.J.W., Davar, R., Megchelenbrink, W., Kim, B., Kuznetsova, T., Kox, M., Zwaag, J., et al. (2016). β-Glucan Reverses the Epigenetic State of LPS-Induced Immunological Tolerance. Cell *167*, 1354-1368.e14.

Oja, M., Peltonen, J., Blomberg, J., and Kaski, S. (2007). Methods for estimating human endogenous retrovirus activities from EST databases. BMC Bioinformatics *8*, S11.

Otto, G.P., Sossdorf, M., Claus, R.A., Rödel, J., Menge, K., Reinhart, K., Bauer, M., and Riedemann, N.C. (2011). The late phase of sepsis is characterized by an increased microbiological burden and death rate. Crit. Care Lond. Engl. *15*, R183.

Pachot, A., Monneret, G., Brion, A., Venet, F., Bohé, J., Bienvenu, J., Mougin, B., and Lepape, A. (2005). Messenger RNA expression of major histocompatibility complex class II genes in whole blood from septic shock patients. Crit. Care Med. *33*, 31–38; discussion 236-237.

Pachot, A., Cazalis, M.-A., Venet, F., Turrel, F., Faudot, C., Voirin, N., Diasparra, J., Bourgoin, N., Poitevin, F., Mougin, B., et al. (2008). Decreased Expression of the Fractalkine Receptor CX3CR1 on Circulating Monocytes as New Feature of Sepsis-Induced Immunosuppression. J. Immunol. *180*, 6421–6429.

Pani, M.A., Seidl, C., Bieda, K., Seissler, J., Krause, M., Seifried, E., Usadel, K.-H., and Badenhoop, K. (2002). Preliminary evidence that an endogenous retroviral long-terminal repeat (LTR13) at the HLA-DQB1 gene locus confers susceptibility to Addison's disease. Clin. Endocrinol. (Oxf.) *56*, 773–777.

Papazian, L., Fraisse, A., Garbe, L., Zandotti, C., Thomas, P., Saux, P., Pierrin, G., and Gouin, F. (1996). Cytomegalovirus. An unexpected cause of ventilator-associated pneumonia. Anesthesiology *84*, 280–287.

Parseval, N. de, Lazar, V., Casella, J.-F., Benit, L., and 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*, 10414–10422.

Paul, M., Shani, V., Muchtar, E., Kariv, G., Robenshtok, E., and Leibovici, L. (2010). Systematic Review and Meta-Analysis of the Efficacy of Appropriate Empiric Antibiotic Therapy for Sepsis. Antimicrob. Agents Chemother. *54*, 4851.

Peronnet, E., Venet, F., Maucort-Boulch, D., Friggeri, A., Cour, M., Argaud, L., Allaouchiche, B., Floccard, B., Aubrun, F., Rimmelé, T., et al. (2017). Association between mRNA expression of CD74 and IL10 and risk of ICU-acquired infections: a multicenter cohort study. Intensive Care Med. *43*, 1013–1020.

Pérot, P., Mugnier, N., Montgiraud, C., Gimenez, J., Jaillard, M., Bonnaud, B., and Mallet, F. (2012). Microarray-Based Sketches of the HERV Transcriptome Landscape. PLOS ONE *7*, e40194.

Pérot, P., Mullins, C.S., Naville, M., Bressan, C., Hühns, M., Gock, M., Kühn, F., Volff, J.-N., Trillet-Lenoir, V., Linnebacher, M., et al. (2015). Expression of young HERV-H loci in the course of colorectal carcinoma and correlation with molecular subtypes. Oncotarget *6*, 40095–40111.

Perron, H., Geny, C., Laurent, A., Mouriquand, C., Pellat, J., Perret, J., and Seigneurin, J.M. (1989). Leptomeningeal cell line from multiple sclerosis with reverse transcriptase activity and viral particles. Res. Virol. *140*, 551–561.

Perron, H., Lalande, B., Gratacap, B., Laurent, A., Genoulaz, O., Geny, C., Mallaret, M., Schuller, E., Stoebner, P., and Seigneurin, J.M. (1991). Isolation of retrovirus from patients with multiple sclerosis. The Lancet *337*, 862–863.

Perron, H., Garson, J.A., Bedin, F., Beseme, F., Paranhos-Baccala, G., Komurian-Pradel, F., Mallet, F., Tuke, P.W., Voisset, C., Blond, J.L., et al. (1997). Molecular identification of a novel retrovirus repeatedly isolated from patients with multiple sclerosis. The Collaborative Research Group on Multiple Sclerosis. Proc. Natl. Acad. Sci. U. S. A. *94*, 7583–7588.

Petersdorf, E.W., Longton, G.M., Anasetti, C., Martin, P.J., Mickelson, E.M., Smith, A.G., and Hansen, J.A. (1995). The significance of HLA-DRB1 matching on clinical outcome after HLA-A, B, DR identical unrelated donor marrow transplantation. Blood *86*, 1606–1613.

Pillay, J., Kamp, V.M., van Hoffen, E., Visser, T., Tak, T., Lammers, J.-W., Ulfman, L.H., Leenen, L.P., Pickkers, P., and Koenderman, L. (2012). A subset of neutrophils in human systemic inflammation inhibits T cell responses through Mac-1. J. Clin. Invest. *122*, 327–336.

Plassais, J., Venet, F., Cazalis, M.-A., Quang, D.L., Pachot, A., Monneret, G., Tissot, S., and Textoris, J. (2017). Transcriptome modulation by hydrocortisone in severe burn shock: ancillary analysis of a prospective randomized trial. Crit. Care *21*, 158.

de Prost, N., Razazi, K., and Brun-Buisson, C. (2013). Unrevealing culture-negative severe sepsis. Crit. Care *17*, 1001.

Prudencio, M., Gonzales, P.K., Cook, C.N., Gendron, T.F., Daughrity, L.M., Song, Y., Ebbert, M.T.W., van Blitterswijk, M., Zhang, Y.-J., Jansen-West, K., et al. (2017). Repetitive element transcripts are elevated in the brain of C9orf72 ALS/FTLD patients. Hum. Mol. Genet. *26*, 3421–3431.

Quenot, J.P., Pavon, A., Fournel, I., Barbar, S.D., and Bruyère, R. (2015). Le choc septique de l'adulte en France : vingt ans de données épidémiologiques. Réanimation *24*, 303–309.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell *159*, 1665–1680.

Renaudineau, Y., Vallet, S., Le Dantec, C., Hillion, S., Saraux, A., and Youinou, P. (2005). Characterization of the human CD5 endogenous retrovirus-E in B lymphocytes. Genes Immun. *6*, 663–671.

Rhoads, A., and Au, K.F. (2015). PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics *13*, 278–289.

Rhodes, A., Evans, L.E., Alhazzani, W., Levy, M.M., Antonelli, M., Ferrer, R., Kumar, A., Sevransky, J.E., Sprung, C.L., Nunnally, M.E., et al. (2017). Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016. Intensive Care Med. *43*, 304–377.

Rimmelé, T., and Kellum, J.A. (2012). High-volume hemofiltration in the intensive care unit: a blood purification therapy. Anesthesiology *116*, 1377–1387.

Rol, M.-L., Venet, F., Rimmele, T., Moucadel, V., Cortez, P., Quemeneur, L., Gardiner, D., Griffiths, A., Pachot, A., Textoris, J., et al. (2017). The REAnimation Low Immune Status Markers (REALISM) project: a protocol for broad characterisation and follow-up of injury-induced immunosuppression in intensive care unit (ICU) critically ill patients. BMJ Open *7*, e015734.

Rowe, H.M., Kapopoulou, A., Corsinotti, A., Fasching, L., Macfarlan, T.S., Tarabay, Y., Viville, S., Jakobsson, J., Pfaff, S.L., and Trono, D. (2013). TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. Genome Res. *23*, 452–461.

RStudio, Inc (2014). shiny: Easy web applications in R.

Rubartelli, A., and Lotze, M.T. (2007). Inside, outside, upside down: damage-associated molecular-pattern molecules (DAMPs) and redox. Trends Immunol. *28*, 429–436.

Rubio, E.D., Reiss, D.J., Welcsh, P.L., Disteche, C.M., Filippova, G.N., Baliga, N.S., Aebersold, R., Ranish, J.A., and Krumm, A. (2008). CTCF physically links cohesin to chromatin. Proc. Natl. Acad. Sci. U. S. A. *105*, 8309–8314.

Saeed, S., Quintin, J., Kerstens, H.H.D., Rao, N.A., Aghajanirefah, A., Matarese, F., Cheng, S.-C., Ratter, J., Berentsen, K., Ent, M.A. van der, et al. (2014). Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. Science *345*, 1251086.

Sarkar, S.N., Ghosh, A., Wang, H.W., Sung, S.S., and Sen, G.C. (1999). The nature of the catalytic domain of 2'-5'-oligoadenylate synthetases. J. Biol. Chem. *274*, 25535–25542.

Scarfò, I., Pellegrino, E., Mereu, E., Kwee, I., Agnelli, L., Bergaggio, E., Garaffo, G., Vitale, N., Caputo, M., Machiorlatti, R., et al. (2016). Identification of a new subclass of ALK-negative ALCL expressing aberrant levels of ERBB4 transcripts. Blood *127*, 221–232.

Schmid, C.D., and Bucher, P. (2010). MER41 Repeat Sequences Contain Inducible STAT1 Binding Sites. PLOS ONE *5*, e11425.

Schmitz-Winnenthal, F.H., Galindo-Escobedo, L.V., Rimoldi, D., Geng, W., Romero, P., Koch, M., Weitz, J., Krempien, R., Niethammer, A.G., Beckhove, P., et al. (2007). Potential target antigens for immunotherapy in human pancreatic cancer. Cancer Lett. *252*, 290–298.

Scicluna, B.P., van Vught, L.A., Zwinderman, A.H., Wiewel, M.A., Davenport, E.E., Burnham, K.L., Nürnberg, P., Schultz, M.J., Horn, J., Cremer, O.L., et al. (2017). Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study. Lancet Respir. Med. *5*, 816–826.

Seifarth, W., Krause, U., Hohenadl, C., Baust, C., Hehlmann, R., and Leib-Mösch, C. (2000). Rapid identification of all known retroviral reverse transcriptase sequences with a novel versatile detection assay. AIDS Res. Hum. Retroviruses *16*, 721–729.

Shankar-Hari, M., and Rubenfeld, G.D. (2016). Understanding Long-Term Outcomes Following Sepsis: Implications and Challenges. Curr. Infect. Dis. Rep. *18*.

Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J.A., Baker, S.C., Collins, P.J., Longueville, F. de, Kawasaki, E.S., Lee, K.Y., et al. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat. Biotechnol. *24*, 1151–1161.

Simpson, G.R., Patience, C., Löwer, R., Tönjes, R.R., Moore, H.D., Weiss, R.A., and Boyd, M.T. (1996). Endogenous D-type (HERV-K) related sequences are packaged into retroviral particles in the placenta and possess open reading frames for reverse transcriptase. Virology *222*, 451–456.

Singer, M., De Santis, V., Vitale, D., and Jeffcoate, W. (2004). Multiorgan failure is an adaptive, endocrine-mediated, metabolic response to overwhelming systemic inflammation. Lancet Lond. Engl. *364*, 545–548.

Singer, M., Deutschman, C.S., Seymour, C.W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G.R., Chiche, J.-D., Coopersmith, C.M., et al. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA *315*, 801–810.

Smit, A., Hubley, R., and Green, P. (2013). RepeatMasker Open-4.0. 2013 - 2015. <http://www.repeatmasker.org>.

Spurgin, L.G., and Richardson, D.S. (2010). How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc. R. Soc. Lond. B Biol. Sci. *277*, 979–988.

Stauffer, Y., Theiler, G., Sperisen, P., Lebedev, Y., and Jongeneel, C.V. (2004). Digital expression profiles of human endogenous retroviral families in normal and cancerous tissues. Cancer Immun. *4*, 2.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.

Suntsova, M., Garazha, A., Ivanova, A., Kaminsky, D., Zhavoronkov, A., and Buzdin, A. (2015). Molecular functions of human endogenous retroviruses in health and disease. Cell. Mol. Life Sci. CMLS *72*, 3653–3675.

Tamori, A., and Kawada, N. (2013). HLA class II associated with outcomes of hepatitis B and C infections. World J. Gastroenterol. *19*, 5395–5401.

Teixeira-Silva, A., Silva, R.M., Carneiro, J., Amorim, A., and Azevedo, L. (2013). The role of recombination in the origin and evolution of Alu subfamilies. PloS One *8*, e64884.

Textoris, J., and Mallet, F. (2017). Immunosuppression and herpes viral reactivation in intensive care unit patients: one size does not fit all. Crit. Care *21*, 230.

Timmermans, K., Kox, M., Vaneker, M., van den Berg, M., John, A., van Laarhoven, A., van der Hoeven, H., Scheffer, G.J., and Pickkers, P. (2016). Plasma levels of danger-associated molecular patterns are associated with immune suppression in trauma patients. Intensive Care Med. *42*, 551–561.

Tomlins, S.A., Laxman, B., Dhanasekaran, S.M., Helgeson, B.E., Cao, X., Morris, D.S., Menon, A., Jing, X., Cao, Q., Han, B., et al. (2007). Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature *448*, 595–599.

Torio, C.M., and Andrews, R.M. (2006). National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2011: Statistical Brief #160. In Healthcare Cost and Utilization Project (HCUP) Statistical Briefs, (Rockville (MD): Agency for Healthcare Research and Quality (US)), p.

Treangen, T.J., and Salzberg, S.L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. *13*, 36–46.

Tu, Y., Stolovitzky, G., and Klein, U. (2002). Quantitative noise analysis for gene expression microarray experiments. Proc. Natl. Acad. Sci. *99*, 14031–14036.

Vargiu, L., Rodriguez-Tomé, P., Sperber, G.O., Cadeddu, M., Grandi, N., Blikstad, V., Tramontano, E., and Blomberg, J. (2016). Classification and characterization of human endogenous retroviruses; mosaic forms are common. Retrovirology *13*, 7.

Venet, F., Tissot, S., Debard, A.-L., Faudot, C., Crampé, C., Pachot, A., Ayala, A., and Monneret, G. (2007). Decreased monocyte human leukocyte antigen-DR expression after severe burn injury: Correlation with severity and secondary septic shock. Crit. Care Med. *35*, 1910–1917.

Venet, F., Chung, C.-S., Kherouf, H., Geeraert, A., Malcus, C., Poitevin, F., Bohé, J., Lepape, A., Ayala, A., and Monneret, G. (2009). Increased circulating regulatory T cells (CD4+CD25+CD127−) contribute to lymphocyte anergy in septic shock patients. Intensive Care Med. *35*, 678–686.

Venet, F., Foray, A.-P., Villars-Méchin, A., Malcus, C., Poitevin-Later, F., Lepape, A., and Monneret, G. (2012). IL-7 restores lymphocyte functions in septic patients. J. Immunol. Baltim. Md 1950 *189*, 5073–5081.

Venkatesh, B., Finfer, S., Cohen, J., Rajbhandari, D., Arabi, Y., Bellomo, R., Billot, L., Correa, M., Glass, P., Harward, M., et al. (2018). Adjunctive Glucocorticoid Therapy in Patients with Septic Shock. N. Engl. J. Med. *378*, 797–808.

Vincendeau, M., Göttesdorfer, I., Schreml, J.M.H., Wetie, A.G.N., Mayer, J., Greenwood, A.D., Helfer, M., Kramer, S., Seifarth, W., Hadian, K., et al. (2015). Modulation of human endogenous retrovirus (HERV) transcription during persistent and de novo HIV-1 infection. Retrovirology *12*, 27.

Vincent, J.-L., Marshall, J.C., Ñamendys-Silva, S.A., François, B., Martin-Loeches, I., Lipman, J., Reinhart, K., Antonelli, M., Pickkers, P., Njimi, H., et al. (2014). Assessment of the worldwide burden of critical illness: the Intensive Care Over Nations (ICON) audit. Lancet Respir. Med. *2*, 380–386.

Volff, J.-N. (2006). Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays *28*, 913–922.

Wallace, A.D., Wendt, G.A., Barcellos, L.F., de Smith, A.J., Walsh, K.M., Metayer, C., Costello, J.F., Wiemels, J.L., and Francis, S.S. (2018). To ERV Is Human: A Phenotype-Wide Scan Linking

Polymorphic Human Endogenous Retrovirus-K Insertions to Complex Phenotypes. Front. Genet. *9*, 298.

Wallace, T.A., Downey, R.F., Seufert, C.J., Schetter, A., Dorsey, T.H., Johnson, C.A., Goldman, R., Loffredo, C.A., Yan, P., Sullivan, F.J., et al. (2014). Elevated HERV-K mRNA expression in PBMC is associated with a prostate cancer diagnosis particularly in older men and smokers. Carcinogenesis *35*, 2074–2083.

Walton, A.H., Muenzer, J.T., Rasche, D., Boomer, J.S., Sato, B., Brownstein, B.H., Pachot, A., Brooks, T.L., Deych, E., Shannon, W.D., et al. (2014). Reactivation of Multiple Viruses in Patients with Sepsis. PLoS ONE *9*, e98819.

Wang, H.E., Szychowski, J.M., Griffin, R., Safford, M.M., Shapiro, N.I., and Howard, G. (2014). Long-term mortality after community-acquired sepsis: a longitudinal population-based cohort study. BMJ Open *4*, e004283.

Wang-Johanning, F., Frost, A.R., Johanning, G.L., Khazaeli, M.B., LoBuglio, A.F., Shaw, D.R., and Strong, T.V. (2001). Expression of human endogenous retrovirus k envelope transcripts in human breast cancer. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *7*, 1553–1560.

Webb, P.M., Merritt, M.A., Boyle, G.M., and Green, A.C. (2007). Microarrays and epidemiology: not the beginning of the end but the end of the beginning.. Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol. *16*, 637–638.

Weitzman, J.B. (2000). Sequence of the major histocompatibility complex. Genome Biol. *1*, reports021.

Wiesner, T., Lee, W., Obenauf, A.C., Ran, L., Murali, R., Zhang, Q.F., Wong, E.W.P., Hu, W., Scott, S.N., Shah, R.H., et al. (2015). Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. Nature *526*, 453–457.

Wildschutte, J.H., Williams, Z.H., Montesion, M., Subramanian, R.P., Kidd, J.M., and Coffin, J.M. (2016). Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proc. Natl. Acad. Sci. *113*, E2326–E2334.

Williams, T.M. (2001). Human Leukocyte Antigen Gene Polymorphism and the Histocompatibility Laboratory. J. Mol. Diagn. JMD *3*, 98–104.

Xue, J., Schmidt, S.V., Sander, J., Draffehn, A., Krebs, W., Quester, I., De Nardo, D., Gohel, T.D., Emde, M., Schmidleithner, L., et al. (2014). Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. Immunity *40*, 274–288.

Young, G.R., Eksmond, U., Salcedo, R., Alexopoulou, L., Stoye, J.P., and Kassiotis, G. (2012). Resurrection of endogenous retroviruses in antibody-deficient mice. Nature *491*, 774–778.

Young, G.R., Stoye, J.P., and Kassiotis, G. (2013). Are human endogenous retroviruses pathogenic? An approach to testing the hypothesis. Bioessays *35*, 794–803.

Young, G.R., Mavrommatis, B., and Kassiotis, G. (2014). Microarray analysis reveals global modulation of endogenous retroelement transcription by microbes. Retrovirology *11*, 59.

Young, G.R., Terry, S.N., Manganaro, L., Cuesta-Dominguez, A., Deikus, G., Bernal-Rubio, D., Campisi, L., Fernandez-Sesma, A., Sebra, R., Simon, V., et al. (2018). HIV-1 Infection of Primary CD4+ T Cells Regulates the Expression of Specific Human Endogenous Retrovirus HERV-K (HML-2) Elements. J. Virol. *92*.

Yu, P., Lübben, W., Slomka, H., Gebler, J., Konert, M., Cai, C., Neubrandt, L., Prazeres da Costa, O., Paul, S., Dehnert, S., et al. (2012). Nucleic acid-sensing Toll-like receptors are essential for the control of endogenous retrovirus viremia and ERV-induced tumors. Immunity *37*, 867–879.

Zarrei, M., MacDonald, J.R., Merico, D., and Scherer, S.W. (2015). A copy number variation map of the human genome. Nat. Rev. Genet. *16*, 172–183.

Züst, R., Cervantes-Barragan, L., Habjan, M., Maier, R., Neuman, B.W., Ziebuhr, J., Szretter, K.J., Baker, S.C., Barchet, W., Diamond, M.S., et al. (2011). Ribose 2'-O-methylation provides a molecular signature for the distinction of self and non-self mRNA dependent on the RNA sensor Mda5. Nat. Immunol. *12*, 137–143.

6 ANNEXES

6.1 ANNEXE 1 : EXEMPLE DE CONTROLE QUALITE DE LA PUCE

EXEMPLE DE RAPPORT DE CONTROLE QUALITE DE LA PUCE SUR LE JEU DE DONNEES DU MODELE A L'ENDOTOXINE TOLERANCE.

BIOMÉRIEUX

MEMO

Recherche & Développement / *Research & Development*
Unité Mixte HCL bioMérieux/ HCL bioMérieux Joint Unit
Hôpital Ed.Herriot, Pl. Arsonval. Pav P – 69347 Lyon - France

Destinataires / *To* : S. BLEIN, K. BRENGEL-PESCE, E. CERRATO, V. CHEYNET, F. MALLET, B.MEUNIER, M. MOMMERT, V. MOUCADEL, J. TEXTORIS, F. VENET

Expéditeur / From : Olivier TABONE **Date :** 29/11/2018
E-mail expéditeur / *E-mail sender* : olivier.tabone@ext.biomerieux.com *No of pages* :
Copies / *cc* :
Relecteurs: S. BLEIN & J. TEXTORIS

Objet / *subject* : Microarray Quality Controls for Endotoxin Tolerance Model study

Executive summary :

In this report, we present quality controls on HERV-V3 chips experiment. After recall of experiment description, we controlled quality of data generated from these arrays. The criteria studied in the report assess quality of sample preparation, quality of untreated (non-normalized) data and quality of normalized data. After evaluation we decided to remove array 10 because it did not passed more than 3 criteria. Then we corrected batching effect using ComBat algorithm on 2 variables: batch of experimentation and healthy volunteer.

Index

1. Experiment description

1.1 Endotoxin Tolerance Model

The Endotoxin Tolerance Model models in vitro the features of monocytes related to sepsis-induced immunosuppression. Ex vivo, the prior exposition of innate immune cells at small quantities of endotoxin (LPS) makes cells insensible at ulterior LPS stimulation. As observed on patients, this phenomenon is associated with monocytes and macrophages presenting functional alterations like a decreased production of pro-inflammatory cytokines (TNF alpha), an increased production of anti-inflammatory cytokines (IL-10), a decreased number of antigen presenting cells and decreased chemotactic and phagocytic properties. More, the absence of LPS-dependence response is correlated with negative regulation of TLR-associated signal pathways.

In the other hand, HERVs (Human Endogenous Retro Viruses), which are endogen element representing around 8 % of our genome (with MALR elements), are known to be mostly inactive in healthy conditions (except for syncytine, *Mallet et al. 2004*). But when unbalance of organism arises, some HERVs can become active. Hervs are involved in auto-immune diseases and cancer context like multiple sclerosis (*Laska et al. 2012*), breast cancer (*Rhyu et al., 2014*), testis cancer (*Gimenez et al. 2010*), colon cancer (*Pérot et al. 2015*). Knowing that, it seems interesting to look at the expression of these HERVs in immunosuppressed context modelized by the endotoxin tolerance model especially as very few data are available on HERV expression in sepsis-like context.

In order to identify the activity of HERVs elements within Endotoxin Tolerance Model (ET model), we used HERV-V3 chips developed by the LCR laboratory, based on Affymetrix technology. We want to define which HERVs are differentially expressed between 3 conditions:

- Non Stimulated (NS)
- LPS stimulated (LPS)
- Endotoxin Tolerance (ET) (LPS-stimulation 24h after the first LPS-stimulation).

1.2 Experiment questions

1.1.1 ***NS vs LPS:*** *Which HERVs are expressed when bacterial infection is simulated by LPS stimulation in cellular model?*

1.1.2 ***ET vs NS:*** *Which HERVs are expressed when monocytes anergy is modelized in cellular model ?*

1.1.3 ***LPS vs ET:*** *Which are the observed differences between infection and anergy ?*

1.3 Dataset description

In order to respond to these questions HERV-V3 chips have been made from 5 Healthy Volunteer's PBMC (Peripheral Blood Mononuclear Cell) samples.

For sample from each HV, 3 treatments have been made: NS, LPS and ET (3 * 5 = 15 different samples).

From each sample, triplicates have been made (15*3 = 45 samples in total).

This leads to a total of **45 arrays** generated.

1.4 Objectives

In this report we will focus on Quality Controls for these 45 HERV-V3 chips.

1.5 Data information

Raw data (.CEL) are available on:

B1493-Immunomonitoring_markers\Raw_Data\ET_model_HERVV3

All other data, source file or figure generated for this study can be found on :

B1493-Immunomonitoring_markers\Studies\HERV_V3\ET_model_HERVV3

2. Quality controls of sample preparation

2.1 RNA quality: RIN values

To assess the quality of samples, we can verify the RNA Integrity Number (RIN). We consider an RNA is of good quality when its RIN is upper 7. We considered the RIN comprised between 7 and 7-sd(RIN _values) are in an intermediate zone (yellow smiley) .RIN values for each sample are presented on (Figure 1).

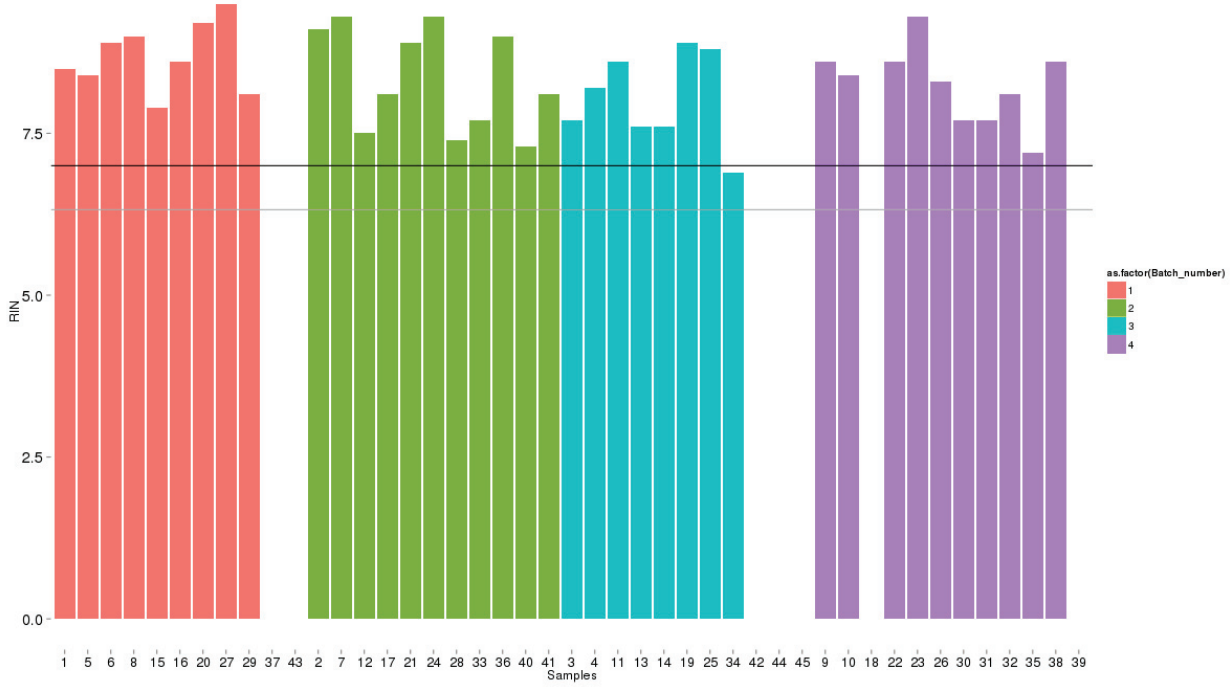


Figure 1: RIN values for the 45 samples. Colors correspond to batch number (hybridization runs).

We can see that RIN values from 7 arrays are missing (37, 43, 42, 44, 45, 18 and 39. The cause is a technical problem on the Agilent chip used to assess these RNA (mostly belongs to HV 7). RIN values for these RNA samples will be recalculated later during PCR validation). For the other arrays, RIN values are available and only 1 array (34) has a RIN value under 7 (6.9). 6.9 is upper the grey line (7 - sd(RIN_values)).

2.2 Amplification controls

The following plots represent the amplification controls on samples. An extract of product of amplification is passed on Agilent chip (as recommended by Nugen) and allow to generate plots . One graph represent the quantity of fluorescence as a function of the size of fragments for one sample. Here we control if theses plots look similar signifying that amplification worked on every sample. The distributions should look like Gaussian log transformed and three pikes should be visible, one around 25pb, one other at 200 nt and a last one between 500 and 1000 nt. Only some example plots are shown but all look similar (Figure 2).

N.B.: The RIN values visible on graphs don't have any meaning in this case since it is DNA (product of amplification) that is passed on Agilent chips.

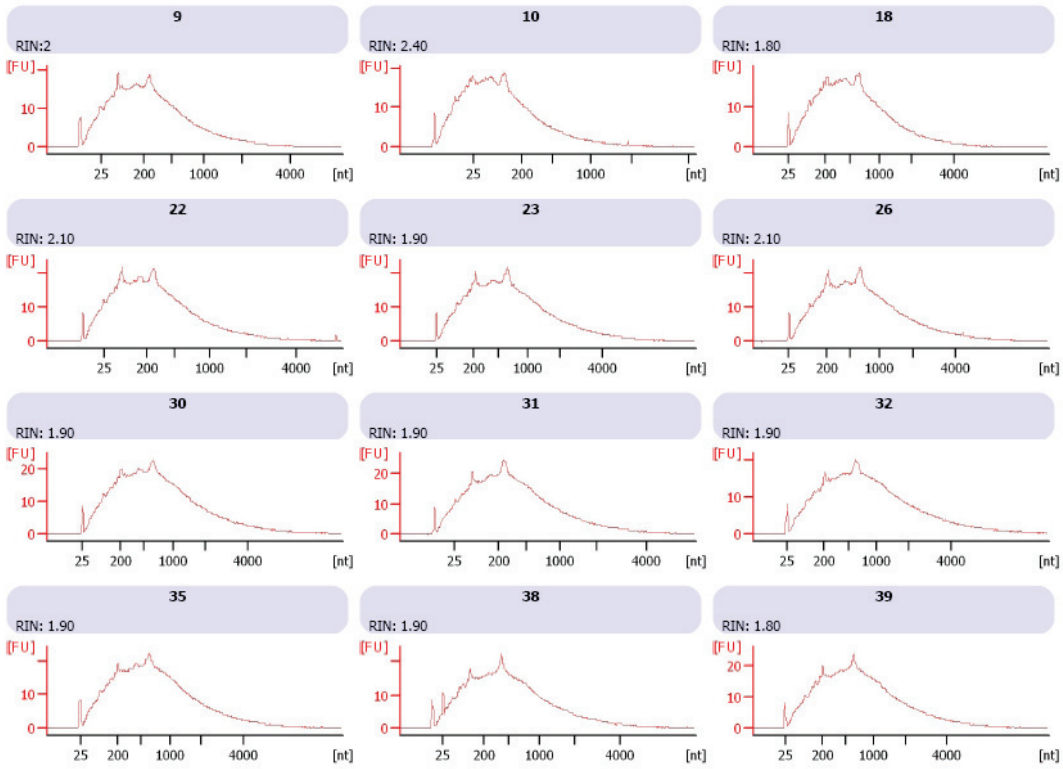


Figure 2: Amplification control plots.

2.3 Fragmentation controls

The following plots represent the fragmentation controls on samples. Like amplification controls, only some example plots are showed but all look similar (Figure 3). One graph represent the quantity of fluorescence as a function of the size of fragments for one sample. Here we control if these plots look similar signifying that fragments have the same size on all samples. Nugen recommends to have around 80% of fluorescence lower than 200 and the average peak near 85 bases. Here it is the case for all fragmentation plots.

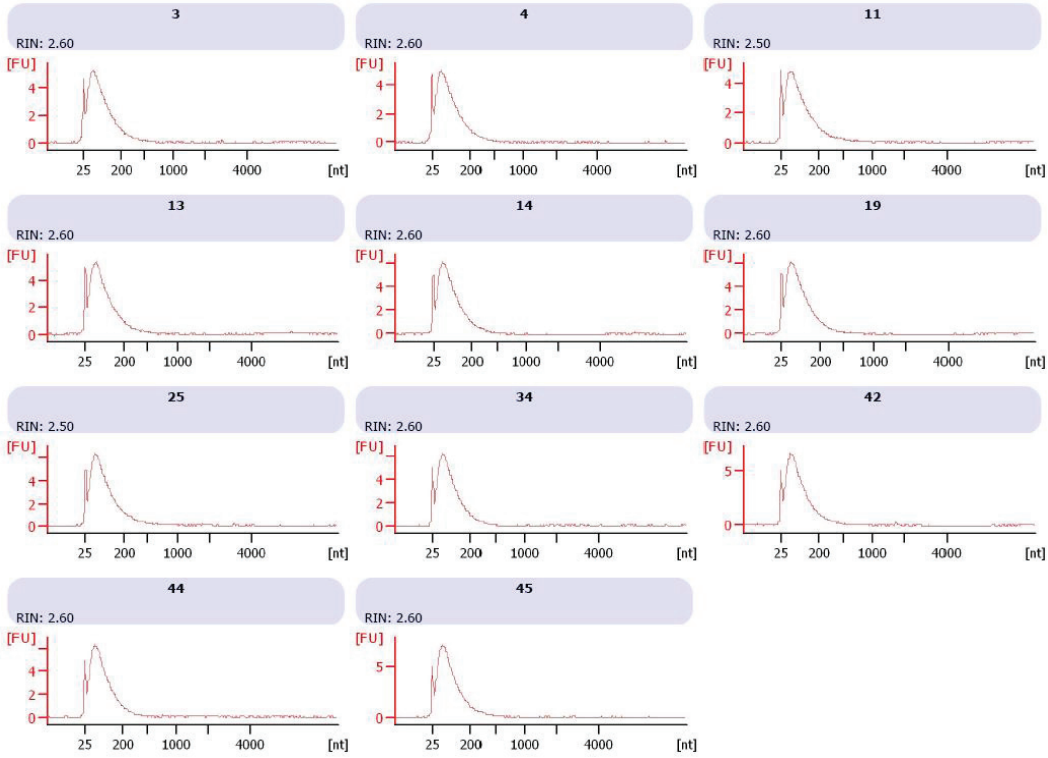


Figure 3: Fragmentation control plots

2.4 Positive and negative controls distributions

These are spots on the outer edge of the arrays that serve as controls for signal intensity and for the automatic setting of the grids. Elements with an intensity greater the 1.2 times the mean for that group are assumed to be positive controls. Elements with a signal less that 0.8 of the mean are assumed to be negative controls. Elements falling in between these cut offs are not used in further calculations. This graph (Figure 4) presents boxplots of the positive and negative distributions. The means and spread of positive elements should be comparable between arrays. Dissimilarity can arise either from non-uniform hybridization or gridding problems (rare). The negative elements represent spots with no hybridization signal, so they are expected to be close to 0. Boxplots that are strongly elevated relative to all the others reflect a higher background level.

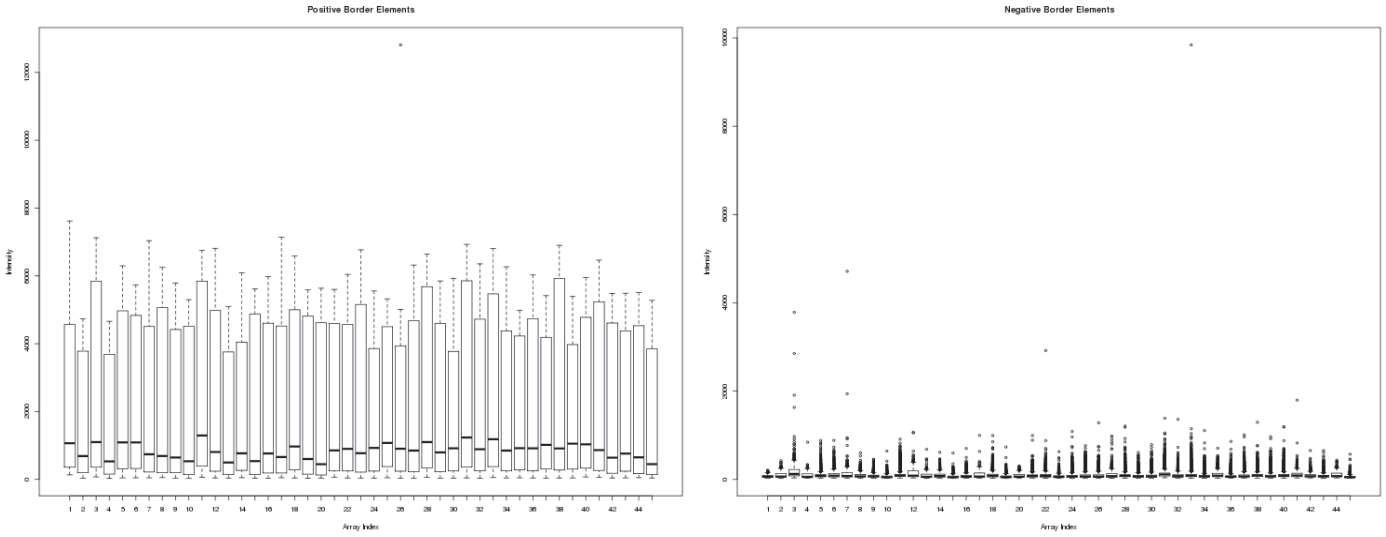


Figure 4: Border elements boxplots. The boxplots on the left represent positive border elements. The boxplots on the right represent negative border elements.

We can see that positive border elements have similar spread and median, sign of uniform hybridization on all arrays. Negative border elements are all close to 0, reflecting that no array have higher background level than the others.

To conclude on sample preparation controls, although RIN values are missing on some arrays (mostly those corresponding at HV 7), no outliers have been detected on RNA quality, amplification and fragmentation step, and hybridization uniformity.

3. Quality controls on raw data

3.1 Signal boxplots

We looked at the intensities boxplots on raw data (Figure 5).

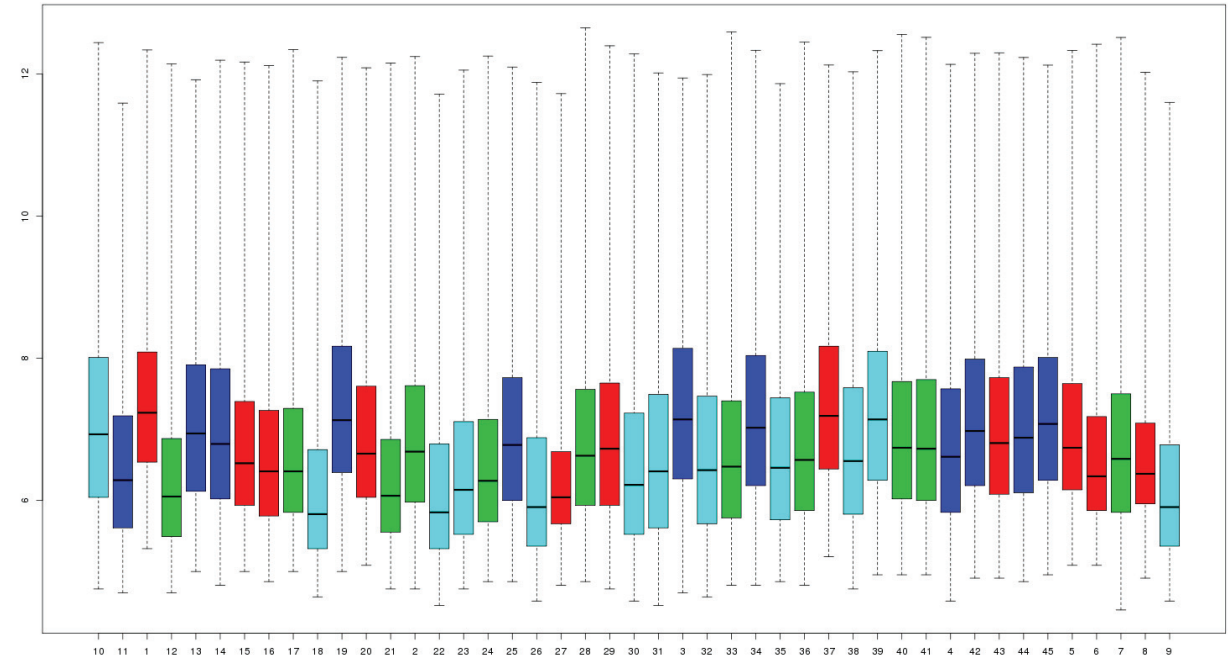


Figure 5 : Boxplots of raw intensities (log2 transformed)

We applied an outlier detection performed by computing the Kolmogorov-Smirnov (KS) statistic between each array's distribution and the distribution of the pooled data. The results are presented on Figure 6.

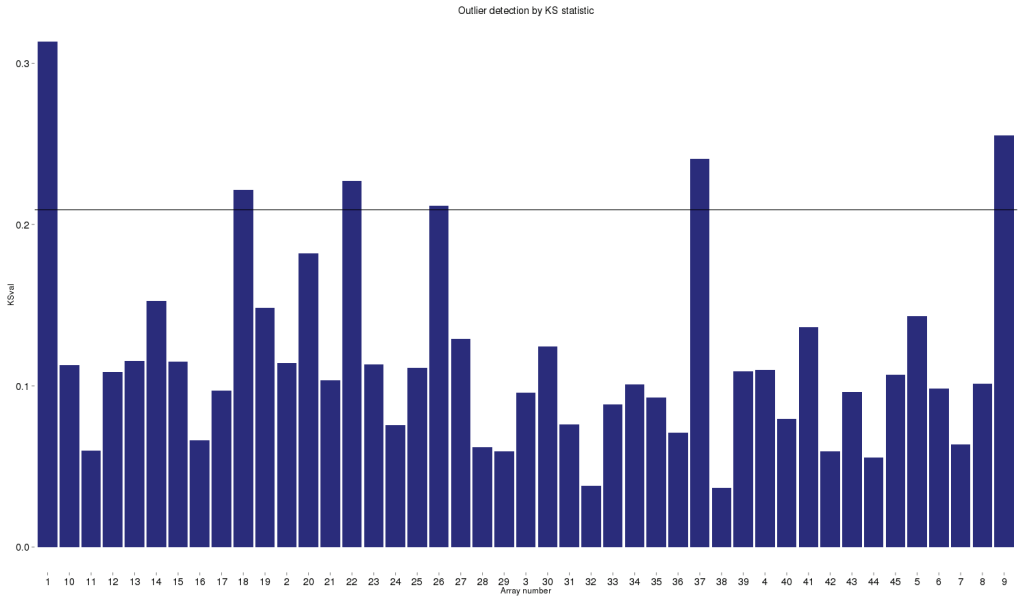


Figure 6 : Bar plot of KS statistics for each array

The horizontal line corresponds to the threshold against which the values of statistic were compared. The threshold represents the *3rd quartile of KS statistics + coef*IQR*. *Coef = 1.5* (default value in R function), *IQR = 3rd quartile – 1st quartile of KS statistic values*. Arrays with a statistic greater than threshold are called outliers. **Arrays 1, 18, 22, 26, 37 and 9 are outliers for this criterion.**

3.2 RLE plots

The Relative Log Expression (RLE) values are computed for each probeset by comparing the expression value on each array to the median expression value for that probeset across all arrays (Figure 7). Since it is assumed that in most experiments only relatively few genes are differentially expressed, the boxes should be similar in range and be centered close to 0. Based on the Kolmogorov-Smirnov statistic across all arrays, only 1 array exceeded the threshold of 0.108: **array 10**.

(for outliers plots see : B1493-Immunomonitoring_markers\Studies\HERV_V3\ET_model_HERVV3\output\QC\outliers_rle)

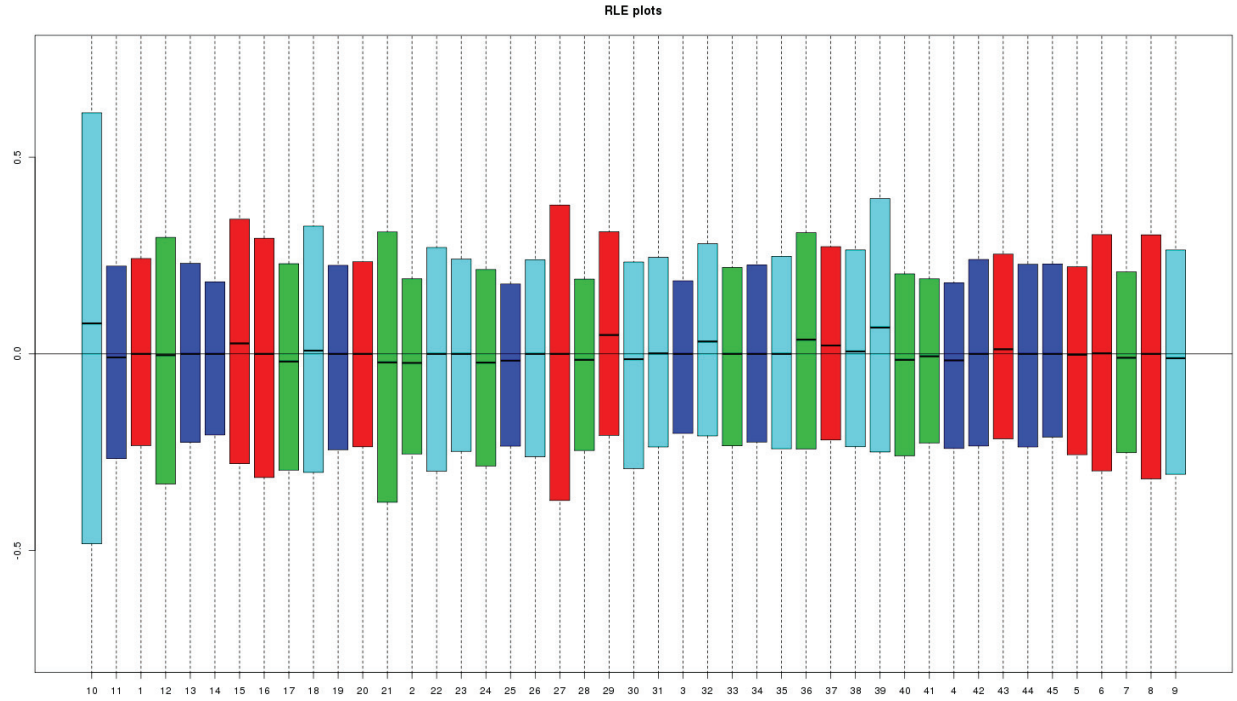


Figure 7 : RLE plots

3.3 Nuse plots

The NUSE plot (Normalized Unscaled Standard Error Plot) is based on standard errors estimated from a linear model. The NUSE plot can help to identify arrays with hybridization problems, a high estimation of standard errors meaning a bad quality. (Figure 8). Outlier detection was performed by computing the 75% quantile of each array's NUSE values. There is 1 outlier (greater than the threshold of 1.136) : **array 10** (for outliers plots see : B1493-Immunomonitoring_markers\Studies\HERV_V3\ET_model_HERVV3\output\QC\outliers_nuse)

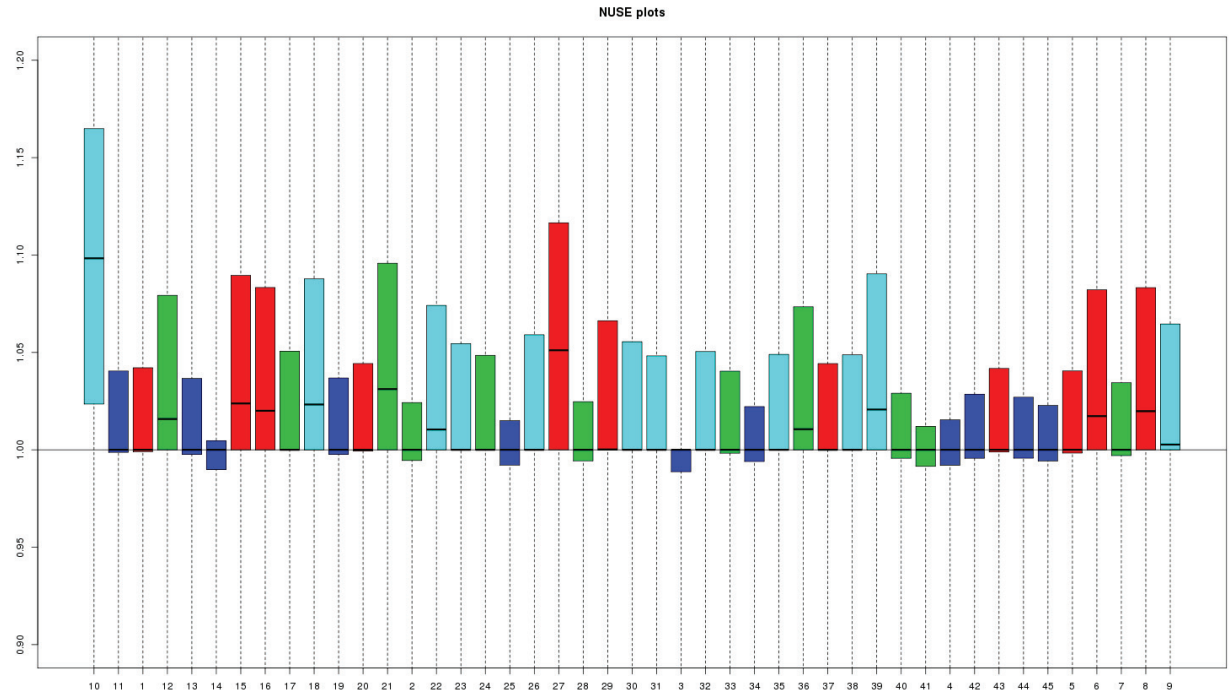


Figure 8 : NUSE plots

3.4 Principal Component Analysis before normalization

Principal Component Analysis is shown in order to detect an eventual outlier array in term of global variability. We will use PCA on normalized data and for batch correction too. We first have to define a threshold for raw data. We decided to determine a threshold based on the distance (d): a threshold of **d = 0.4** on at least one of the two principal axes of the analysis (arbitrary). With this criterion, **array 10** appears as outlier (Figure 9). We will check later if normalization step corrected this.

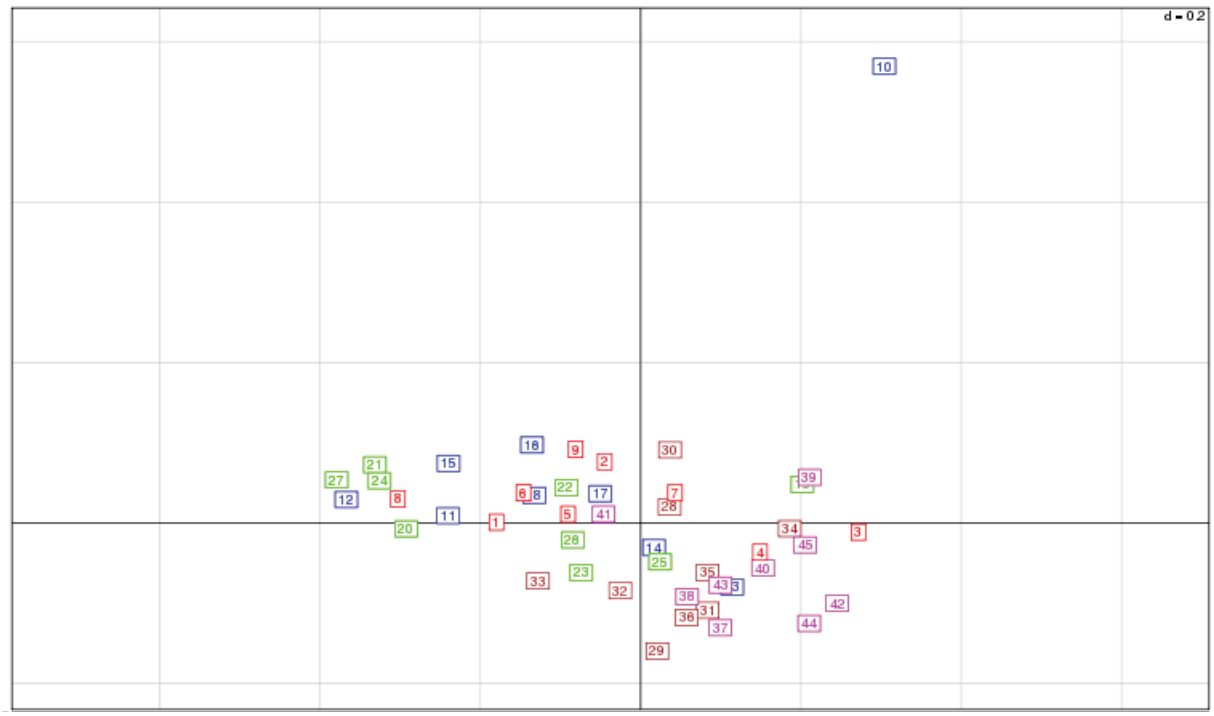


Figure 9: Principal Component Analysis on raw data

To conclude on quality controls of raw data, we detected 6 outliers on intensity distribution criterion, 1 outlier for RLE, NUSE, and PCA. This outlier is array 10. Now we can check quality on normalized data (by RMA). Thus we will be able to verify if the normalization step removed outliers for boxplots intensities and PCA.

4. Quality controls after RMA normalization

4.1 Boxplots

We show the boxplots of intensities after RMA normalization (Figure 10). We can verify if RMA normalization worked and also detect eventual outliers.

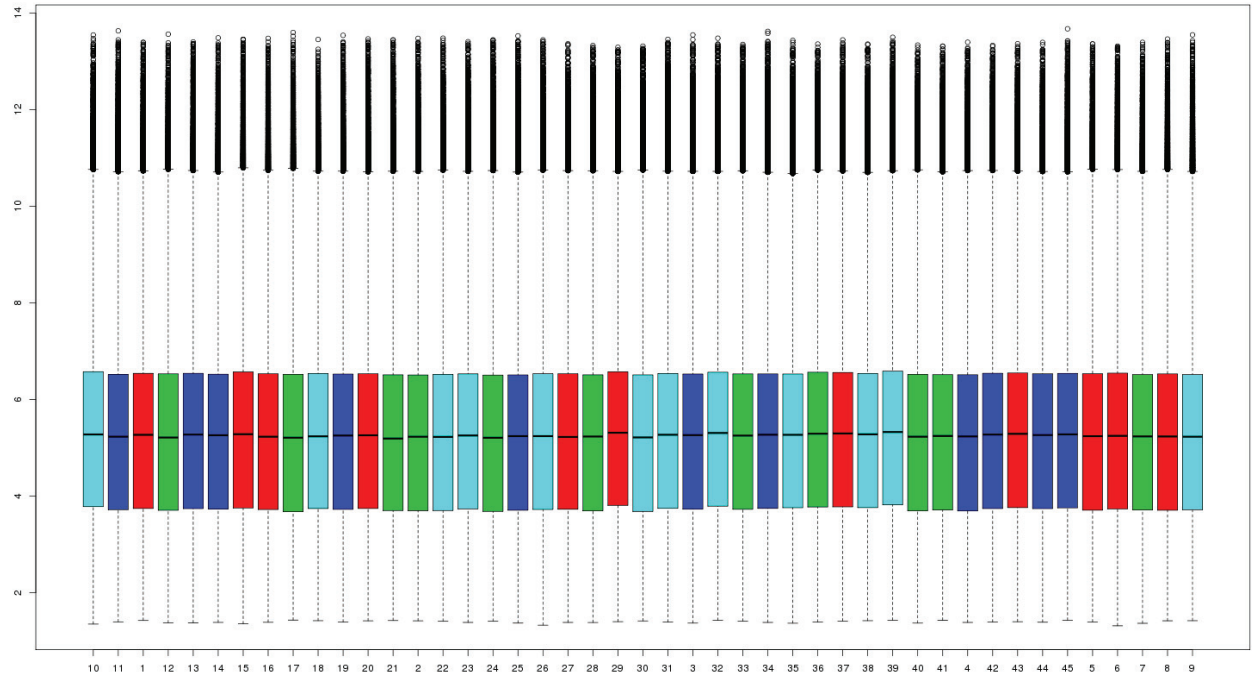


Figure 10 : Boxplots of normalized intensities

We can see that the boxplots are almost identical indicating that the normalization worked well. All outliers detected on raw data are not outliers after the normalization step.

4.2 Correlation between arrays

We show the between array Pearson's correlation on normalized intensities. On the **Figure 11**, more the color is red, less the two arrays are correlated. We defined (arbitrarily) a threshold of 0.85. If the mean correlation for one array is under this threshold, this is an outlier.

Here we can clearly see that array 10 - marked with '*' on first column and last line - is an outlier (mean correlations = 0.828) .

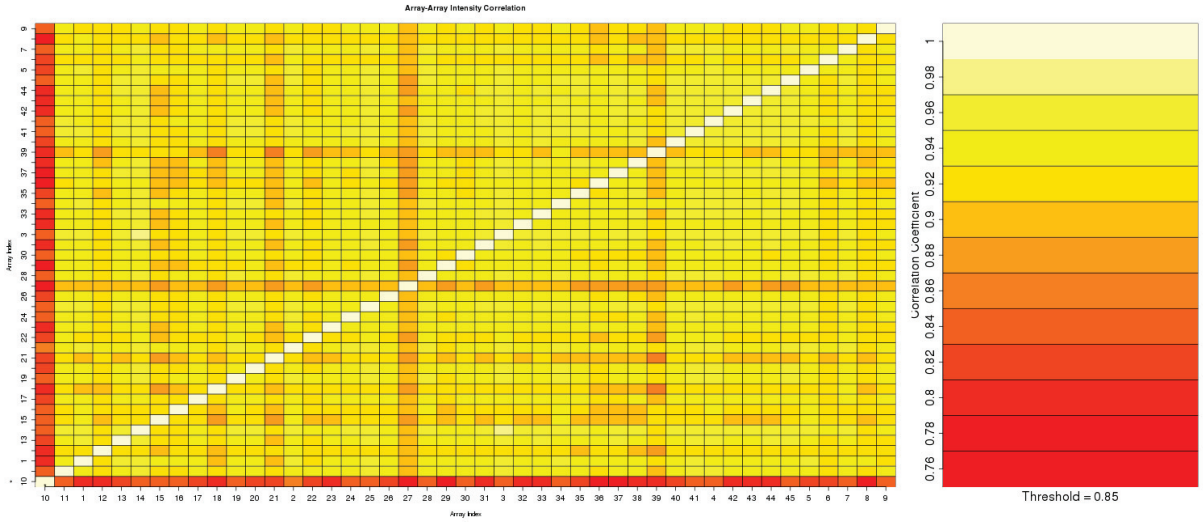


Figure 11 : Pearson's correlation between array

4.3 Principal Component Analysis after normalization

Finally Principal Component Analysis are shown in order to detect an eventual outlier array in term of global variability. We will use PCA for batch correction too. We first have to define a threshold for normalized data. We decided to determine a threshold based on the distance (d): a threshold of **d = 0.1** on at least one of the two principal axes of the analysis (arbitrary). With this criterion, **array 10** still appears as outlier, even after normalization step (Figure 12) .

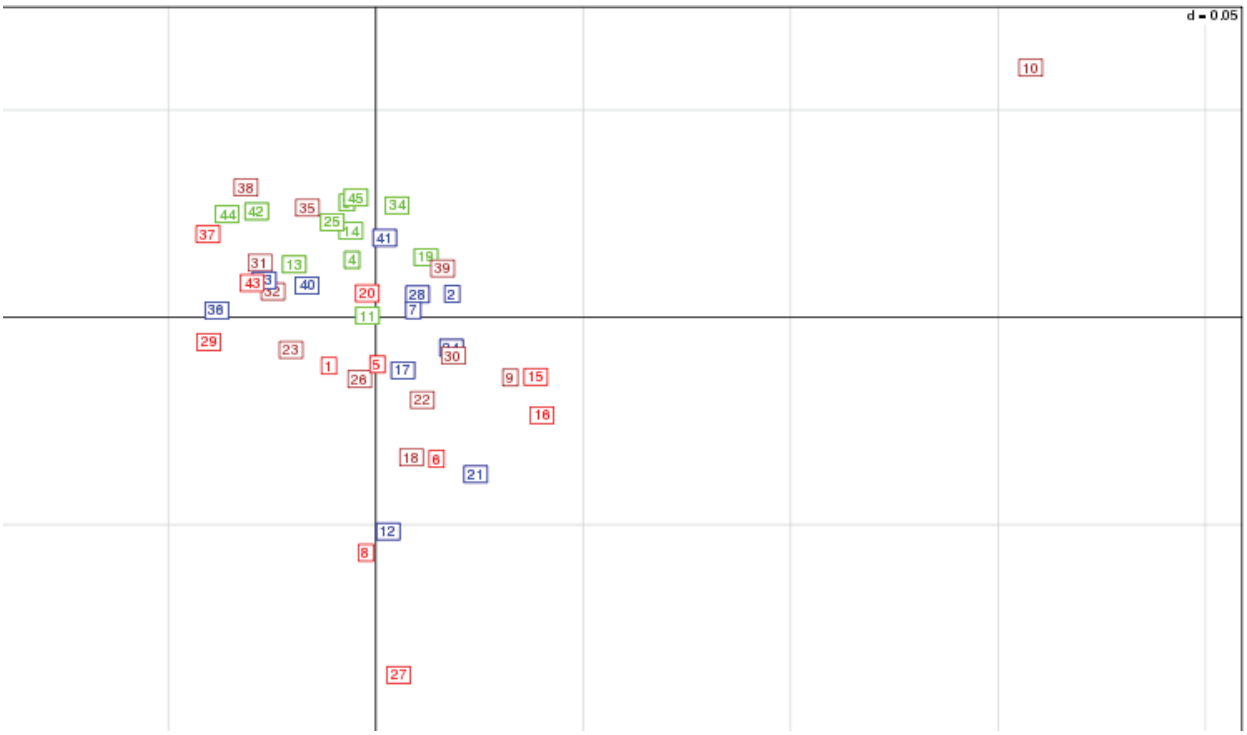


Figure 12: Principal Component Analysis on normalized data

4.4 Decision table

Only arrays with one or more bad criterion are presented in the table. 8 arrays on 45 have at least one bad quality criterion.

Array number	HV	Status	Batch number	RIN	AMPLIF & FRAGM	Border elements	Boxplots RAW	NUSE plot	RLE	Boxplots normalized	Correlation	PCA (2)	Sum	Minimum of 3
1	HV2	NS	1	☺	☺	☺	☹	☺	☺	☺	☺	☺	1	
9	HV2	ET	4	☺	☺	☺	☹	☺	☺	☺	☺	☺	1	
10	HV3	NS	4	☺	☺	☺	☺	☹	☹	☺	☹	☹	4	x
18	HV3	ET	4	☹	☺	☺	☹	☺	☺	☺	☺	☺	2	
22	HV4	LPS	4	☺	☺	☺	☹	☺	☺	☺	☺	☺	1	
26	HV4	ET	4	☺	☺	☺	☹	☺	☺	☺	☺	☺	1	
34	HV6	ET	3	😐	☺	☺	☺	☺	☺	☺	☺	☺	1	
37	HV7	NS	1	☹	☺	☺	☹	☺	☺	☺	☺	☺	2	
39	HV7	NS	4	☹	☺	☺	☺	☺	☺	☺	☺	☺	1	
42	HV7	LPS	3	☹	☺	☺	☺	☺	☺	☺	☺	☺	1	
43	HV7	ET	1	☹	☺	☺	☺	☺	☺	☺	☺	☺	1	
44	HV7	ET	3	☹	☺	☺	☺	☺	☺	☺	☺	☺	1	
45	HV7	ET	3	☹	☺	☺	☺	☺	☺	☺	☺	☺	1	

Table 1: Decision Table

To conclude on quality controls, we decided to remove array 10 from the dataset because it has more than 3 bad quality criteria (4). We kept the others. We also noticed that missing RIN values corresponded to RNA samples passed on the same Agilent chip which

did not work. To avoid another defrost cycle for these RNA samples, the RIN values will be computed later, during another experiment using these same RNA samples.

Batch correction with ComBat

ComBat for COMbining BATches of gene expression microarray data is an algorithm based on parametric and nonparametric empirical Bayes frameworks for adjusting data for batch effects that is robust to outliers in small sample sizes and performs comparable to existing methods for large samples (*Johnson, WE, Rabinovic, A, and Li, C (2007). Adjusting batch effects in microarray expression data using Empirical Bayes methods. Biostatistics **8**(1):118-127.)*

5.1 First run: Batch number

Using PCA, we can visualize and determine if there are non-desirable effects of some variables.

First, using several known variables, we can color in different ways the same PCA.

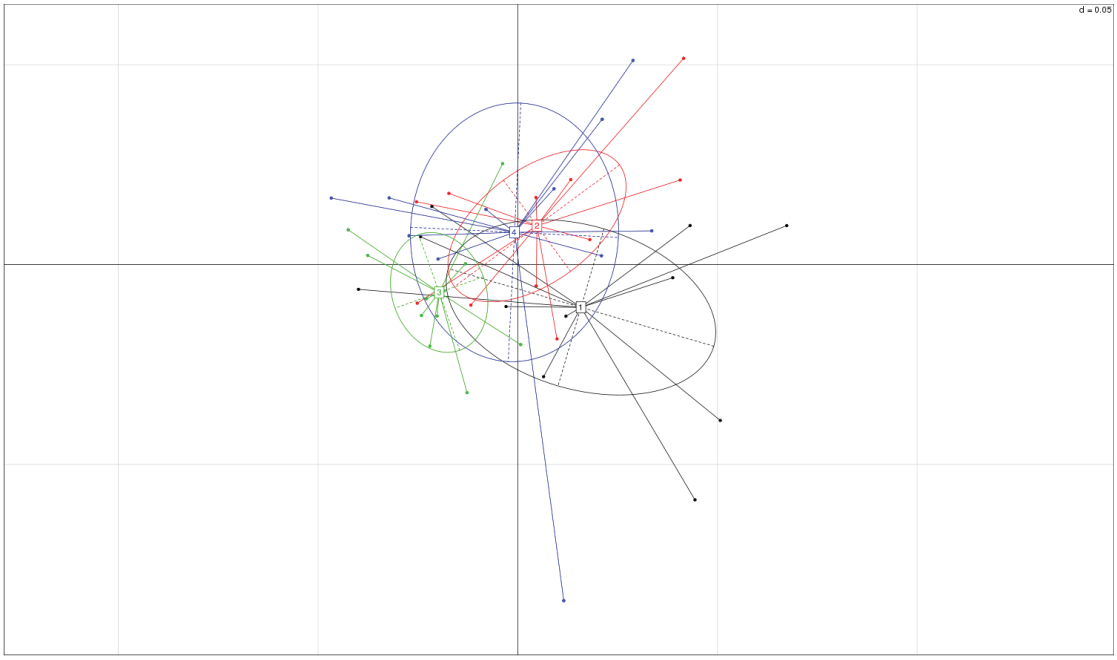


Figure 13 : PCA colored by batch number before correction

On Figure 13, the different centroids correspond to different batches. We can see a difference between these batches, so we decided to apply ComBat algorithm on this variable. After batch correction with ComBat, this is what we obtain (Figure 14):

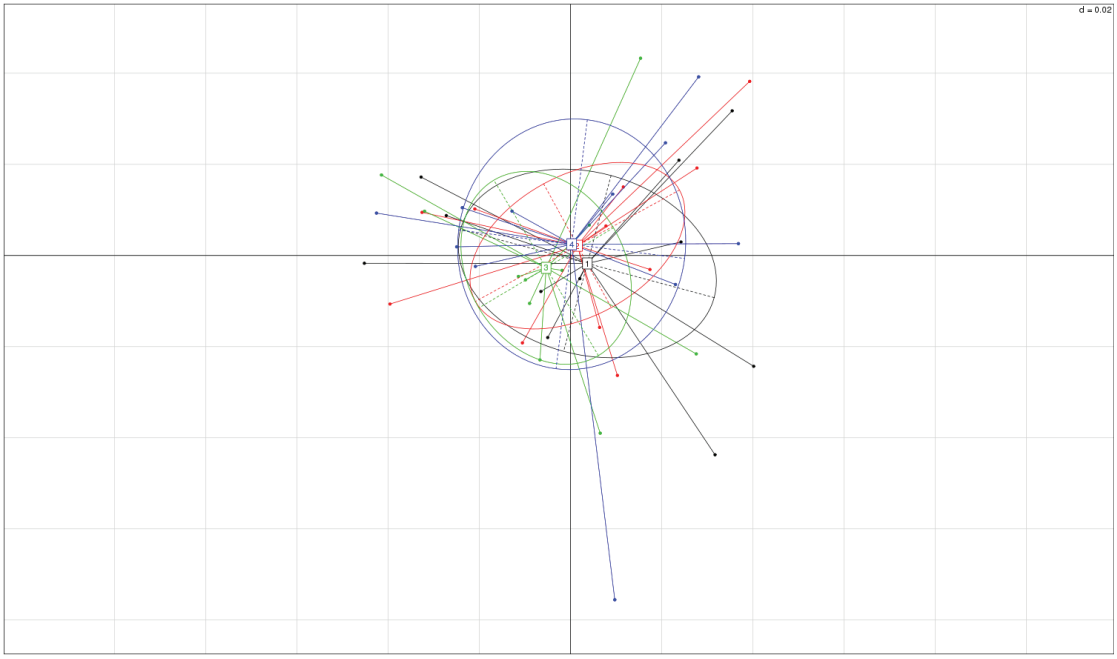


Figure 14 : PCA colored by batch number after correction

NB: The different batch numbers correspond to the different hybridization runs and the different sample preparation dates.

5.2 Second run: Healthy volunteers

Using the same PCA, we colored arrays by healthy volunteer (Figure 15).

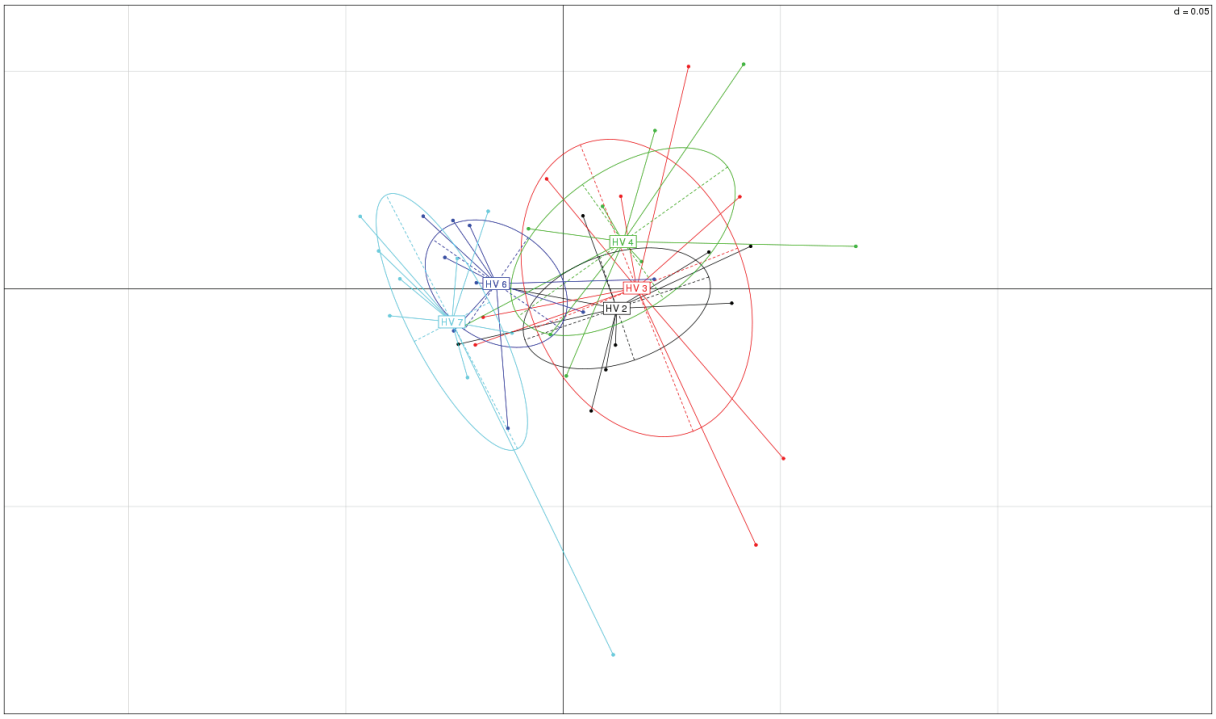


Figure 15 : PCA colored by healthy volunteer before correction

The variability is different depending on the volunteer and we do not want to see the effect of individual variability when we will analyze the dataset. So we decided to apply a second run of ComBat algorithm on the variable healthy volunteer. The resulting PCA is shown on Figure 16.

N.B : Batch correction is usually applied to correct technical bias, not for inter-individual variability. Regarding the inter-individual variability, because we applied different treatments on the same biological samples and because we want to get rid of individual variability, we decided to correct effect of volunteer.

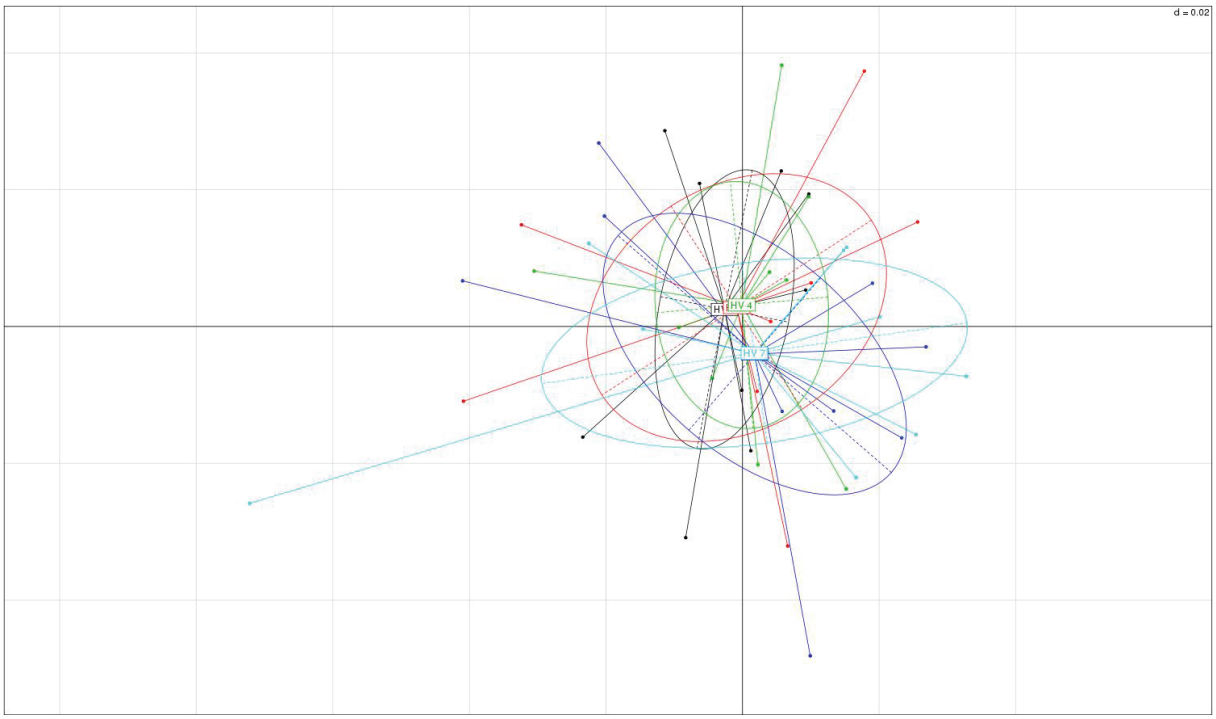


Figure 16 : PCA colored by healthy volunteer after correction

We also checked the pca colored by batch number after the second run of ComBat and we did not see difference with those after the first run.

To conclude, we applied ComBat to remove batch effects on 2 variables (Batch number and healthy volunteer). This will allow us to conclude in further analysis on the variable of interest which is the type of stimulation (NS, LPS, ET).

6.2 ANNEXE 2 : ARTICLE REALISM, ROL ET AL.

THE REANIMATION LOW IMMUNE STATUS MARKERS (REALISM) PROJECT : A PROTOCOL FOR BROAD CHARACTERISATION AND FOLLOW-UP OF INJURY-INDUCED IMMUNOSUPPRESSION IN INTENSIVE CARE UNIT (ICU) CRITICALLY ILL PATIENTS

Rol ML, Venet F, Rimmele T, Moucadel V, Cortez P, Quemeneur L, Gardiner D, Griffiths A, Pachot A, Textoris J, Monneret G; **REALISM study group**.

Publié dans *BMJ Open*, 2017

BMJ Open

The REAnimation Low Immune Status Markers (REALISM) project: a protocol for broad characterisation and follow-up of injury-induced immunosuppression in intensive care unit (ICU) critically ill patients

Mary-Luz Rol,[1,2] Fabienne Venet,[2,3] Thomas Rimmele,[2,4] Virginie Moucadel,[5] Pierre Cortez,[6] Laurence Quemeneur,[7] David Gardiner,[13] Andrew Griffiths,[8] Alexandre Pachot,[2,5] Julien Textoris,[2,4,5] Guillaume Monneret,[2,3] On behalf of the REALISM study group

To cite: Rol M-L, Venet F, Rimmele T, *et al*. The REAnimation Low Immune Status Markers (REALISM) project: a protocol for broad characterisation and follow-up of injury-induced immunosuppression in intensive care unit (ICU) critically ill patients. *BMJ Open* 2017;**7**:e015734. doi:10.1136/bmjopen-2016-015734

► Prepublication history and additional material are available. To view these files please visit the journal online (http://dx.doi.org/ 10.1136/bmjopen-2016-015734).

Received 2 January 2017
Revised 22 February 2017
Accepted 8 March 2017

For numbered affiliations see end of article.

Correspondence to
Dr Julien Textoris; julien.textoris@biomerieux.com

ABSTRACT

Introduction The host response to septic shock is dynamic and complex. A sepsis-induced immunosuppression phase has recently been acknowledged and linked to bad outcomes and increased healthcare costs. Moreover, a marked suppression of the immune response has also been partially described in patients hospitalized in intensive care unit (ICU) for severe trauma or burns. It has been hypothesized that immune monitoring could enable identification of patients who might most benefit from novel, adjunctive immune-stimulating therapies. However, there is currently neither a clear definition for such injury-induced immunosuppression nor a stratification biomarker compatible with clinical constraints.

Methods and analysis We set up a prospective, longitudinal single-centre clinical study to determine the incidence, severity and persistency of innate and adaptive immune alterations in ICU patients. We optimized a workflow to describe and follow the immunoinflammatory status of 550 patients (septic shock, severe trauma/burn and major surgery) during the first 2 months after their initial injury. On each time point, two immune functional tests will be performed to determine whole-blood TNF-α production in response to *ex vivo* lipopolysaccharide stimulation and the T lymphocyte proliferation in response to phytohaemagglutinin. In addition, a complete immunophenotyping using flow cytometry including monocyte HLA-DR expression and lymphocyte subsets will be obtained. New markers (ie, levels of expression of host mRNA and viral reactivation) will be also evaluated. Reference intervals will be determined from a cohort of 150 age-matched healthy volunteers. This clinical study will provide, for the first time, data describing the immune status of severe ICU patients over time.

Ethics and dissemination Ethical approval has been obtained from the institutional review board (no 69HCL15_0379) and the French National Security agency for drugs and health-related products. Results will be disseminated through presentations at scientific meetings and publications in peer-reviewed journals.

Trial registration number Clinicaltrials.gov Registration number: NCT02638779. Pre-results.

Strengths and limitations of this study

- This is the first prospective study to provide a broad immune status characterisation in a large cohort of intensive care unit (ICU) patients.
- There is a mid-term assessment (D60) of the immune status in ICU patients, which has never been done before.
- Long-term follow-up will not be addressed here and should be examined in future studies.
- New biomarkers of the immune status will be assessed in comparison to standardised tools and immune functional assays.
- Whether such biomarkers would permit to stratify patients for immunomodulatory treatments should be addressed in future studies.
- The role of host genomics, microbiota as well as checkpoint inhibitor expressions will not be assessed in this study.

INTRODUCTION

Sepsis is a major health problem and the main aetiology for intensive care unit (ICU) admissions.[1 2] Its incidence is increasing over the years due to several factors, including a better awareness and an ageing population.[3] Hospital admissions for sepsis have thus overtaken those for stroke and myocardial infarction.[4] Despite advances on its management, mortality of sepsis has remained stable over the last 20 years, reaching 30%–40% in

case of septic shock, the most severe form, and it is the leading cause of death in ICU.

Sepsis is a severe infection, defined as a 'life-threatening organ dysfunction caused by a dysregulated host response to infection'.[5] Besides circulatory and metabolic abnormalities, the multifaceted host response to the invading pathogen is amplified by comorbid conditions.[6 7] It is now acknowledged that the pro-inflammatory response, which can lead to organ failure, comes with a compensatory anti-inflammatory response. Recovery occurs when inflammation resolves quickly. However, in numerous patients, the anti-inflammatory response lingers on and leads to an immunosuppression state, associated with secondary infections, and increased morbidity and mortality.[8] This sepsis-induced immunosuppression could explain the failure of several previous clinical trials and support new innovative trials testing immune adjuvant drugs in septic shock.[9]

Therefore, several studies and case reports now support the rational of boosting the immune system, in order to avoid the occurrence of healthcare-associated infection and therefore reduce the associated morbidity.[10 11] However, to avoid reproducing the errors from the past, such innovative treatments should be administered only to those individuals identified as immunosuppressed.[11] Some studies have already demonstrated that the concept of biomarker-guided therapeutic stratification can lead to clinical improvements.[12]

A marked immunosuppression has been partially described in other patients admitted to the ICU for severe trauma/burns and other major surgeries.[13–16] In these 'sterile' injuries, signs of injury-induced immune alterations have also been associated with increased susceptibility to secondary infections and mortality.

Given the complexity and heterogeneity of ICU patients, it is unlikely that any single biomarker will be sufficient to describe and diagnose injury-induced immunosuppression. On the contrary, a panel of validated biomarkers may bring enough information to accomplish such complex endeavour.

Rationale of the study

From a clinical perspective, no specific clinical signs or symptoms are associated with a state of altered immune response to allow prospective identification of at risk patients. Further, the outcomes of sustained immunosuppression are best defined by clinical relevant endpoints such as the occurrence of opportunistic and secondary infections. However, waiting for such a healthcare-associated infection to occur does not facilitate implementation of preventive strategies. Thus, diagnosis will rely on biomarkers.

From a biological perspective, sepsis-induced immunosuppression may be best identified by immune *functional* assays (such as cytokine release or lymphocyte proliferations after *ex vivo* stimulation) or by cell count parameters (such as number of lymphocytes or level of expression of mHLA-DR) but both approaches present drawbacks. Indeed, such functional assays are not suitable to stratify patients in a prospective interventional clinical trial due to (1) the long time to results (up to 5 days for lymphocytes proliferation) and (2) poor reproducibility due to standardisation issues and cumbersome technique. Due to such complexity, these reference tests are rarely performed in clinical studies evaluating biomarkers associated with deleterious outcomes in ICU.

On the other hand, HLA-DR expression on monocytes is currently the best biomarker available for such a routine use,[17] and it is being employed for patient stratification in a large multicentre interventional trial assessing the administration of GM-CSF in patients with septic shock.[18] However, its measurement requires flow cytometry analysis within 4 hours of blood sampling which may not be available in all centres, making interlaboratory standardisation challenging.

As a consequence of the previously discussed challenges, numerous biomarkers proposed to monitor injury-induced immune alterations have yet to be compared with these *reference* assays.

Hypothesis

Although several studies have shown an association between markers related to the immune system (eg, HLA-DR) and the occurrence of healthcare-associated infections in septic patients,[14 15 19] we still do not have a clear and operational definition of the immune deficiency that occurs in severely injured ICU patients. Precise description of injury-induced immunosuppression incidence and its characteristics are lacking. In the REALISM (REAnimation Low Immune Status Markers) project, we propose to broadly assess immune parameters over time and to correlate these findings with clinical epidemiological data and outcomes in order to identify and define immunosuppression in ICU patients in terms of both magnitude and time duration.

To this aim, we have established two standardised functional immune assays (whole-blood TNF-α release after *ex vivo* stimulation with LPS (lipopolysaccharides)[20] and lymphocyte proliferation in response to *ex vivo* stimulation with PHA (phytohaemagglutinin).[21] We propose to define the status of immunosuppression on the basis of an abnormal result (values outside the reference intervals) obtained in at least one of the two 'reference' tests.

The REALISM project aims to provide a validated operational definition of injury-induced immunosuppression predicting clinically relevant outcomes. This will facilitate development of new tools and biomarkers with the goal of introducing diagnosis of immune suppression into routine clinical practice and will allow patient stratification for the evaluation of new individual immunotherapies.

It may also enable the identification of new targets and the development of new innovative therapeutics to treat ICU patients and prevent opportunistic infections in the future

Primary aim

The primary objective of the study is to determine the incidence of injury-induced immunosuppression in ICU patients, during the first 2 months after injury.

Secondary aims

The secondary objectives of the study are as follows:

- To describe the occurrence of immunosuppression, its depth and impact on innate and adaptive immune responses and its evolution during the first 2 months after injury.
- To assess the strength of the proposed definition, in particular, by evaluating its association with secondary infections and mortality.
- To assess the accuracy of new biomarkers and immune functional assays to diagnose immunosuppression.

These new biomarkers / immune functional assays could therefore replace assays such as the T cell proliferation assay, the current protocol of which is not suited to the routine management of ICU patients. We therefore expect to provide data to validate simpler diagnostic tools to determine and follow the immune status in hospitalised patients.

METHODS AND ANALYSIS

REALISM is a prospective longitudinal, single-centre observational study, conducted in the anaesthesiology and intensive care department at the Edouard Herriot Hospital (University Hospital, Lyon, France; capacity of approximately 1000 beds).

Study population

REALISM will include healthy volunteers (n=150) and patients at risk of injury-induced immunosuppression: (1) septic shock patients (n=160), (2) severe trauma patients (n=180), (3) severe burns patients (n=30) and (4) patients admitted to the ICU after major surgery (n=180).

Septic shock inclusion criteria follow the current definition[5] and require a state of shock defined by vasopressors administration and plasma lactate level above 2 mmol/L (18 mg/dL). An infection must be suspected, and microbiological sampling should have been performed, along with the administration of antimicrobials. Only primary septic shock will be considered (vasopressors should have been started within the first 48 hours after ICU admission).[5]

Patients with severe trauma, defined by an ISS (injury severity score, Baker *et al*, 1974) >15,[22] will be included in the study. As we hypothesised that the depth of

Box 1 Inclusion and exclusion criteria for patients

Inclusion criteria
Male or female aged over 18 years
Patient hospitalised for:
Septic shock, defined by:
Infection site suspected, and microbiological analysis sampling carried out
Vasopressor therapy needed to elevate mean arterial pressure ≥65 mm Hg and lactate >2 mmol/L (18 mg/dL) despite adequate fluid resuscitation[27]
Norepinephrine >0.20 µg/kg/min for at least 2 hours
Norepinephrine started within 48 hours after intensive care unit (ICU) admission
Serious trauma, defined by:
Patient admitted directly to the recruiting ICU
ISS, Baker *et al*, 1974 >15[22]
Severe burns, defined by:
Total burned surface area >30%
Major surgery, defined by:
Surgery set for one of the following indications: (1) eso-gastrectomy, (2) Bricker's bladder resection (total bladder resection with reconstruction from small bowel), (3) cephalic pancreaticoduodenectomy (Whipple's procedure) and (4) abdominal aortic aneurysm surgery by laparotomy. Categories 1–3 concern management of solid tumours, while category 4 concerns non-cancerous pathologies
Induction of anaesthesia before 11:00 (to permit same-day processing of all samples)
Exclusion criteria
Patient with severe neutropenia (neutrophil count <0.5 G/L)
Patients receiving immunosuppressive therapy
Corticosteroids (intravenously or per os).
Use of therapeutic antibodies
Onco-haematological disease (eg, lymphoma, leukaemia…) under treatment or treated within 5 years before inclusion
End of chemotherapy within the 6 months prior to inclusion date
Patient with innate or acquired immune deficiency (eg, severe combined immunodeficiency, HIV or AIDS, any stage)
Patients with a 'do not resuscitate order' or a 'withdraw of care' decision, at time of inclusion
Patient whose anticipated duration of hospitalisation in the ICU is estimated at less than 48 hours
Participation in any interventional study
Extra-corporeal circulation in the month preceding inclusion in the case of cardiac surgery
Pregnant or breastfeeding women
Patient with no social security insurance, with restricted liberty or under legal protection

Box 2 Inclusion and exclusion criteria for healthy volunteers

Inclusion criteria
- Male or female aged over 18 years
- Normal clinical examination
- Signed informed consent form
- Person with social security insurance

Exclusion criteria
- Person with an infectious syndrome during the last 90 days
- Extreme physical stress within the last week

Person having received within the last 90 days, a treatment based on:
- Antivirals
- Antibiotics
- Antiparasitics
- Antifungals
- Person having received within the last 15 days, a treatment based on non-steroidal anti-inflammatory drugs

Person having received within the last 24 months, a treatment based on:
- Immunosuppressive therapy
- Corticosteroids (intravenously or per os)
- Therapeutic antibodies
- Chemotherapy

History of:
- Innate or acquired immune deficiency
- Haematological disease
- Solid tumour
- Severe chronic disease
- Surgery or hospitalisation within the last 2 years
- Pregnancy within the last year
- Participation to a phase I clinical assay during the last year
- Participation to a phase I clinical assay during the last year
- Pregnant or breastfeeding women
- Person with restricted liberty or under legal protection

immunosuppression might be related to severity, we will limit the group of patients between ISS[15–17 19–26] values to 90 patients to ensure that, at least, 50% of the cohort includes patients with an ISS >25. Severe burn patients will be selected for inclusion based on a total burn surface area over 30%.

Surgical patients will be screened according to the planned surgical procedure. This study will include patients undergoing: (1) eso-gastrectomy, (2) Bricker's bladder resection (total bladder resection with reconstruction from small bowel), (3) cephalic pancreaticoduodenectomy (Whipple's procedure) and (4) abdominal aortic aneurysm surgery by laparotomy.

Exclusion criteria are mainly related to factors that might impact the immune status and bias the results such as the following: severe neutropenia (neutrophil count $<0.5\,10^9$/L), administration of immunosuppressive therapy, corticosteroids (IV or oral administration), use of therapeutic antibodies (such as anti-TNF-α), onco-haematological disease (eg, lymphoma, leukaemia) under treatment or treated within 5 years before inclusion and end of chemotherapy within the 6 months prior to inclusion date. Patients with congenital/hereditary or acquired immune deficiency (eg, severe combined immunodeficiency, HIV or AIDS, at any stage) and patients that have received extracorporeal circulation in the month preceding inclusion will be excluded as well.

Considering the possible influence of gender bias on measured parameters, we will recruit healthy donors from both genders, following the age and gender distribution of the French population.

Complete lists of the inclusion and exclusion criteria for patients and healthy volunteers are presented in box 1 and box 2, respectively.

Table 1 Age and gender distribution for the reference group

Age range	Male	Female
(19–30)	14	14
(30–50)	25	25
(50–65)	18	19
(65–100)	15	20
Total	72	78

Sampling schedule

Samples and clinical data will be collected 3–4 times within the first week (early time points) with the aim to evaluate the modulation of the immune status early after injury. Samples will be collected at day 1 (the morning following injury), at day 2 (for the severe trauma group) and at day 3/4 and day 5/7 (table 1). Samples will also be collected before surgery, at day 0, as surgical patients are the only group for which sampling can be performed before injury. Additional samples will be collected during late time points to evaluate the recovery of the immune status, at day 14 (between day 13 and 18), day 28 (between day 26 and 36) and day 60 (between day 52 and 68), depending on patient availability and technical constraints (figure 1). Total volume of sampling will be 30 mL at each time point.

Definition of immunosuppression

The REALISM project will monitor the immune function of the patients and healthy volunteers using two standardised immune functional tests: one reference test to evaluate the innate immune response (whole-blood production of TNF-α in response to *ex vivo* stimulation by LPS) and a second reference test for the adaptive immune response (the lymphocyte proliferation in response to *ex vivo* T cell stimulation with PHA). Immunosuppression will be defined in comparison to the values as obtained in a group of healthy volunteers for the two reference tests using the following methodology. First, reference intervals will be derived from the independent set of healthy volunteers. Second, immunosuppression will be defined in a patient when an abnormal result (value outside the reference intervals)

is obtained in at least one of the two 'reference' tests over at least two consecutive time points

Definition of secondary infection

During the ICU stay, patients will be screened daily for exposure to invasive devices (intubation, indwelling urinary catheter and central venous line) and occurrence of secondary infection. Information referent to infections will be collected, reviewed and validated by a dedicated adjudication committee, composed of three physicians not involved in the recruitment of the patients with confirmation of secondary infection made according to the definitions used by the European Centre for Disease Prevention and Control[24] and the Infectious Diseases Society of America.

Immune functional assays

Innate immune response: TNF-α release after LPS whole-blood stimulation

Innate immune response will be evaluated by measuring the production of TNF-α in response to *ex vivo* stimulation of whole blood by LPS.[20] The stimulation will be performed through the use of standardised TruCulture tubes from MYRIAD RBM (MYRIAD RBM, Austin, USA) (the concentration, quality and activity of the LPS is guaranteed by the manufacturer MYRIAD RBM).[20] The tubes contain the medium alone (Null) or the medium with LPS 100 ng/mL (LPS from *Escherichia coli* O55:B5) (LPS-R; Null-R; MYRIAD RBM). The blood samples will be collected on heparin and transported to the laboratory where 1 mL of heparinized blood will be transferred to each TruCulture tube and incubated for 24 hours at 37°C. Following incubation, the supernatant (medium+plasma) will be collected using a separation valve (according to manufacturer instructions) and stored at −80°C until batch quantification of TNF-α by ELISA (BE55001; BL International-Tecan, Männedorf, Switzerland).

Adaptive immune response: T lymphocyte proliferation after *ex vivo* peripheral blood mononuclear cells mitogenic stimulation

Adaptive immune response will be assessed by measuring T lymphocyte proliferation in response to *ex vivo* stimulation with a mitogen.[21] Briefly, peripheral blood mononuclear cells isolated by Ficoll density gradient centrifugation (U-04; Eurobio, Les Ulis, France) will be stimulated with PHA at 4 µg/mL (HA16; Remel, Lenexa, USA), at 37°C for 72 hours. Following incubation, the cells will be harvested and cell's proliferation will be determined by the incorporation of EdU (5-ethynyl-2'-deoxyuridine, 10 µM for 2 hours) in T cells using the commercial kit Click-It EdU AF488 flow kit (C10420; Life Technologies, Carlsbad, California, USA). Cell proliferation is measured as the percentage of EdU-positive T cells (gated as CD3+ using a CD3-APC staining) using flow cytometry.[21]

Cellular immunophenotyping

Complete blood cell count report from the haematology laboratory will be collected on each time point, this information will be compared with our cell counts results by flow cytometry. Beside phenotypic immune cells, characterisation and cell counting will be completed by flow cytometry and we will count the number of B lymphocytes (CD45+, CD3−, CD19+), T lymphocytes, CD4+ (CD45+, CD3+, CD8−, CD4+) and CD8+ (CD45+, CD3+, CD8+, CD4−), NK cells (CD45+, CD3−, CD56+), regulatory T lymphocytes (gated on T CD4+, CD25high, CD127low) and mature (CD10High, CD16High, CD14−, CRTH2−) and immature mature (CD10dim, CD16dim, CD14−, CRTH2−) polymorphonuclear cells, as previously published.[25 26] In addition, the number of HLA-DR molecules per monocyte will be determined using the BD quantibrite standardised method (HLA-DR:340827; QuantiBRITE:340495; Becton Dickenson, New Jersey, USA).[27] It is well known that the flow cytometry is highly sensitive to variation between laboratories and instruments; therefore, a validation with the routine hospital immunology laboratory was performed to guarantee that all the protocols are reproducible and standardised. All procedures generated results with less than 20% of variation when compared with reference protocols.

Biobanking

This study will provide the opportunity to establish four different types of biobanks to preserve the material collected, enabling exploration of innovative biomarkers:

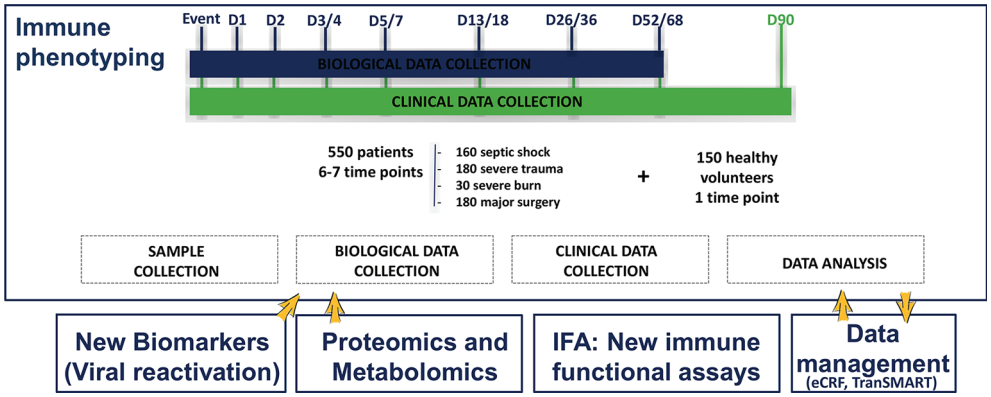


Figure 1 Schematic design of the REALISM project illustrating the type of patients included in the study, the various time points and major planned analysis. REALISM, REAnimation Low Immune Status Markers.

(1) TruCulture plasma biobank from whole blood stimulated with LPS, SEB *(Staphylococcus aureus* enterotoxin B) or not stimulated, to study cytokines release; (2) EDTA plasma biobank to study viral reactivation markers and soluble host biomarkers; (3) heparin plasma biobank for metabolomics/proteomics soluble host biomarkers studies and (4) RNA biobank to study new transcriptomic host biomarkers (RNA will be extracted from whole blood collected in PAXgene tubes).

Innovative immune functional assays and exploration of new biomarkers

Regarding the immune functional tests, other stimulants (eg, SEB) and read-outs (eg, interleukin 2, interferon gamma) will be tested using the TruCulture tubes. The cytokine production levels in the supernatants of the functional assays will be quantified using commercial IVD or RUO assays. Finally, a metabolomics and proteomics study will be performed using frozen (heparin) plasma. Biomarkers potentially associated to immune deficiency will be identified by liquid chromatography–mass spectrometry on high-resolution mass spectrometry and ^{1}H nuclear magnetic resonance, after polar and non-polar samples extraction.

Sample size and data analysis plan

Population sizing

The number of healthy volunteers required to determine the reference intervals for the two immune reference tests was defined according to the methodology recommended by the Clinical and Laboratory Standards Institute C28-A3 guidelines.[28] The minimal number of subjects recommended being 120, after exclusion of aberrant values (CI of 90%), we decided to include 150 healthy volunteers to take into account exclusions related to technical reasons, aberrant values or consent withdrawal.

For this reference population, the age range of healthy volunteers group has been carefully calculated to include the expected age range and gender distribution from ICU patients in France (table 1).

The main objective being descriptive, the computation of the sample size was based on secondary objectives, especially for (1) the analysis of the occurrence of immunosuppression, its depth and impact on innate and adaptive immune responses (Cohen's d is 0.55) and (2) the correlation between new biomarkers and immune functional assays to diagnose immunosuppression (r>0.4). A Student's t-test was used to approximate the number of patients needed and a minimum of 150 patients per group was required for a standardised Cohen's d effect=0.55, if we get the recommended number of healthy volunteers of 120. It was therefore decided to include 160 septic shock patients, 180 severe trauma patients and 180 patients with a major surgery, to overcome secondary exclusions for technical causes or consent withdrawal. The severe burn patients group is an ancillary group that was arbitrary fixed at 30 subjects in order to collect data with the intent to inform a dedicated study on this population in the future.

Statistical analysis

First, the percentage of patients meeting the definition of injury-induced immunosuppression will be computed in each patients group to answer the main objective. Second, the occurrence of immunosuppression will be further described. The proportion of patients with at least one abnormal test will be computed for both immune reference tests and each patients group. The correlation between the two reference tests will be established from a Spearman correlation test. A mixed model will be constructed to describe the extent of the changes in the innate and adaptive measures over time, taking groups and time points into account. Third, a comparison of each biomarker or new functional tests with the two reference tests will be performed using a Spearman correlation test. For correlated biomarkers or functional tests, the performance for prediction of secondary infection will be estimated from a receiver operating characteristic curve. A Fine & Grey predictive model will be constructed[29] for the biomarkers harbouring the best areas under curve, taking into account the competing risk of mortality. Finally, multiple imputations will be taken into consideration in the case of a relevant amount of missing values.

ETHICS AND DISSEMINATION

Ethics approval

The protocol, information documents and consent forms received approval by the local institutional review board (Comité de Protection des Personnes Sud-Est II, Bron, France) and the French National Security agency for drugs and health-related products (Approval code: 69HCL15_0379, 30November 2015). An amendment has been filled to extend sampling time points over the first week and add the metabolomics and proteomics study. This amendment has been approved on the 22July 2016 (protocol version 3). This study complies with the Declaration of Helsinki, principles of Good Clinical Practice and the French personal data protection act.

Informed consent

The free and informed consent of each patient and healthy volunteer will be obtained following a complete and faithful information, in comprehensive words, of the objectives, the proceedings and the constrains of the study, the right to refuse the enrollment or the possibility to withdraw at any time, when he/she is in capacity to understand. The patient (or next of kin) will also be informed of (1) the existence of processing system for data concerning them, (2) Their right to access and rectify these data (accessible through the physician of their choice) and (3) the possibility of the use of remaining biological material and associated data stored following the end of the study and their possible transfer to another academic or private party. This information is part of the written notice and the informed consent.

If the patient is not in capacity to understand and/or express his/her consent, the informed consent will be

Table 2 Clinical and biological data collection planning

	D0*	D1 (D2†)	D2‡	D3/4	D5/7	D13/18	D26/36	D52/68	D90
Inclusion/exclusion criteria		x§							
Consent form		x§							
Demography		x§							
Weight		x§							
Size		x§							
Description of hospital stay		x§							
IGS II score		x§							
McCabe score		x§							
CHARLSON score		x§							
Documentation of the septic shock, surgery, burn or trauma		x§							
SOFA score	x	x	x	x	X				
Treatments against infections			Steadily			x	x	x	x
Therapeutic management			Steadily						
Exposition to medical devices			Steadily						
Surveillance of healthcare associated infections			Steadily			x¶	x¶	x¶	x¶
Concomitant events			Steadily			x	x	x	x
Vital status**						x	x	x	x
Life quality (EQ5D)									x
Biology									
PAXgene tube sampling	x	x	x	x	X	x	x	x	
EDTA tubes sampling	x	x	x	x	X	x	x	x	
Heparin tubes sampling	x	x	x	x	X	x	x	x	
Haematology	x	x	x	x	X	x	x	x	
Lactate	x††	x††	x††	x††	x††				
pH	x††	x††	x††	x††	x††				
Liver results (ASAT, ALAT, PAL)	x††	x††	x††	x††	x††				
Procalcitonin	x††	x††	x††	x††	x††	x††	x††	x††	x††
Serology (CMV, HSV1)		X§							

*Only for patients of the surgery group
†For the septic shock and burn patients: The enrollment at D2 will be accepted if D1 is not available
‡Only for patients of the trauma group
§Evaluation on day 0 for patients of the surgery group (not repeated on day 1)
¶Only if related to a new hospitalisation
**Also at ICU release and Hospital release
††If available

obtained from a next of kin. In the event that only the informed consent of a third party has been sought at the time of inclusion, the patients should be informed as soon as possible of their participation in this study and be asked to give their own consent to continue the study.

If the next of kin is not present and not available by phone, the patient may be included in emergency situation. The investigator will be required to record all steps for calling the next of kin in the medical record (contact attempts with date, time and phone number) and justify patient inclusion in medical emergencies in accordance with French legislation. The written consent of the next of kin and the patient should be obtained as soon as the person is available and as soon as the patient's clinical condition allows. The consent form contains the possibility to refuse the storage of samples after the end of the study.

Safety of participants

This study includes no serious foreseeable risk to the health of the persons involved. The only potential risk is related to blood sample collection (maximum 192 mL collected over all time points — 2 months). However, this aspect of nursing is part of daily practice. Blood samples will be taken under the same conditions of safety as currently used for common diagnostic tests.

Study management

The study is managed by BIOASTER and a dedicated team composed from members of all the consortium partners. The promoter of the study is the Hospices Civils de Lyon. The principal investigator is Dr Thomas Rimmelé.

Data management

Clinical data

For each patient, an electronic case report form including socio-demographic, clinical and para-clinical information will be completed by clinical research assistants (table 2): adescription of the hospital stay, the documentation on the type of injury (surgery, burn, trauma or septic shock) and the severity as defined by the ASA classification, SOFA score[30] and SAPSII score.[30] In addition, we will collect routine laboratory results about the CMV, HSV1 serology and complete blood count. Moreover, we will document if there is any specific treatments administered to the patient, such as antibiotics, exposure to invasive medical devices and secondary infections. All data will be transferred to a TranSMART[30] database following curation for data exploration and analysis.

Duration of the study

The study is planned to run for 30 months, starting December 2015. The expected end date for recruitment is June 2018. Some biomarkers will be quantified by batch analysis, at the end of the study. Primary data analysis is expected to be completed with subsequent dissemination of results by December 2018.

Author affiliations
[1]BIOASTER Technology Research Institute, Lyon, France
[2]EA7426 "Pathophysiology of Injury-induced immunosuppression", Université Claude Bernard Lyon 1 - Hospices Civils de Lyon - bioMérieux, Lyon, France
[3]Immunology Laboratory, Hospices Civils de Lyon - Université Claude Bernard Lyon 1, Lyon, France
[4]Anesthesiology and Critical Care Medicine, Hospices Civils de Lyon - Université Claude Bernard Lyon 1, Lyon, France
[5]Medical Diagnostic Discovery Department (MD3), bioMérieux, Marcy-l'Étoile, France
[6]R&D, Sanofi Aventis, Chilly-Mazarin, France
[7]Sanofi-Pasteur SA, Lyon, France
[8]ESPCI Paris, PSL Research University, Paris, France
[13]GlaxoSmithKline, Collegeville, PA, USA

Correction notice This paper has been amended since it was published Online First. Owing to a scripting error, some of the publisher names in the references were replaced with 'BMJ Publishing Group'. This only affected the full text version, not the PDF. We have since corrected these errors and the correct publishers have been inserted into the references.

Acknowledgements The project is funded by a consortium: bioMérieux, SANOFI, GlaxoSmithKline, Ecole Supérieure de Physique Chimie Industrielles de la ville the Paris – PSL Research University, the University Hospital Hospices Civils de Lyon and the microbiology technological institute BIOASTER. The project is financially supported in part by public funding through BIOASTER and Hospices Civils de Lyon. The project will be audited annually by the French National Research Agency ("Investissement d'Avenir" program; grant no ANR7107AIRT703).

Collaborators For Hospices Civils de Lyon: Asma BEN AMOR, André BOIBIEUX, Julien DAVIDSON, Laure FAYOLLE-PIVOT, Charline GENIN, Arnaud GREGOIRE, Alain LEPAPE, Anne Claire LUKASZEWICZ, Guillaume MARCOTTE, Delphine MAUCORT-BOULCH, Boris MEUNIER, Guillaume MONNERET, Nathalie PANEL, Thomas RIMMELE, Hélène VALLIN and Fabienne VENET. For bioMérieux: Sophie BLEIN, Karen BRENGEL-PESCE, Elisabeth CERRATO, Valérie CHEYNET, Emmanuelle GALLET-GORIUS, Audrey GUICHARD, François MALLET, Virginie MOUCADEL, Marine MOMMERT, Guy ORIOL, Alexandre PACHOT, Claire SCHREVEL, Olivier TABONE, Julien TEXTORIS and Javier YUGUEROS MARCOS. For BIOASTER: Jérémie BECKER, Frédéric BEQUET, Yacine BOUNAB, Nathalie GARCON, Irène GORSE, Cyril GUYARD, Fabien LAVOCAT, Philippe LEISSNER, Karen LOUIS, Maxime MISTRETTA, Yoann MOUSCAZ, Laura NOAILLES, Magali PERRET, Frédéric REYNIER, Cindy RIFFAUD, Mary Luz ROL, Nicolas SAPAY, Trang TRAN and Christophe VEDRINE. For Sanofi: Nicolas BURDIN, Christophe CARRE, Pierre CORTEZ, Aymeric DE MONFORT, Karine FLORIN, Laurent FRAISSE, Isabelle FUGIER, Sandrine PAYRARD, Annick PELERAUX and Laurence QUEMENEUR. For ESPCI Paris: Andrew GRIFFITHS and Stephanie TOETSCH. For GSK: Theresa ASHTON, Peter GOUGH, Scott BERGER, Lionel TAN, Iain GILLESPIE and David GARDINER.

Contributors All authors (M-LR, FV, TR, VM, PC, LQ, DG, AG, AP, JT and GM) fulfilled ICMJE guidelines and provided substantial contributions to conception, design and acquisition of data; drafted and revised critically the manuscript; and approved the final version of the manuscript.

Competing interests AP, JT and VM are employees of bioMérieux SA, an in vitro diagnostic company. PC, LQ and DG are employees of Sanofi-Aventis R&D, Sanofi-Pasteur SA and GlaxoSmithKline, three pharmaceutical companies.

Ethics approval Comit de Protection des Personnes Lyon Sud-Est 2.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Results will be disseminated through presentations at scientific meetings and publications in peer-reviewed journals. New markers and immune functional tests will be evaluated for the diagnostic immune deficiency and may be patentable.

REFERENCES

1. Jawad I, Lukšić I, Rafnsson SB. Assessing available information on the burden of sepsis: global estimates of incidence, prevalence and mortality. *J Glob Health* 2012;2:4.
2. Vincent JL, Marshall JC, Namendys-Silva SA, *et al*. Assessment of the worldwide burden of critical illness: the Intensive Care Over Nations (ICON) audit. *Lancet Respir Med* 2014;2:380–6.
3. Walkey AJ, Lagu T, Lindenauer PK. Trends in sepsis and infection sources in the United States. A population-based study. *Ann Am Thorac Soc* 2015;12:216–20.
4. Seymour CW, Rea TD, Kahn JM, *et al*. Severe sepsis in pre-hospital emergency care: analysis of incidence, care, and outcome. *Am J Respir Crit Care Med* 2012;186:1264–71.
5. Singer M, Deutschman CS, Seymour CW, *et al*. The Third International Consensus Definitions for Sepsis and Septic shock (Sepsis-3). *JAMA* 2016;315:801.
6. Angus DC, van der Poll T, Sepsis S, *et al*. Severe sepsis and septic shock. *N Engl J Med* 2013;369:840–51.
7. Cohen J, Vincent JL, Adhikari NK, *et al*. Sepsis: a roadmap for future research. *Lancet Infect Dis* 2015;15:581–614.
8. Hotchkiss RS, Monneret G, Payen D. Sepsis-induced immunosuppression: from cellular dysfunctions to immunotherapy. *Nat Rev Immunol* 2013;13:862–74.
9. Hotchkiss RS, Monneret G, Payen D. Immunosuppression in sepsis: a novel understanding of the disorder and a new therapeutic approach. *Lancet Infect Dis* 2013;13:260–8.
10. Delsing CE, Gresnigt MS, Leentjens J, *et al*. Interferon-gamma as adjunctive immunotherapy for invasive fungal infections: a case series. *BMC Infect Dis* 2014;14:166.
11. Venet F, Lukaszewicz AC, Payen D, *et al*. Monitoring the immune response in sepsis: a rational approach to administration of immunoadjuvant therapies. *Curr Opin Immunol* 2013;25:477–83.
12. Meisel C, Schefold JC, Pschowski R, *et al*. Granulocyte-macrophage colony-stimulating factor to reverse sepsis-associated immunosuppression: a double-blind, randomized, placebo-controlled multicenter trial. *Am J Respir Crit Care Med* 2009;180:640–8.
13. Gentile LF, Cuenca AG, Efron PA, *et al*. Persistent inflammation and immunosuppression: a common syndrome and new horizon for surgical intensive care. *J Trauma Acute Care Surg* 2012;72:1491–501.
14. Angele MK, Chaudry IH. Surgical trauma and immunosuppression: pathophysiology and potential immunomodulatory approaches. *Langenbecks Arch Surg* 2005;390:333–41.
15. Kimura F, Shimizu H, Yoshidome H, *et al*. Immunosuppression following surgical and traumatic injury. *Surg Today* 2010;40:793–808.
16. Timmermans K, Kox M, Vaneker M, *et al*. Plasma levels of danger-associated molecular patterns are associated with immune suppression in trauma patients. *Intensive Care Med* 2016;42:551–61.
17. Gossez M, Malcus C, Demaret J, *et al*. Evaluation of a novel automated volumetric flow cytometer for absolute CD4+ T lymphocyte quantitation. *Cytometry B Clin Cytom* 2016:n/a.
18. http://www.clinicaltrial.gov/.
19. Duffy D, Rouilly V, Libri V, *et al*. Functional analysis via standardized whole-blood stimulation systems defines the boundaries of a healthy immune response to complex stimuli. *Immunity* 2014;40:436–50.
20. Poujol F, Monneret G, Friggeri A, *et al*. Flow cytometric evaluation of lymphocyte transformation test based on 5-ethynyl-2'deoxyuridine incorporation as a clinical alternative to tritiated thymidine uptake measurement. *J Immunol Methods* 2014;415:71–9.
21. Baker SP, O'Neill B, Haddon W, *et al*. The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. *J Trauma* 1974;14:187–96.
22. European Centre for Disease Prevention and Control. *Point prevalence survey of healthcareassociated infections and antimicrobial use in european acute care hospitals*: ECDC, 2012.
23. van Vught LA, Klein Klouwenberg PM, Spitoni C, *et al*. Incidence, risk factors, and attributable mortality of secondary infections in the intensive care unit after admission for sepsis. *JAMA* 2016;315:1469–79.
24. Demaret J, Venet F, Friggeri A, *et al*. Marked alterations of neutrophil functions during sepsis-induced immunosuppression. *J Leukoc Biol* 2015;98:1081–90.
25. Venet F, Chung CS, Kherouf H, *et al*. Increased circulating regulatory T cells (CD4(+)CD25 (+)CD127 (-)) contribute to lymphocyte anergy in septic shock patients. *Intensive Care Med* 2009;35:678–86.
26. Döcke WD, Höflich C, Davis KA, *et al*. Monitoring temporary immunodepression by flow cytometric measurement of monocytic HLA-DR expression: a multicenter standardized study. *Clin Chem* 2005;51:2341–7.
27. Horowitz GL, Altaie S, Boyd JC, *et al*. *CLSI C28-A3 defining, establishing and verifying reference intervals in the Clinical Laboratory Approved Guideline*, 2008.
28. Fine JP, Gray RJ. A proportional hazards Model for the subdistribution of a competing risk. *J Am Stat Assoc* 1999;94:496–509.
29. Vincent JL, Moreno R, Takala J, *et al*. The SOFA (Sepsis-related organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related problems of the European Society of Intensive Care Medicine. *Intensive Care Med* 1996;22:707–10.
30. Le Gall JR, Lemeshow S, Saulnier F. A New Simplified Acute Physiology Score (SAPS II) based on a European/North American Multicenter study. *JAMA* 1993;270:2957–63.
31. TranSMART. http://transmartfoundation.org/.

6.3 ANNEXE 3 : RAPPORT D'ANALYSES DE LA COHORTE MIP

RAPPORT D'ANALYSES DE L'EXPRESSION DANS LE SANG DE 102 PATIENTS EN CHOC SEPTIQUE (MIP) A PARTIR DE LA PUCE HERV-V3.

HERVs expression in septic shock patients

REALISM Project

Olivier Tabone, Marine Mommert, Virginie Moucadel, François Mallet, Julien Textoris

December 2016

Abstract

The study intends to identify transcriptomic markers of immunosuppression. We aim to validate the interest of these markers for immunodeficiency diagnosis on REALISM cohort and valorize them through patents and publications. We focused on Human Endogenous Retroviruses or HERV and Mammalian Apparent LTR-Retrotransposons or MalR which are type I mobile elements.

In order to identify markers associated to immunosuppression, we measured expression of more than 350 000 HERV and MalR, on 102 septic shock patients from MIP_REA cohort, using custom Affymetrix DNA microarray. For each patient, we had 3 samples collected on day 1, day 3 and day 6 after the shock diagnosis. As we did not have the immunosuppression status of these patients, we first used the occurrence of Hospital Acquired Infection (HAI) as a proxy of immunosuppression. Indeed, HAI can be considered as an accepted indirect marker of immunosuppression state because several biomarkers linked to the immune system have been associated to an increased incidence of HAI (e.g. monocytic HLA-DR). Another approach was to correlate any HERV or MalR biomarker with the expression of CD74 (the gene coding the invariant chain of the HLA-DR complex) - available from RTqPCR experiments made on the same cohort - CD74 Day 3/Day 1 ratio is indeed inversely correlated with the occurrence of HAI. Finally, a secondary objective was the association with mortality at day 28, to explore if some HERVs or MalR could be survival prognostic markers.

After a descriptive section, a differential expression analysis was performed in order to identify a list of potential biomarkers for all considered endpoints. We selected markers according to defined thresholds (fold changes and adjusted p_values). We did not find any differentially expressed probesets for HAI and mortality endpoints.

However, we found 115 differentially expressed probesets for the CD74 endpoint. Among these probesets, some targeted genes and other targeted HERV/MALRs elements. In total, 10 different genes and 31 HERV/MALRs elements were found differentially expressed. Interestingly, several differentially expressed HERV/MALRs could be mapped in close proximity of or within conventional genes that were also found differentially expressed. Eight HERVs elements were differentially expressed and co-localized with IL18R1, IL18RAP, IL1R1 and IL1R2, all belonging to the interleukine 1 receptors family on chromosome 2, and also found differentially expressed. Other differentially expressed HERVs could be localized close to or within a gene, like OLAH or FKBP5, but differentiall expression of these last could not be evaluated as they were not present on HERV-V3 chip.

Finally we realized a longitudinal differential expression analysis and we found 16 genes and 166 HERVs/MALR elements differentially expressed between D1 and late time points (D1 vs D3/D6). Among them, we identified 3 genes, IL10, MME and MERTK that were co-localized with differentially expressed HERV/MALRs.

Contents

```
##############
# load data #
##############
 #filtered data
load(paste0('../data/cdf',icdf,'/fltr_RMA_CV',CVthr,'.RData')) #fltr_RMA object
#unfiltered data (batch corrected)
load(paste0('../data/cdf',icdf,'/edata_RMA_batchcorrected.RData')) #HERVV3_RMA object
#unfiltered data (no batch corrected)
#load(paste0("../data/cdf",icdf,"/HERVV3_RMA.RData"))
#covdesc
load('../data/covdesc.RData') #covdesc object
```

Introduction

Goal of the study

The goal of this study is to identify transcriptomic markers of immunosuppression (genes or HERV/MALRs. In MIPrea retrospective study, we tried to identify lists of biomarkers associated to Hopital Acquired Infection (HAI), ratio between D3 and D1 of CD74 and death at D28 as proxies of immunosuppression in septic shock patients.

Scientific context

HERV (for Human Endogenous RetroViruses) were initially exogenous retroviruses which have infected germ cells and become integrated in our genome million years ago. MALR (for Mammalian-Apparent Long Terminal Repeat Retrotransposons) are ancesters of retroviruses. They are LTR retrotranspons of type I. They represent around 8% of our genome and they are known to be mostly inactive in physiological conditions (except for a few of them, e.g. syncytine (Pérot et al. 2012)). But, following injury (trauma, burn, infection or other aggression) some HERVs can be expressed. For example modulation of HERV expression has been shown in auto-immune diseases and cancer context like multiple sclerosis (Laska et al. 2012), breast cancer (Rhyu et al. 2014), testis cancer (Gimenez et al. 2010), colon cancer (Pérot et al. 2015). Knowing that, it seems interesting to look at the expression of these potential markers of immunosuppression state.

For this purpose, we studied HERV / MalR expression in patients after a septic shock. Indeed, septic shock is the worst outcome of septic syndrome, characterized by an acute inflammatory phase, followed by a deep immunosuppression phase. We dispose of samples collected at day 1, day 3 and day 6 after the diagnosis for each patient. HERV-V3 DNA microarray is a customized Affymetrix chip designed within the joint research unit bioMérieux/HCL, addressing specifically more than 350 000 HERVs and MalR elements, but also 1500 genes known to be involved in immunity.

The outcome of this study is the immunosuppression state of the patient. However we did not have directly access to this status. We had the HAI outcome (Hospital Acquired Infection) as an indirect marker of this status. This was the principal outcome. Moreover we used CD74 mRNA expression (RTqPCR experiment, from previous intra LCR study). As it is the invariant chain of HLA-DR (MHC class II), a well-known marker of immunosuppression, we could test the correlation of its expression with HAI outcome. Ratio between D3 and D1 of CD74 expression was used as outcome, by splitting the cohort into two groups, patients with high ratio and patients with low ratio (threshold is equal to 1.23, it was defined in previous study) . Another secondary outcome was the death at D28 after septic shock.

Repertoire	Number of probesets	Number of elements	Description
HERV_proto	90 654	24 387	Well annotated HERVs, different sub regions are defined.
HERV_Dfam	567 168	150 685	HERVs from Dfam database, sub regions not clearly identified.
MALR_Dfam	622 396	179 283	MALRs from Dfam database. Ancestors of retroviruses.
Lines	2 826	664	Retrotransposons without LTR.
lncRNA	7 552	3769	Long Non Coding RNA.
Centromeric	58	23	HERVs localized on centromeres.
Virus	734	289	Exogenous viruses.
HTA	70 632	1557	Probesets from HTA v2 chips. Target conventional genes.
U133	7 704	1684	Probesets from U133 plus2 chips. Target conventional genes.
Opti	3032	1516	Home made designed probesets. Target conventional genes.

Figure 1: **HERV-V3 microarray description**. The differences observed between number of probesets and number of elements is because multiple probesets target different regions (or exons for genes) of the same element (HERV/MALR or gene). HTA, U133 and Opti repertoires target the same subset of genes

Material and Methods

Cohort

We selected a subset of the septic shock cohort of MIP REA with all three time points D1, D3 and D6. It represents 102 patients and 306 samples.

HERV-V3 microarray

See Figure 1

```
knitr::include_graphics("../chip_description.png")
```

HERV-V3 chips generation

Total RNA was extracted from whole blood using PAXgeneTM Blood RNA Kit (PreAnalytix, Hilden, Germany), using an amended version of the manufacturer's guidelines. RNA integrity was assessed with the RNA 6000 Nano Kit on a Bioanalyzer (Agilent Technologies, Santa Clara, California).

The cDNA synthesis and amplification steps were performed from 16 ng of RNA using the Ovation Pico WTA System V2 kit (Nugen). Briefly, cDNA synthesis was done using a mixture of random and polydT primers, followed by the synthesis of the complementary strand. The Single Primer Isothermal Amplification (SPIA) was then performed with hybrid DNA/RNA primers sensitive to RNAse-H digestion, in the presence of a DNA polymerase with strong strand displacement activity. The resulting amplified cDNA was purified using the QIAquick purification kit (Qiagen), from which, total DNA concentration was measured using the NanoDrop 1000 spectrophotometer (Thermo Scientific) and the product quality was checked on the Bioanalyser 2100. Five micrograms of purified dsDNA were fragmented by enzymatic activity into 50-200 bp

fragments by using Encore Biotin Module (Nugen). The product quality was checked also on the Bioanalyser 2100. 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 99°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-V3 microarrays were prehybridized with 200 µL of pre-hybridization buffer and placed under stirring (60 rpm) in an oven at 50°C for 10 minutes. The pre-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 finally scanned using a GeneChip scanner 3000 7G, Affymetrix.

R libraries

Output from sessionInfo() command in R:

```
sessionInfo()
```

```
## R version 3.3.0 (2016-05-03)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.5 LTS
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] grid      parallel  stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] sva_3.20.0            mgcv_1.8-16           nlme_3.1-127
##  [4] png_0.1-7             pheatmap_1.0.8        xtable_1.8-2
##  [7] tidyr_0.6.0           dplyr_0.5.0           gridExtra_2.2.1
## [10] made4_1.46.0          scatterplot3d_0.3-37  gplots_3.0.1
## [13] RColorBrewer_1.1-2    ade4_1.7-4            limma_3.28.19
## [16] ggplot2_2.1.0         pcaMethods_1.64.0     affyQCReport_1.50.0
## [19] lattice_0.20-34       yaqcaffy_1.32.0       affyPLM_1.48.0
## [22] preprocessCore_1.34.0 simpleaffy_2.48.0     gcrma_2.44.0
## [25] genefilter_1.54.0     affy_1.50.0           Biobase_2.32.0
## [28] BiocGenerics_0.18.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.6          Biostrings_2.40.0    gtools_3.5.0
##  [4] assertthat_0.1       digest_0.6.9         R6_2.1.2
##  [7] plyr_1.8.4           stats4_3.3.0         RSQLite_1.0.0
## [10] evaluate_0.9         BiocInstaller_1.22.0 zlibbioc_1.18.0
## [13] annotate_1.50.0      gdata_2.17.0         S4Vectors_0.10.0
```

```
## [16] Matrix_1.2-6          rmarkdown_1.0         splines_3.3.0
## [19] stringr_1.1.0         munsell_0.4.3         htmltools_0.3.5
## [22] tibble_1.1            codetools_0.2-15      IRanges_2.6.0
## [25] XML_3.98-1.4          bitops_1.0-6          gtable_0.2.0
## [28] DBI_0.5               magrittr_1.5          formatR_1.4
## [31] scales_0.4.0          KernSmooth_2.23-15    stringi_1.0-1
## [34] XVector_0.12.0        affyio_1.42.0         tools_3.3.0
## [37] survival_2.39-2       yaml_2.1.13           AnnotationDbi_1.34.0
## [40] colorspace_1.3-1      caTools_1.17.1        knitr_1.14
```

Pre processing steps

Batching

306 microarrays were hybridized in 13 batches. We defined batches to ensure an even distribution of the following variables, as described below:

Batching was done according to HAI (Yes or No), Death at D28 (Survivors or Not survivors), Severity (IGSII score binarized, severe or not severe), Sex (Male or Female), Center of inclusion (from 1 to 6) and microarray lots (4260848, 4260849 and 4276952).

```
knitr::include_graphics("../batching_summary.png")
```

On the figure 2, each column represent a batch. Within each batch, are represented the relative proportions of each modalities of the variables used for the batching. No major bias was observed, the variable distributions were globally similar in all batches.

Quality check

In this section are presented some key figures of the quality check analysis. The script and all figures are available in appendix. 306 chips have been processed corresponding to the 3 samples (D1, D3, D6) for the 102 patients of the cohort. 305 chips could be analyzed, the remaining one had no signal at all. First we checked images quality of all microarrays. The majority presented no visual artefacts. Two images were of bad quality. Then we checked several quality controls.

This is a list of the criteria checked:

- Chips image: Visual check of chips images
- Hybridization controls
- RIN: RNA Integrity Number, check RNA quality
- Amplification control
- Fragmentation control
- Boxplots RAW: intensities before data *normalization*
- Boxplots normalized: intensities after data *normalization*
- NUSE plot: Comparison of Normalized Unscaled Standard Error between chips
- RLE plots: Comparison of Relative Log expression between chips.
- Correlation raw: Between arrays correlations before data normalization.
- Correlation RMA: Between arrays correlation after data normalization.
- PCA: Principal Component Analysis

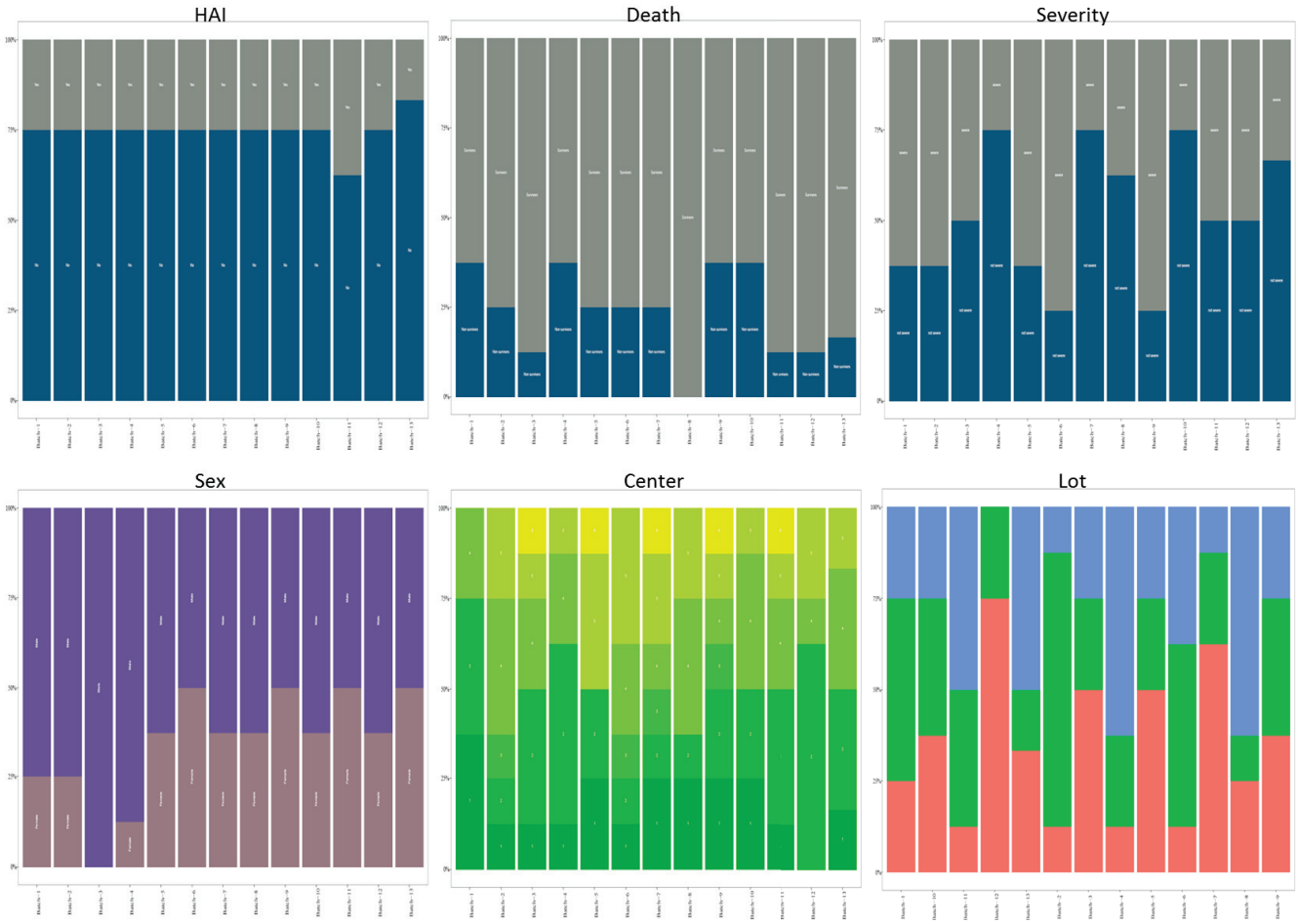


Figure 2: **Batching results**. Each column represent a batch. Within each batch, are represented the relative proportions of each modality of the variables used for the batching

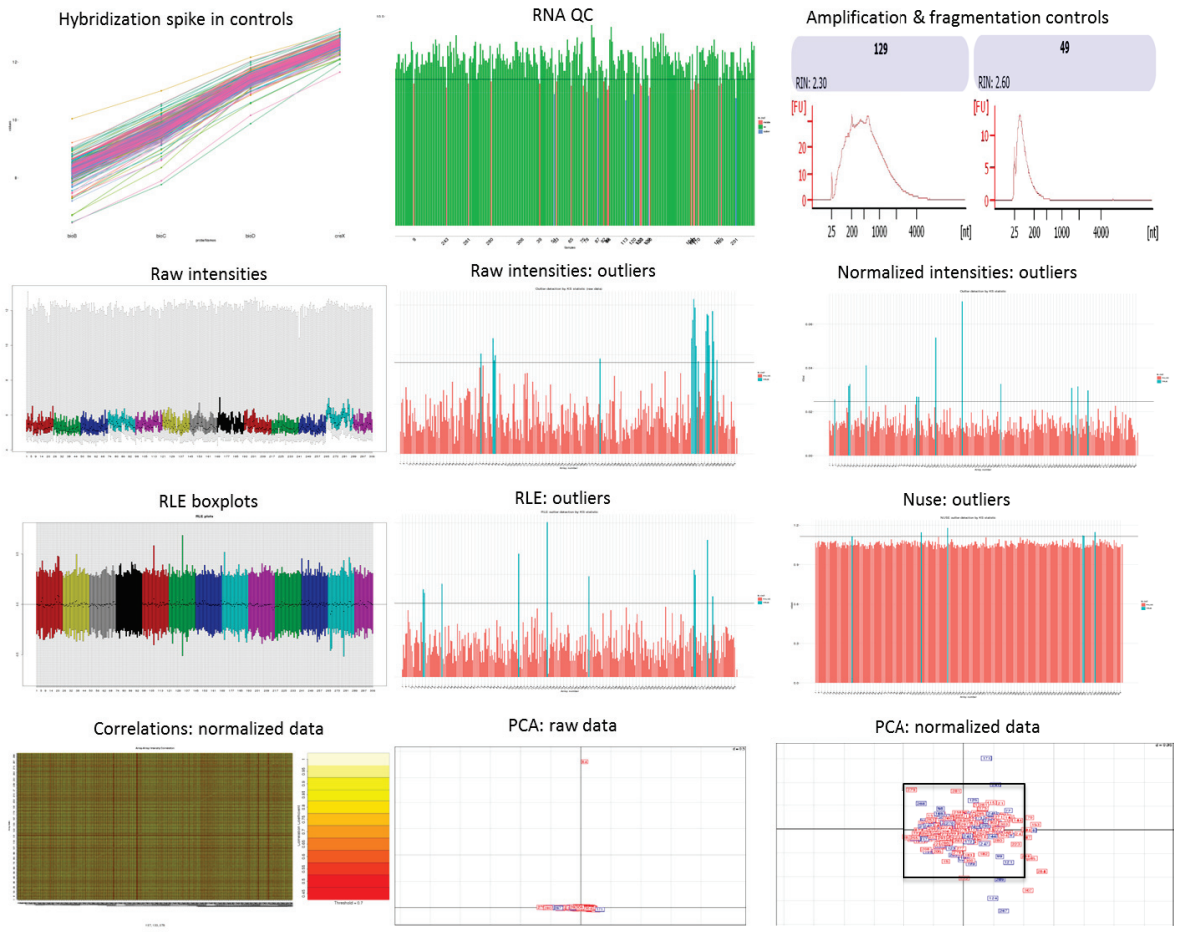


Figure 3: **Quality Check**. Key figures of quality check analysis are shown. All figures are avalaible in appendix.

```
knitr::include_graphics("../output/QC/QC_figures.png")
```

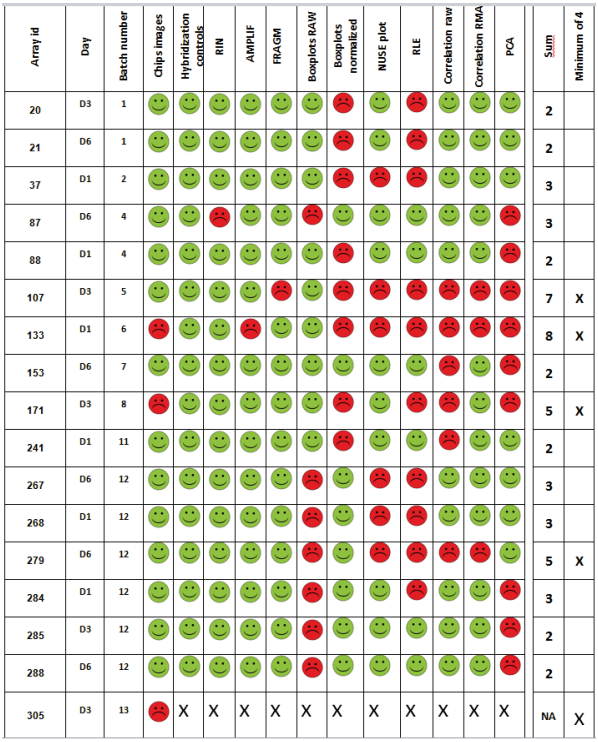
Some of the criteria checked are presented in the figure 3.

In order to decide which chips were of bad quality, we counted how many criteria a chip failed for quality check. If a chip had 4 bad criteria or more, we removed it for analysis. In this figure only chips with at least 2 bad criteria are shown. Meaning that the other chips had 1 or 0 bad criterion.

```
knitr::include_graphics("../output/QC/dec_table_2min.PNG")
```

As we can see on Figure 5, 5 samples were removed from the data analysis, representing 1.6 % of the total samples.

```
knitr::include_graphics("../output/QC/tab_smpl_removed.png")
```



Array id	Day	Batch number	Chips images	Hybridization controls	RIN	AMPLIF	FRAGM	Boxplots RAW	Boxplots normalized	NUSE plot	RLE	Correlation raw	Correlation RMA	PCA	Sum	Minimum of 4
20	D3	1	☺	☺	☺	☺	☺	☺	☹	☺	☹	☺	☺	☺	2	
21	D6	1	☺	☺	☺	☺	☺	☺	☹	☺	☹	☺	☺	☺	2	
37	D1	2	☺	☺	☺	☺	☺	☺	☹	☹	☹	☺	☺	☺	3	
87	D6	4	☺	☺	☹	☺	☺	☹	☺	☺	☺	☺	☺	☹	3	
88	D1	4	☺	☺	☺	☺	☺	☺	☹	☺	☺	☺	☺	☹	2	
107	D3	5	☺	☺	☺	☺	☹	☺	☹	☹	☹	☹	☹	☹	7	X
133	D1	6	☹	☺	☺	☹	☺	☺	☹	☹	☹	☹	☹	☹	8	X
153	D6	7	☺	☺	☺	☺	☺	☺	☺	☺	☺	☹	☺	☹	2	
171	D3	8	☹	☺	☺	☺	☺	☺	☹	☺	☹	☹	☺	☹	5	X
241	D1	11	☺	☺	☺	☺	☺	☺	☹	☺	☺	☹	☺	☺	2	
267	D6	12	☺	☺	☺	☺	☺	☹	☺	☹	☹	☺	☺	☺	3	
268	D1	12	☺	☺	☺	☺	☺	☹	☺	☹	☹	☺	☺	☺	3	
279	D6	12	☺	☺	☺	☺	☺	☹	☺	☹	☹	☹	☹	☺	5	X
284	D1	12	☺	☺	☺	☺	☺	☹	☺	☺	☹	☺	☺	☹	3	
285	D3	12	☺	☺	☺	☺	☺	☹	☺	☺	☺	☺	☺	☹	2	
288	D6	12	☺	☺	☺	☺	☺	☹	☺	☺	☺	☺	☺	☹	2	
305	D3	13	☹	X	X	X	X	X	X	X	X	X	X	X	NA	X

Figure 4: **Decision table**. Five arrays are outliers for four or more criteria checked. They have been removed of the dataset.

Num tubes	Num batch	Id	HAI	DEATH_D28	Sex	Severity	Center	Lot puces
107	Batch 5	2031 D3	No	Survivor	F	Severe	2	4260849
133	Batch 6	3016 D1	No	Survivor	F	Severe	3	4260849
171	Batch 8	2160 D6	Yes	Survivor	F	Not severe	2	4276962
279	Batch 12	5236 D6	No	Survivor	F	Severe	5	4260848
305	Batch 13	1022 D3	No	Non Survivor	F	Severe	1	4276962

Figure 5: **Samples removed from QC**. The main clinical or technical variables are shown for each of the sample removed.

Figure 6: **PCA before correction**. Each point represents a sample. On the left samples are colored by batch and corresponding centroids; on the right samples are colored by HAI status and corresponding ellipses.

Batch correction

ComBat for COMbining BATches of gene expression microarray data is an algorithm based on parametric and nonparametric empirical Bayes frameworks for adjusting data for batch effects that is robust to outliers in small sample sizes and performs comparable to existing methods for large samples (Johnson, Li, and Rabinovic 2007)

Using PCA, we can visualize and determine if there are non-desirable effects of some variables. We saw an effect of the "batch variable", corresponding of the different sessions of experiment. So we corrected it using COMBAT.

PCA before correction

```
#colored by batch (left) or hai (right)
knitr::include_graphics("../output/Batch_Correction/pca_RMA_batch_hai.png")
```

The graphics are results of PCA on the 2 first components. Each point represents a sample. On the left each sample is colored by batch number. On the right by HAI status.

PCA after correction with COMBAT

```
#colored by batch (left) or hai (right)
knitr::include_graphics("../output/Batch_Correction/pca_combat_batch_hai.png")
```

We can observe that before correction with Combat, the ellipses and centroids were very different between batches. After correction, they were almost identical (left part of the figures 6 and 7). The right parts show that it did not change the ellipses for the variable of interest HAI. We tried to detect eventual bias with other technical variables (Operator, lot of chips, ...) but we did not detect another evident bias.

Data filtering

Before starting analysis, filtering low intensity probesets is necessary in order to gain power and time. We computed Coefficient of Variation (CV) by range of intensity (Figure 8. The intensity threshold has been

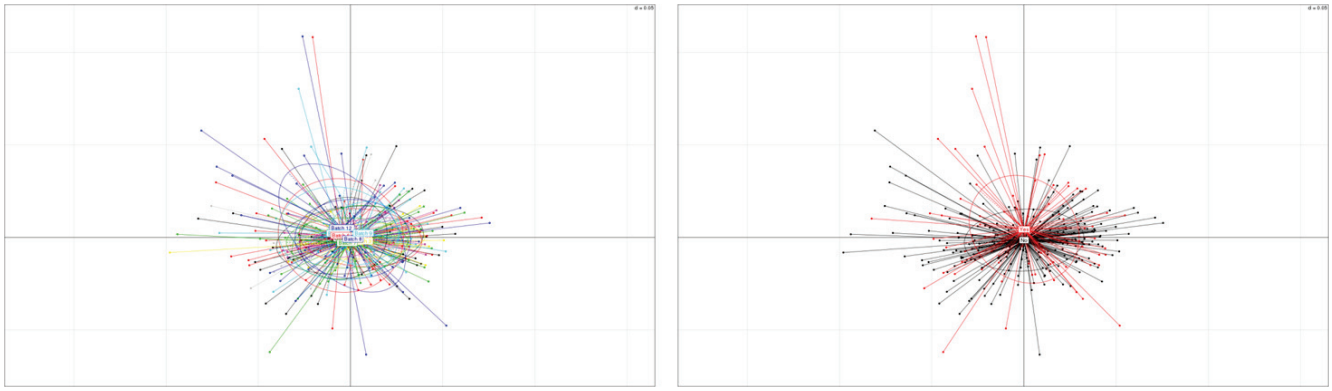


Figure 7: **PCA after correction**. Each point represents a sample. On the left samples are colored by batch and corresponding centroids; on the right samples are colored by HAI status and corresponding ellipses.

defined as the lowest intensity for which the third quartile of CV (for HERV repertoires) is under a CV=0.2. From this figure, we choose an intensity threshold equal to 3.5.

```
knitr::include_graphics("../output/Analysis/cdf1/CVs0.2_rep2.png")
```

Then we kept probesets for which at least 24 samples (the smaller group,HAI patients at Day 6, is composed of 25 samples.) were over the intensity threshold (3.5). This way we obtain a dataset with 462649 probesets (33.5 % of total dataset) and 301 samples.

Cohort description

Clinical variables

Here are shown some clinical variables used for batching and those for which there was a difference between HAI and No HAI patients (10).

```
knitr::include_graphics("../output/cohort1.png")
```

```
knitr::include_graphics("../output/cohort2.png")
```

- Charlson score : Morbidity score. Sign of preexisting failure
- Glasgow score : between 3 and 15. Coma score. Eyes, Voice, Motricity.
- PaO2/FiO2 D1: Arterial Pressure / Inhaled fraction. Capacity of lungs to transfer oxygen into blood. If <200: critical respiratory disorder.

We saw lower morbidity and coma scores in HAI group compared to No HAI group. The ratio PaO2/FiO2 at day1 was critical in both groups, but even more critical for HAI group. And we observed the consequences of critical state of the HAI patients as they had higher duration of catecholamines and dialysis than the NO HAI group.

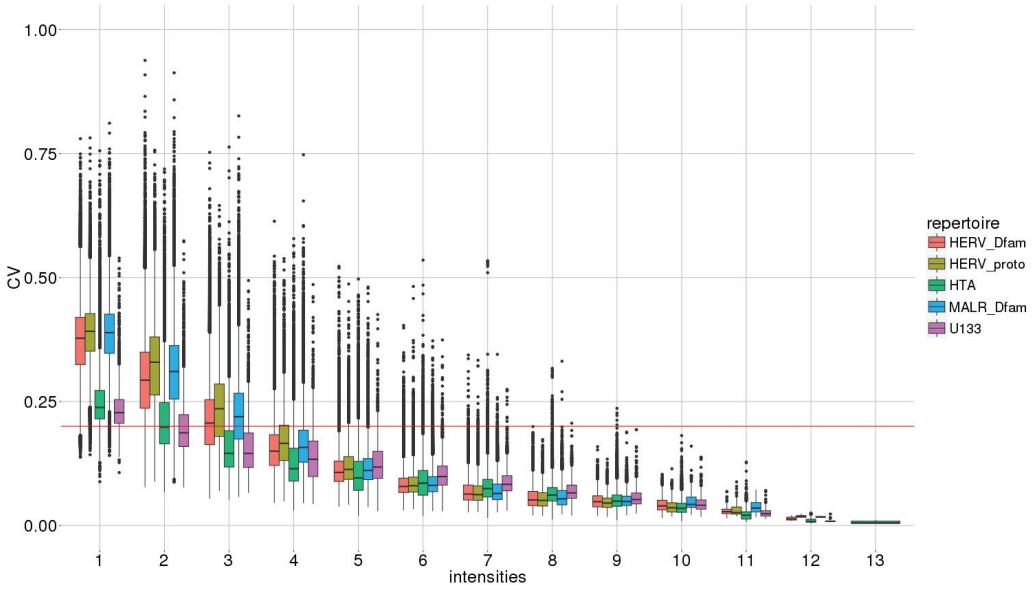


Figure 8: **Coefficient of Variations by range of intensities for each repertoire**. The red line corresponds to the CV threshold equal to 0.2. The intensity for which the 3rd quartile of all repetoires are under the red line corresponds to the range [3.5;4.5].

		HAI (n=26)	NO HAI (n=76)	Total (n=102)	P-value
Gender	Female	7 (26.92 %)	28 (36.84 %)	35 (34.31 %)	0.496
	Male	19 (73.08 %)	48 (63.16 %)	67 (65.69 %)	
Death D28	Non survivors	4 (15.38 %)	20 (26.32 %)	24 (23.53 %)	0.298
	Survivors	22 (84.62 %)	56 (73.68 %)	78 (76.47 %)	
SAPS II	Mean (SD)	63.38 (15.02)	63.21 (16.36)	63.25 (15.96)	0.962
	Median [Q1,Q3]	66.50 [55.25,73.75]	62.00 [51.00,72.00]	63.00 [51.25,72.75]	
	min - max	31.00 - 85.00	30.00 - 98.00	30.00 - 98.00	
Charlson score	Mean (SD)	1.38 (1.42)	2.36 (2.03)	2.11 (1.93)	0.023
	Median [Q1,Q3]	1.00 [0.00,2.00]	2.00 [1.00,3.00]	2.00 [1.00,3.00]	
	min - max	0.00 - 5.00	0.00 - 8.00	0.00 - 8.00	

Figure 9: **Clinical variables**

		HAI (n=26)	NO HAI (n=76)	Total (n=102)	P-value
Glasgow score	Mean (SD)	9.58 (4.48)	12.12 (4.08)	11.47 (4.31)	0.032
	Median [Q1,Q3]	9.50 [6.00,14.00]	14.00 [9.00,15.00]	14.00 [8.00,15.00]	
	min - max	3.00 - 15.00	3.00 - 15.00	3.00 - 15.00	
PaO2/FiO2 D1	Mean (SD)	130.88 (58.46)	178.67 (87.08)	166.84 (83.29)	0.013
	Median [Q1,Q3]	127.00 [83.00,168.00]	164.50 [106.50,243.50]	155.00 [100.00,215.00]	
	min - max	59.00 - 262.00	44.00 - 490.00	44.00 - 490.00	
Duration of catecholamines, Days	Mean (SD)	6.66 (6.46)	4.14 (3.43)	4.78 (4.51)	0.045
	Median [Q1,Q3]	3.80 [2.80,9.12]	3.10 [1.70,5.05]	3.15 [1.83,5.95]	
	min - max	0.20 - 27.20	0.50 - 17.60	0.20 - 27.20	
Dialysis duration, Days	Mean (SD)	22.62 (18.05)	6.41 (4.53)	11.60 (12.97)	0.049
	Median [Q1,Q3]	22.50 [7.50,31.25]	5.00 [3.00,9.00]	6.00 [3.00,15.00]	
	min - max	2.00 - 51.00	1.00 - 17.00	1.00 - 51.00	

Figure 10: **Clinical variables**

Contingency tables

Some description between variables of interest. The goal is to cross two by two some variables of interest (with contingency tables) and see if there is a relationship between the two variables. For example: "are the patient with HAI are preferentially Non survivors ?".

The variable ratioCD74 is the ratio of CD74 between D3 and D1. The value is High for patients with ratio $>$ 1.23 and Low for patients with ratio $<$ 1.23. This variable could be a proxy of immunosuppression as the patients, after an initial pro-inflammation, can enter into immunosuppression state (D3/D1 $<$1.23 $=>$ Low ratio) or return to the "normal" (D3/D1 $>$ 1.23, High ratio).

First, a summary of HAI and ratioCD74 variables:

```
#recoding of hai and ratioCD74 variables
covdesc$hai=as.factor(sapply(as.character(covdesc$HAI),
                             function(x){if(x=="Yes") return("HAI") else return("No_HAI")}))
covdesc$ratioCD74 = as.factor(sapply(as.character(covdesc$youden_cd74),
                                     function(x){if(x=="FALSE") return("Low") else return("High")}))
#comparison between patients with/without HAI and patients classified
# with ratio cd74 D3/D1 over youden value (1.23)
summary(covdesc$hai[!duplicated(covdesc$patient)])
```

```
##    HAI No_HAI
##     26     76
```

```
summary(covdesc$ratioCD74[!duplicated(covdesc$patient)])
```

```
## High  Low
```

```
##    50    52
```

- HAI vs. ratioCD74:

```
patients=covdesc[!duplicated(covdesc$patient),]
#table(patients$hai, patients$ratioCD74)
knitr::include_graphics("../hai_cd74.png", dpi=200)
```

	High	Low
HAI	**7** (27%) (14%)	**19** (73%) (37%)
No_HAI	**43** (57%) (86%)	**33** (43%) (63%)

```
chisq.test(patients$hai, patients$ratioCD74)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  patients$hai and patients$ratioCD74
## X-squared = 5.6826, df = 1, p-value = 0.01713
```

With the Chi-squared test, we see a statistically significant relationship between the 2 variables. HAI is associated with low ratio of CD74, and No HAI is associated with high ratio of CD74, which is coherent as they are both proxy of immunosuppression.

- Mortality vs. HAI:

```
#mortality vs hai
# table(patients$DEATH_D28, patients$hai)
knitr::include_graphics("../death_hai.png", dpi=200)
```

	HAI	No_HAI
Non survivors	**4** (17%) (15%)	**20** (83%) (26%)
Survivors	**22** (28%) (85%)	**56** (72%) (74%)

```
#chisq.test(patients$DEATH_D28, patients$hai)
```

- Sex vs. HAI:

```
#sex vs hai
# table(patients$SEX, patients$hai)
knitr::include_graphics("../sex_hai.png", dpi=200)
```

	HAI	No_HAI
Female	7 (20%) (27%)	28 (80%) (37%)
Male	19 (28%) (73%)	48 (72%) (63%)

```
#chisq.test(patients$SEX, patients$hai)
```

There is no significant association between HAI status and mortality or sex. Chi squared tests not shown.

- Mortality vs. ratioCD74:

```
#mortality vs ratio cd74
# table(patients$DEATH_D28, patients$ratioCD74)
knitr::include_graphics("../death_cd74.png", dpi=200)
```

	High	Low
Non survivors	10 (42%) (20%)	14 (58%) (27%)
Survivors	40 (51%) (80%)	38 (49%) (63%)

```
# chisq.test(patients$DEATH_D28, patients$ratioCD74)
```

NB: There were 24 patients who died over the 102 patients in the cohort. It represents 23.5%, which is low for septic shocks patients. But we recall that the patients were choosen because D1, D3 and D6 samples were available.

- Sex vs. ratioCD74:

```
#sex vs ratio cd74
# table(patients$SEX, patients$ratioCD74)
knitr::include_graphics("../sex_cd74.png", dpi=200)
```

	High	Low
Female	16 (46%) (32%)	19 (54%) (37%)
Male	34 (51%) (68%)	33 (49%) (63%)

```
# chisq.test(patients$SEX, patients$ratioCD74)
```

There was no significant association between cd74 ratio and mortality or sex. Chi squared tests not shown.

Comparison RT-qPCR / HERV_V3

The goal of this section is to gain confidence on HERVV3 chip results by comparing the microarray dataset with RTqPCR dataset made on the same cohort in a previous study. 7 genes are both in RTqPCR dataset and HERV_V3 chips dataset: CD74, CX3CR1, CD3D, IL10, TNFa and IL1b.

```
gene_list_chip=c( "CD74", "CX3CR1", "CD3D", "IL10", "TNF", "IL1B")
gene_list_chip2=c( "CD74", "CX3CR1", "CD3D", "IL10", "TNFa", "IL1b")

load('../data/cdf1/edata_rtqpcr.RData')
load('../data/MipRea_D1.Rdata')
names(edata_rtqpcr)=gene_list_chip2 #alias names correspond with rtqpcr names

i_list=list()
for( i in seq(1,length(gene_list_chip2)))
{
  alias=gene_list_chip2[i]
  j=grep(paste0('^RQ_', alias), colnames(MipRea_D1))
  i_list[[i]]=j
}
col_index=unlist(i_list)
rq=MipRea_D1[,c(1,col_index)]

#selection of the good patients for wp2
id_patients=covdesc$patients
patients=unique(as.character(id_patients))

i=match(patients, rq$Ncenincl)
rq_wp2=rq[i[which(!is.na(i))],]
#length(i[which(!is.na(i))])

df_rq=rq_wp2 %>% gather(Ncenincl)
#head(df_rq)
colnames(df_rq)[1]="patient"
colnames(df_rq)[2]="type_alias_day"

#some values to remove
df_rq= df_rq %>% filter(type_alias_day!="RQ_IL10_D3_cens" &
                          type_alias_day!="RQ_IL10_D6_cens")
#dim(df_rq) #coherent with nmesuretypes*nbpatients of rq_wp2 -> 102*18

day_rq=sapply(df_rq$type_alias_day, FUN=function(x)
{
  return(strsplit(x,'_')[[1]][3])
})
type_rq=sapply(df_rq$type_alias_day, FUN=function(x)
{
  return(strsplit(x,'_')[[1]][1])
})
alias_rq=sapply(df_rq$type_alias_day, FUN=function(x)
{
  return(strsplit(x,'_')[[1]][2])
})
df_rq$type=type_rq
```

```
df_rq$alias=alias_rq
df_rq$Day=day_rq
```

```
#To compute a correlation, we need first to compute the mean of probesets,
#as we need the same length of vectors.

df_mean_ps=df_ggchip %>% group_by(alias, Day, patient) %>% summarize(mean=mean(value))
df_mean_ps = df_mean_ps %>% filter(alias!="IL7R")

# dim(df_mean_ps)
# dim(df_rq)
#dim are not equal because of samples removed in chips analysis
#the goal here is to reduce df_rq in order to have the same samples in each dataset
chip_test=paste(df_mean_ps$alias, df_mean_ps$Day, df_mean_ps$patient, sep='_')
rq_test=paste(df_rq$alias, df_rq$Day, df_rq$patient, sep='_')
i=match(chip_test, rq_test)
#identical(chip_test, rq_test[i[which(!(is.na(i)))]])
#identical(chip_test, rq_test[i])

df_rq=df_rq[i[which(!(is.na(i)))],]
# dim(df_mean_ps)
# dim(df_rq)
#now same dim and same order we can cbind it
```

```
df_gg=cbind(df_rq[,c("patient", "alias", "Day", "value")], df_mean_ps$mean)
colnames(df_gg)[4]="rq_values"
colnames(df_gg)[5]="chip_values"

df_ggchip=df_ggchip %>% filter(alias!="IL7R")
#plots on the same graph
#Intensity boxplots for each gene on HERVV3 chip
# and Correlation of intensities of these genes between HERVV3 and RTqPCR
p1=ggplot(df_ggchip)
p1=p1+geom_boxplot(aes(x=alias, fill=Day, y=value))+
  theme(panel.background = element_blank(),
        axis.text.x = element_text(size=12, colour='black', angle=45, hjust=1),
        axis.text.y = element_text(size=12, colour='black'),
        axis.line=element_line(colour="black"),
        panel.border = element_rect(color = "black", fill = NA, size = 1),
        panel.grid.major= element_line(colour="grey"),
        panel.grid.minor= element_line(colour="white"),
        legend.key.size=unit(0.5,"cm"),
        legend.text = element_text(size=10),
        plot.title=element_text(size=14))+ggtitle("Intensities on HERVV3 chip")

p2=ggplot(df_gg)
p2=p2+geom_point(aes(x=log2(rq_values), y=chip_values, color=alias), size=0.5)+
  theme(panel.background = element_blank(),
        axis.text.x = element_text(size=12, colour='black'),
        axis.text.y = element_text(size=12, colour='black'),
        axis.line=element_line(colour="black"),
        panel.border = element_rect(color = "black", fill = NA, size = 1),
        panel.grid.major= element_line(colour="grey"),
```

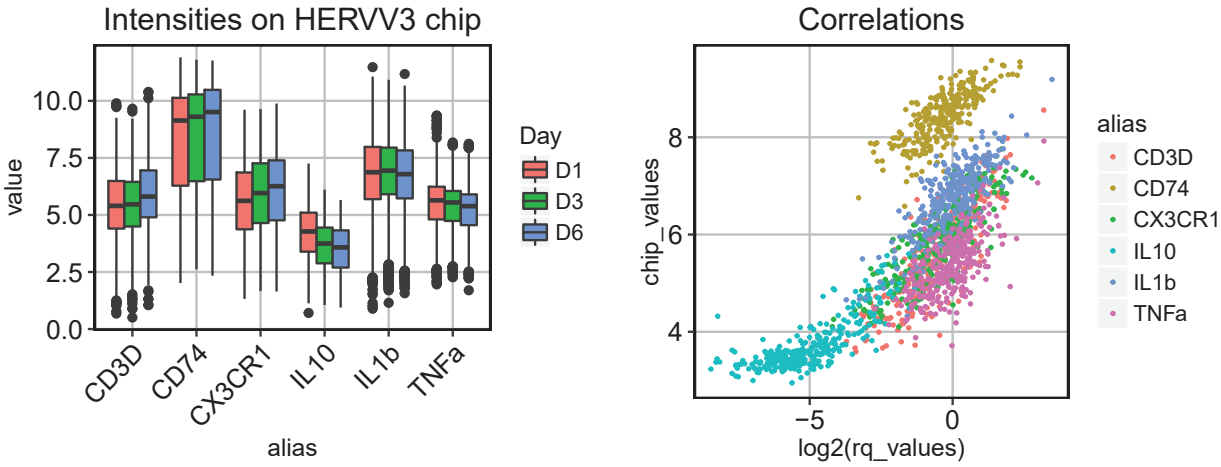


Figure 11: **RTqPCR vs. HERV-V3**. On the left, the boxplots represent the intensities of the 6 genes from HERV-V3 dataset, by day. On the right the graphic is a scatter plot of the intensity values from HERV-V3 dataset versus RTqPCR dataset.

```
        panel.grid.minor= element_line(colour="white"),
        legend.key.size=unit(0.5,"cm"),
        legend.text = element_text(size=10),
        plot.title=element_text(size=14))+ggtitle("Correlations")
```

```
grid.arrange(p1,p2, ncol=2)
```

On Figure 11 on the left are represented the log2(intensities) of the 6 genes from HERVV3 chip. The expression of these genes are quite high (intensity threshold = 3.5) in microarray data and are not very different between D1, D3 and D6. On the right are represented the correlation between HERVV3 chip and RTqPCR for the 6 genes present in both studies, colored by gene.

```
# Correlation by gene:
df_gg %>% group_by(alias) %>%
  summarize(cor=cor(log2(rq_values), chip_values, method="spearman"))
```

```
## # A tibble: 6 x 2
##    alias       cor
##    <chr>     <dbl>
## 1   CD3D 0.8254849
## 2   CD74 0.7739368
## 3 CX3CR1 0.8160645
## 4   IL10 0.7814619
## 5   IL1b 0.7543201
## 6   TNFa 0.6315800
```

We can see on the right graphic (11) and on the table above that the correlations between RTqPCR and HERV_V3 chips are pretty good. We can now enter in data exploration and analysis with more confidence.

Unsupervised analysis

Data description

```
i=getAnnotMatches(rownames(fltr_RMA))
#test before adding repertoire column in data

d_fltr=data.frame(psNames=rownames(fltr_RMA), repertoire = ann$repertoire[i])
myrep=apply(d_fltr,1,FUN=gen_rep2,n=ncol(d_fltr))
myrep=as.factor(myrep)
d_fltr=cbind(d_fltr, myrep)
```

Distributions by repertoire

Density curves. Distributions of the main repertoires.

```
load(paste0('../data/cdf',icdf,'/data_CV.RData'))
#only some repertories
sub_dcv=subset(d_CV, d_CV$repertoire=="HERV_Dfam" | d_CV$repertoire=="HERV_proto" |
                 d_CV$repertoire=="MALR_Dfam" |
                 d_CV$repertoire=="U133" | d_CV$repertoire=="HTA" |
                 d_CV$repertoire=="Opti")
#ordering levels
sub_dcv$repertoire=factor(sub_dcv$repertoire,
                          levels=c("HERV_proto", "HERV_Dfam", "MALR_Dfam",
                                   "HTA", "U133", "Opti"))
p1=ggplot(data=sub_dcv)
p1=p1+geom_density( aes(x=mean, fill=repertoire, colour=repertoire), alpha=0.2)+
  ggtitle("Distribution of intensities by repertoire")+
  theme(panel.background = element_blank(),
        axis.text.x = element_text(size=10, colour='black'),
        axis.text.y = element_text(size=10, colour='black'),
        panel.grid.major= element_line(colour="grey"),
        panel.grid.minor= element_line(colour="white"),
        legend.position="none")
```

```
p2=ggplot(data=sub_dcv)
p2=p2+geom_boxplot( aes(x=repertoire, y=mean, fill=repertoire))+
  ggtitle("Boxplots of intensities by repertoire")+
  theme(panel.background = element_blank(),
        axis.text.x = element_text(size=10, colour='black', angle=45, hjust=1),
        axis.text.y = element_text(size=10, colour='black'),
        panel.grid.major= element_line(colour="grey"),
        panel.grid.minor= element_line(colour="white"))
grid.arrange(p1,p2, ncol=2)
```

As seen in previous works, HERVs and MALRs have similar intensities, lower than HTA, U133 and Opti (Figure 12). We can see the usual shape of distributions for each repertoire. For hervs / Malr, the majority of intensities are between 1 and 2. For genes we see a bimodal distribution as expected.

Note: it seems that hervs and malr have more values at high intensity compared to previous datasets.

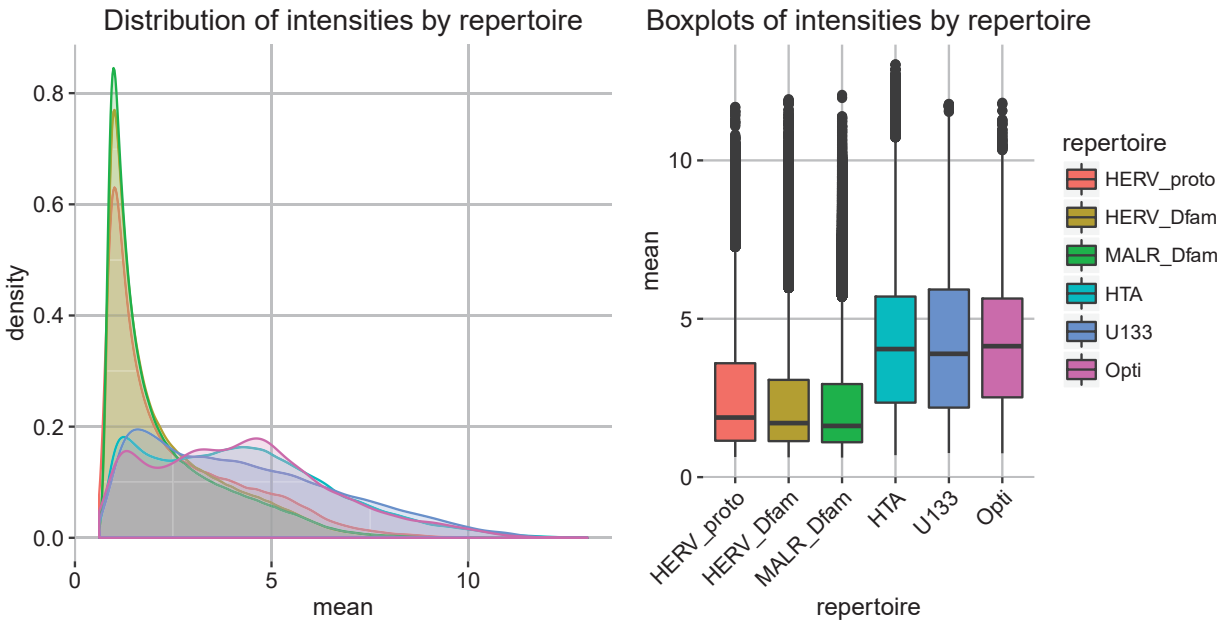


Figure 12: **Intensities by repertoire**. On the left the densities of intensities for the main repertoires. On the right the boxplots

Repartition by repertoire: piechart

```
#HERVV3 chip composition
i=getAnnotMatches(rownames(exprs_RMA))

d_RMA=data.frame(psNames=rownames(exprs_RMA), repertoire = ann$repertoire[i])
myrep=apply(d_RMA,1,FUN=gen_rep2,n=ncol(d_RMA))
myrep=as.factor(myrep)
d_RMA=cbind(d_RMA, myrep)

y=summary(d_RMA$myrep)
percent=paste(round((summary(d_RMA$myrep)/nrow(d_RMA))*100,digits=1),'%',sep='')
percent2=sapply(percent, function(x)
  {
  if(as.numeric(substr(x,1, nchar(x)-1)) <5)
  {
    return(" ")
  }
  else return(x)
  })
pos=cumsum(y)-y/2
p1 = ggplot(data=d_RMA, aes(x=factor(1), fill = myrep))

p1 = p1 + geom_bar(width = 1,stat='count')+ylab("")+
theme(panel.background = element_blank(),
      axis.text.x = element_blank(),
      axis.text.y = element_blank(),
      panel.grid.major= element_line(colour="white"),
      panel.grid.minor= element_line(colour="white"),
```

```
      legend.key.size=unit(0.5,"cm"), legend.text = element_text(size=10),
      plot.title=element_text(size=14))+
  coord_polar(theta="y")+
  ggtitle("Repartition of probesets by repertoire \n on data before filtering")+
annotate(geom = "text", y = pos, x = 1.2, label = percent2,size=6)
```

```
#filtered data composition
y=summary(d_fltr$myrep)
percent=paste(round((summary(d_fltr$myrep)/nrow(d_fltr))*100,digits=1),'%',sep='')
#avoid overlapping percentage on graphics
percent2=sapply(percent, function(x)
  {
  if(as.numeric(substr(x,1, nchar(x)-1)) <5)
  {
    return(" ")
  }
  else return(x)
  })

pos=cumsum(y)-y/2
p2 = ggplot(data=d_fltr, aes(x=factor(1), fill = myrep))

p2 = p2 + geom_bar(width = 1,stat='count')+ylab("")+
theme(panel.background = element_blank(),
      axis.text.x = element_blank(),
      axis.text.y = element_blank(),
      panel.grid.major= element_line(colour="white"),
      panel.grid.minor= element_line(colour="white"),
      legend.key.size=unit(0.5,"cm"), legend.text = element_text(size=10),
      plot.title=element_text(size=14))+
  coord_polar(theta="y")+
  ggtitle("Repartition of probesets by repertoire \n on filtered data")+
annotate(geom = "text", y = pos, x = 1.2, label = percent2,size=6)

grid.arrange(p1,p2, ncol=2)
```

Compared with the composition on the chip, the filtered data have a lower proportion of MALR and HERVs Dfam and have higher proportion of genes and HERVs proto (Figure 13).

Family approach

In this section we focused on proto repertoire. It is the repertoire composed of HERVs which are well annotated. We have among others the information of their family.

```
nbcores=4
colseq=seq(1,ncol(exprs_RMA))
rowseq=seq(1,nrow(exprs_RMA))
thr=3.5

f_data=apply(exprs_RMA, c(1,2), FUN=function(x)
  {
  if(x>=thr){return(TRUE)}
```

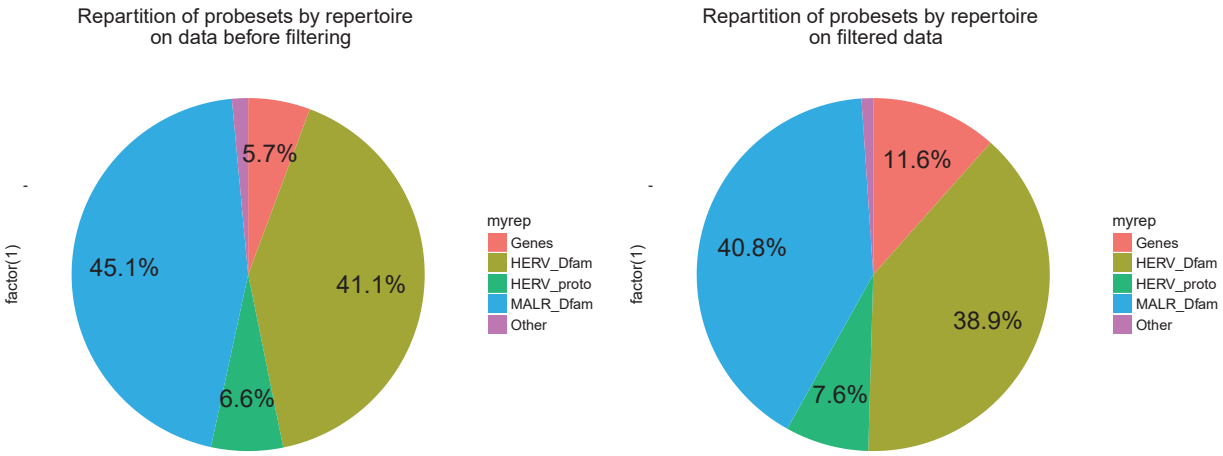


Figure 13: **Repartition of probesets by repertoire before (on the left) and after filtering step (on the right).**

```
  else return(FALSE)
  })

colnames(f_data)=covdesc$haiday

j=getAnnotMatches(rownames(exprs_RMA))
ann_chip=ann[j[which(!(is.na(j)))],]
count_all=ann_chip %>% group_by(repertoire) %>% summarise(count_all=n())
count_all=na.omit(count_all)

cpt_over_thr_samples=mclapply(c(1:ncol(f_data)),mc.cores=nbcores, FUN=function(y)
  {
  dat=f_data[,y][which(f_data[,y]==TRUE)]
  len=length(dat)
  j=getAnnotMatches(names(dat))
  length(which((is.na(j))))

  ann_over_thr=ann[j[which(!(is.na(j)))],]
  count_over_thr=ann_over_thr %>% group_by(repertoire) %>% summarise(count=n())
  count_over_thr=na.omit(count_over_thr)
  count_over_thr$count_all=count_all$count_all
  count_over_thr$prop = round(count_over_thr$count/count_over_thr$count_all, 6)
  count_over_thr$percentage=count_over_thr$prop*100
  count_over_thr$rep=factor(count_over_thr$repertoire,
                            levels=sort(count_over_thr$repertoire))
  #removing genes
  count_over_thr=count_over_thr[which(count_over_thr$repertoire!='U133'),]
  count_over_thr=count_over_thr[which(count_over_thr$repertoire!='HTA'),]
  count_over_thr=count_over_thr[which(count_over_thr$repertoire!='Opti'),]

  ann_proto_all = filter(ann_chip, repertoire=="HERV_proto")
  proto_all=ann_proto_all %>% group_by(simple_family) %>% summarise(count_all=n())
  ann_proto_over_thr=filter(ann_over_thr, repertoire=="HERV_proto")
```

```
  count_proto_over_thr=ann_proto_over_thr %>% group_by(simple_family) %>%
    summarise(count=n())
  count_proto_over_thr$count_all=proto_all$count_all
  count_proto_over_thr$prop =
    round(count_proto_over_thr$count/count_proto_over_thr$count_all, 6)
  count_proto_over_thr$percentage=count_proto_over_thr$prop*100
  count_proto_over_thr$sfam=factor(count_proto_over_thr$simple_family,
                                   levels=sort(count_proto_over_thr$simple_family))

  #nb family + dfam
  count_dfam_over_thr = rbind(filter(count_over_thr, rep=="HERV_Dfam"),
                              filter(count_over_thr, rep=="MALR_Dfam"),
                              filter(count_over_thr, rep=="Centromeriques"))
  colnames(count_dfam_over_thr)=colnames(count_proto_over_thr)
  cpt_over_thr=rbind(count_dfam_over_thr, count_proto_over_thr)
  cpt_over_thr$id_patient=covdesc$Identifiant_patients[y]
  cpt_over_thr$subgroup=covdesc$haiday[y]
  return(cpt_over_thr)
  })
names(cpt_over_thr_samples)=covdesc$Identifiant_patients
```

```
icdf=1
output_dir = paste0("output/Analysis/cdf",icdf,"/")
#load(paste0('data/cdf',icdf,'/cpt_over_thr_samples.RData'))

df_cpt=do.call("rbind",cpt_over_thr_samples)
df_cpt$subgroup=as.factor(df_cpt$subgroup)

df_subgroups=df_cpt %>% group_by(subgroup, simple_family) %>%
  summarise(count=round(median(count),0), count_all=round(median(count_all),0))
df_subgroups$percentage=round(df_subgroups$count/df_subgroups$count_all*100,2)

df_subgroups$class_retro=sapply(as.character(df_subgroups$simple_family), FUN=function(x)
  {
    if(x=="ERV9" || x=="HARLEQUIN" || x=="HERV30" || x=="HERV46" || x=="HERV-ADP" ||
       x=="HERV-E4.1" || x=="HERV-F" || x=="HERV-Fb" || x=="HERV-Fc1" || x=="HERV-Fc2" ||
       x=="HERV-FRD" || x=="HERV-H" || x=="HERV-HS49C23" || x=="HERV-I" || x=="HERV-P" ||
       x=="HERV-Pb" || x=="HERV-R" || x=="HERV-Rb" || x=="HERV-T" || x=="HERV-V" ||
       x=="HERV-W" || x=="HERV-XA34" || x=="MER52A" || x=="MER84" || x=="PABL-A" ||
       x=="PRIMA4" || x=="PRIMA41" || x=="RRHERVI")
    {
      return("Gammaretrovirus")
    }
  else if(x=="HERV-K14C" || x=="HML-1" || x=="HML-2" || x== "Centromeriques"||
          x=="HML-3" || x=="HML-4" || x=="HML-5" || x=="HML-6" || x=="HML-7" ||
          x=="HML-8" || x=="HML-9" || x=="HML-10")
  {
    return("Betaretrovirus")
  }
  else if (x=="HERV18" || x=="HERV-L" || x=="HERV-L66")
  {
    return("Spuma_epsilon_like")
  }
```

```
  else return("Other")
  })

df_subgroups$simple_family=
  factor(df_subgroups$simple_family,
         levels=c("ERV9" , "HARLEQUIN", "HERV30", "HERV46", "HERV-ADP", "HERV-E4.1",
                  "HERV-F", "HERV-Fb", "HERV-Fc1", "HERV-Fc2", "HERV-FRD", "HERV-H",
                  "HERV-HS49C23", "HERV-I", "HERV-P", "HERV-Pb", "HERV-R", "HERV-Rb",
                  "HERV-T", "HERV-V","HERV-W", "HERV-XA34", "MER52A", "MER84", "PABL-A",
                  "PRIMA4", "PRIMA41", "RRHERVI", "HERV-K14C", "HML-1", "HML-2",
                  "Centromeriques", "HML-3", "HML-4", "HML-5" , "HML-6" , "HML-7" ,
                  "HML-8" , "HML-9" , "HML-10", "HERV18", "HERV-L", "HERV-L66",
                  "HERV_Dfam", "MALR_Dfam"))
df_subgroups$class_retro=
  factor(df_subgroups$class_retro,
         levels=c("Betaretrovirus", "Gammaretrovirus", "Spuma_epsilon_like", "Other"))

g = ggplot(data=df_subgroups, aes(x=simple_family, y=percentage,
                                  fill=class_retro, colour=class_retro)) +
  facet_wrap(~subgroup, ncol=2, dir="v")+
  theme(panel.background = element_blank(),
        axis.text.x = element_text(size=11, colour='black',angle=45, hjust=1),
        panel.border = element_rect(color = "black", fill = NA, size = 1),
        axis.text.y = element_text(size=15, colour='black'),
        axis.line=element_line(colour="black"),
        panel.grid.major= element_line(colour="grey"),
        panel.grid.minor= element_line(colour="white"),
        strip.text=element_text(size=14),
        legend.position="none")+
  geom_bar(stat='identity') +
  geom_text(aes(x=simple_family, y=percentage, label=count),
            colour="black", size=6, vjust=0.5, hjust= 1,  angle=90)+
  xlab("Family") + ylab("Percentage (%)")+
  ggtitle(
paste0('Count and proportion of probesets that passed the filter by family + Dfam\n cdf: ',
          icdf))
```

Proportions and counts of proto by group

- No_D1 represents the samples at D1 for patients without HAI.
- No_D3 represents the samples at D3 for patients without HAI.
- No_D6 represents the samples at D6 for patients without HAI.
- Yes_D1 represents the samples at D1 for patients with HAI.
- Yes_D3 represents the samples at D3 for patients with HAI.
- Yes_D6 represents the samples at D6 for patients with HAI.

The graphic (Figure 14) represents the proportion and the count of proto that passed the intensity threshold compared to the total of protos.

```
df_proto=df_subgroups %>% filter(simple_family!="HERV_Dfam" & simple_family!="MALR_Dfam" &
                                   simple_family!="Centromeriques") %>%
  group_by(subgroup) %>% summarise(count = sum(count))
```

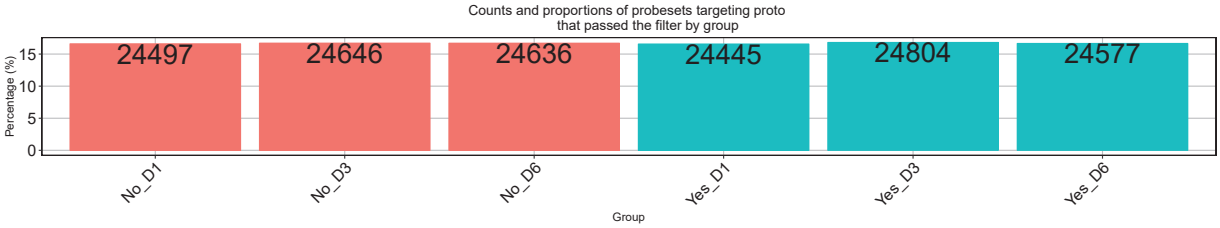


Figure 14: **Counts and proportions of proto probesets over intensity threshold**. On red the groups with no HAI, on blue the HAI groups.

```
HAI=sapply(df_proto$subgroup, FUN=function(x)
  {
    if(substr(x,1,2)=="No")
    {
      return("No")
    }
    else return("Yes")
  })
df_proto$HAI=HAI
df_proto$percentage=df_proto$count/sum(df_proto$count)*100

gProto = ggplot(data=df_proto, aes(x=subgroup, y=percentage,  colour=HAI, fill=HAI)) +
  theme(panel.background = element_blank(),
        axis.text.x = element_text(size=15, colour='black',angle=45, hjust=1),
        panel.border = element_rect(color = "black", fill = NA, size = 1),

        axis.text.y = element_text(size=15, colour='black'),
        axis.line=element_line(colour="black"),
        panel.grid.major= element_line(colour="grey"),
        panel.grid.minor= element_line(colour="white"), legend.position="none")+
  geom_bar(stat='identity') +
  geom_text(aes(x=subgroup, y=percentage,
                label=round(count,0)), colour="black", size=9, vjust=1)+
  xlab("Group") + ylab("Percentage (%)")+
  ggtitle('Counts and proportions of probesets targeting proto
          that passed the filter by group')
gProto
```

Proportions and counts by group for each family of proto

```
g
```

On the figures 14 & 15, we did not detect differences between groups. Chi squared test results are not shown but we could not detect any difference between groups.

```
#table made by hand from barplots summary figures
data=read.table("../data/cdf1/summary_tab.txt", sep='\t', header=TRUE )

####first calculate chisq on data compared to theorical
```

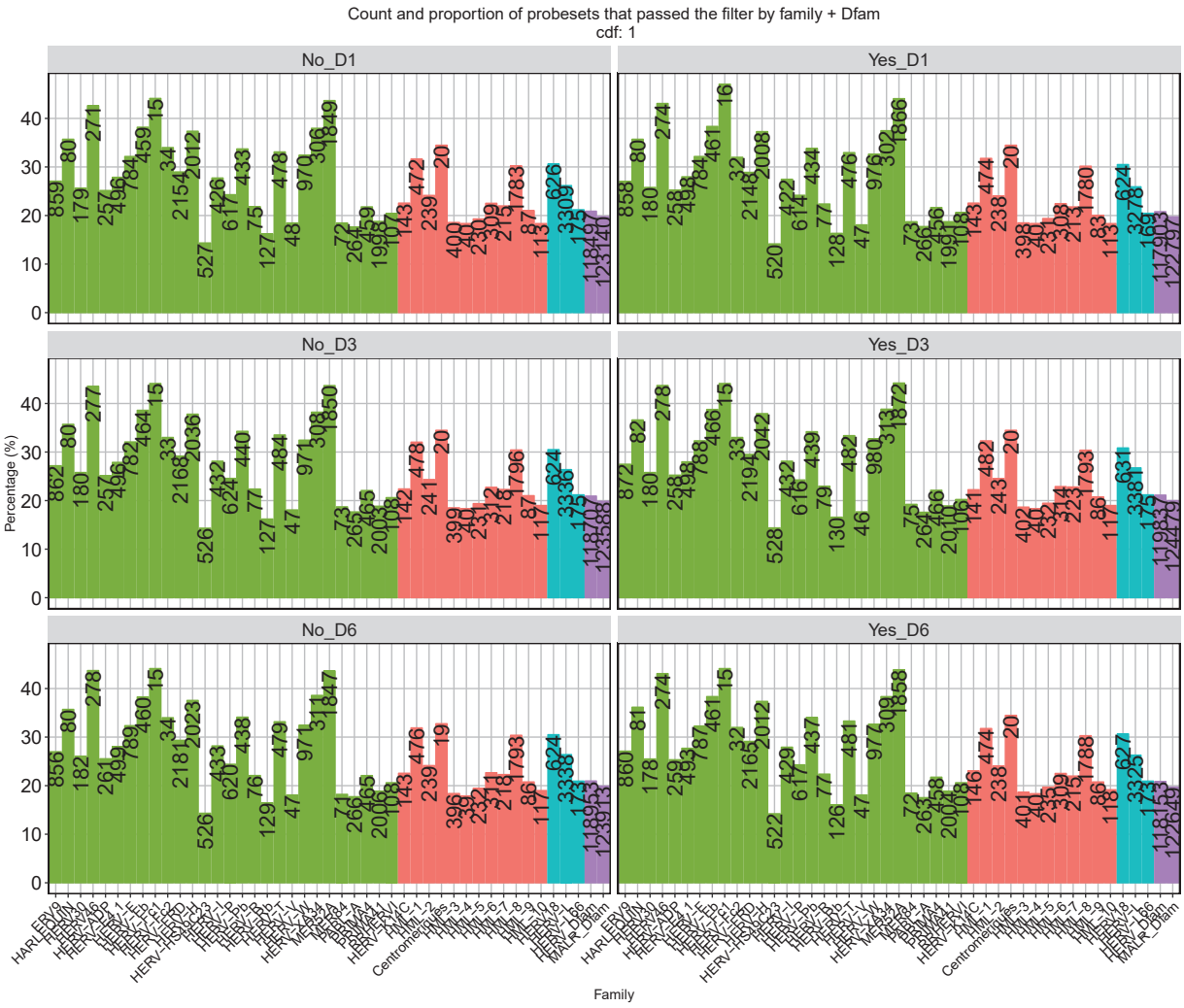


Figure 15: **HERV families over intensity threshold for each group of patients**. The y-axis represents the proportion of probesets that passed the filter compared to the total of probesets avalaible on the chip, for each proto family, for each group of patients. The numbers represent the counts. The colors represent the classes of retroviruses (from left to right: Gammaretrovirus, Betaretrovirus, Spuma / epsilon-like and Other).

```
#tab=matrix(c(863, 3172-863,23632, 83482-23632), ncol=2); chisq.test(tab)
ch_nod1=sapply(seq(1,nrow(data)),FUN=function(x)
  {
    fam_overthr = data[x,"No_D1"]
    fam_underthr = data[x,"Total_puce"]-fam_overthr
    reste_overthr = data[x,"reste_No_D1"]
    reste_underthr = data[x,"reste_puce"] - reste_overthr
    tab=matrix(c(fam_overthr,fam_underthr, reste_overthr, reste_underthr), ncol=2)
    chsq=chisq.test((tab))
    return(chsq$p.value)
  })
ch_nod3=sapply(seq(1,nrow(data)),FUN=function(x)
{
  fam_overthr = data[x,"No_D3"]
  fam_underthr = data[x,"Total_puce"]-fam_overthr
  reste_overthr = data[x,"reste_No_D3"]
  reste_underthr = data[x,"reste_puce"] - reste_overthr
  tab=matrix(c(fam_overthr,fam_underthr, reste_overthr, reste_underthr), ncol=2)
  chsq=chisq.test((tab))
  return(chsq$p.value)
})
ch_nod6=sapply(seq(1,nrow(data)),FUN=function(x)
{
  fam_overthr = data[x,"No_D6"]
  fam_underthr = data[x,"Total_puce"]-fam_overthr
  reste_overthr = data[x,"reste_No_D6"]
  reste_underthr = data[x,"reste_puce"] - reste_overthr
  tab=matrix(c(fam_overthr,fam_underthr, reste_overthr, reste_underthr), ncol=2)
  chsq=chisq.test((tab))
  return(chsq$p.value)
})
ch_yesd1=sapply(seq(1,nrow(data)),FUN=function(x)
{
  fam_overthr = data[x,"Yes_D1"]
  fam_underthr = data[x,"Total_puce"]-fam_overthr
  reste_overthr = data[x,"reste_Yes_D1"]
  reste_underthr = data[x,"reste_puce"] - reste_overthr
  tab=matrix(c(fam_overthr,fam_underthr, reste_overthr, reste_underthr), ncol=2)
  chsq=chisq.test((tab))
  return(chsq$p.value)
})
ch_yesd3=sapply(seq(1,nrow(data)),FUN=function(x)
{
  fam_overthr = data[x,"Yes_D3"]
  fam_underthr = data[x,"Total_puce"]-fam_overthr
  reste_overthr = data[x,"reste_Yes_D3"]
  reste_underthr = data[x,"reste_puce"] - reste_overthr
  tab=matrix(c(fam_overthr,fam_underthr, reste_overthr, reste_underthr), ncol=2)
  chsq=chisq.test((tab))
  return(chsq$p.value)
})
ch_yesd6=sapply(seq(1,nrow(data)),FUN=function(x)
{
```

```
  fam_overthr = data[x,"Yes_D6"]
  fam_underthr = data[x,"Total_puce"]-fam_overthr
  reste_overthr = data[x,"reste_Yes_D6"]
  reste_underthr = data[x,"reste_puce"] - reste_overthr
  tab=matrix(c(fam_overthr,fam_underthr, reste_overthr, reste_underthr), ncol=2)
  chsq=chisq.test((tab))
  return(chsq$p.value)
})
data$Chisq_No_D1=ch_nod1
data$Chisq_No_D3=ch_nod3
data$Chisq_No_D6=ch_nod6
data$Chisq_Yes_D1=ch_yesd1
data$Chisq_Yes_D3=ch_yesd3
data$Chisq_Yes_D6=ch_yesd6
```

Enrichments by family

To go further we computed enrichment scores for HERVs families compared to theorical effectives per family. (i.e. the number of probesets targetting each proto family on the chip)

The heatmap (Figure 16) shows the enrichment of each family for each group of patient compared to the distribution of the chip. The more intense the color intense is, the lower p.value of chi-squared test. It is blue when the family is depleted compared to theorical distribution, and in red when the family is enriched.

```
limits=50
rownames(data)=as.character(as.vector(data$FAMILLE))
htp = -log10(data[,36:41])

htp[which(data$ratio_No_D1 < 1),1] = -1*htp[which(data$ratio_No_D1 < 1 ),1]
htp[which(data$ratio_No_D3 < 1),2] = -1*htp[which(data$ratio_No_D3 < 1 ),2]
htp[which(data$ratio_No_D6 < 1),3] = -1*htp[which(data$ratio_No_D6 < 1 ),3]
htp[which(data$ratio_Yes_D1 < 1),4] = -1*htp[which(data$ratio_Yes_D1 < 1 ),4]
htp[which(data$ratio_Yes_D3 < 1),5] = -1*htp[which(data$ratio_Yes_D3 < 1 ),5]
htp[which(data$ratio_Yes_D6 < 1),6] = -1*htp[which(data$ratio_Yes_D6 < 1 ),6]

htp[which(htp[,1]>limits),1]=limits
htp[which(htp[,2]>limits),2]=limits
htp[which(htp[,3]>limits),3]=limits
htp[which(htp[,4]>limits),4]=limits
htp[which(htp[,5]>limits),5]=limits
htp[which(htp[,6]>limits),6]=limits

htp[which(htp[,1]< -limits),1]=-limits
htp[which(htp[,2]< -limits),2]=-limits
htp[which(htp[,3]< -limits),3]=-limits
htp[which(htp[,4]< -limits),4]=-limits
htp[which(htp[,5]< -limits),5]=-limits
htp[which(htp[,6]< -limits),6]=-limits

pheatmap(as.matrix(htp),
         color = colorRampPalette(c("navy", "white", "firebrick3"))(2*limits),
         breaks = seq(-limits,limits,1),cluster_rows = F,cluster_cols = F, fontsize = 14)
```

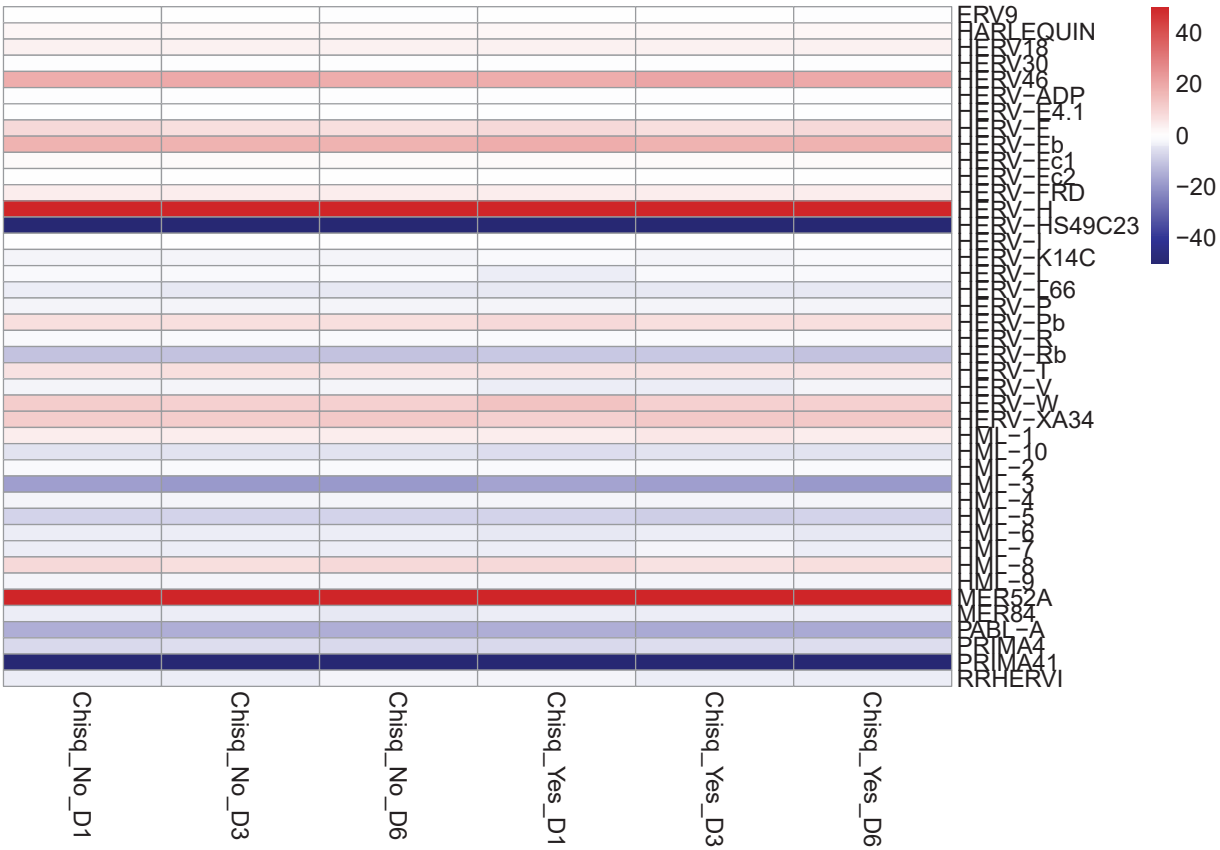


Figure 16: **Enrichments**. Each line represent a family of HERV protos, each column the chi-squared test made for each group of patients compared to the theorical distribution (Probesets available on the chip). The more intense the color intense is, the lower p.value of chi-squared test. It is blue when the family is depleted compared to the theorical distribution, and it is red when the family is enriched.

As we can see, there was no difference of family representation between groups. But some families were specifically over represented in filtered data, no matter the group, like HERV-H and MER52A. Whereas some families were under represented, like HERV HSC49C23 and PRIMA41.

Chromosome approach

We wanted to verify if there are bias of expression by chromomsome, on all dataset or on protos only.

```
#all dataset
i=getAnnotMatches(rownames(fltr_RMA))
ann_fltr=ann[i[which(!is.na(i))],]

j=getAnnotMatches(rownames(exprs_RMA))
ann_chip=ann[j[which(!is.na(j))],]

df_chip=ann_chip %>% group_by(CHR) %>% summarize(count_chip=n())
df_fltr=ann_fltr %>% group_by(CHR) %>% summarize(count_fltr=n())
#the pb here is that we have chromosome annotation we want to simplify
#we remove all annot which are not 1 to 22, and X & Y
df_chip=df_chip[c(1:19, 27:29, 53,54),]
df_fltr=df_fltr[c(1:19, 27:29, 44,45),]

df=cbind(df_chip, count_fltr = df_fltr$count_fltr)
df$prop=df$count_fltr/df$count_chip

df$CHR=factor(df$CHR, levels=c(as.character(seq(1,22)), "X", "Y"))

p1=ggplot(df, aes(x=CHR, y=prop))+
  geom_bar(stat='identity')+
  theme(panel.background = element_blank(),
        axis.text.x = element_text(size=15, colour='black'),
        panel.border = element_rect(color = "black", fill = NA, size = 1),
        axis.text.y = element_text(size=15, colour='black'),
        axis.line=element_line(colour="black"),
        panel.grid.major= element_line(colour="grey"),
        panel.grid.minor= element_line(colour="white"))+
  geom_text(aes(x=CHR, y=prop, label=count_fltr),
            colour="white", size=8, vjust=0.5, hjust= 1,  angle=90)+
  xlab("Chromosome") + ylab("Proportions")+
  ggtitle('Counts and proportions of probesets that passed the filter by chromosome')
```

```
#only with proto
ann_p_chip = ann_chip %>% filter(repertoire=="HERV_proto")
ann_p_fltr = ann_fltr %>% filter(repertoire=="HERV_proto")

df_chip=ann_p_chip %>% group_by(CHR) %>% summarize(count_chip=n())
df_fltr=ann_p_fltr %>% group_by(CHR) %>% summarize(count_fltr=n())
#the pb here is that we have chromosome annotation we want to simplify
#we remove all annot which are not 1 to 22, and X & Y


df=cbind(df_chip, count_fltr = df_fltr$count_fltr)
df$prop=df$count_fltr/df$count_chip
```

```
df$CHR=factor(df$CHR, levels=c(as.character(seq(1,22)), "X", "Y"))

p2=ggplot(df, aes(x=CHR, y=prop))+
  geom_bar(stat='identity')+
  theme(panel.background = element_blank(),
        axis.text.x = element_text(size=15, colour='black'),
        panel.border = element_rect(color = "black", fill = NA, size = 1),
        axis.text.y = element_text(size=15, colour='black'),
        axis.line=element_line(colour="black"),
        panel.grid.major= element_line(colour="grey"),
        panel.grid.minor= element_line(colour="white"))+
  geom_text(aes(x=CHR, y=prop, label=count_fltr),
            colour="white", size=8, vjust=0.5, hjust= 1,  angle=90)+
  xlab("Chromosome") + ylab("Proportions")+
  ggtitle('Count and proportion of proto probesets that passed the filter by chromosome')

grid.arrange(p1,p2, ncol=1)
```

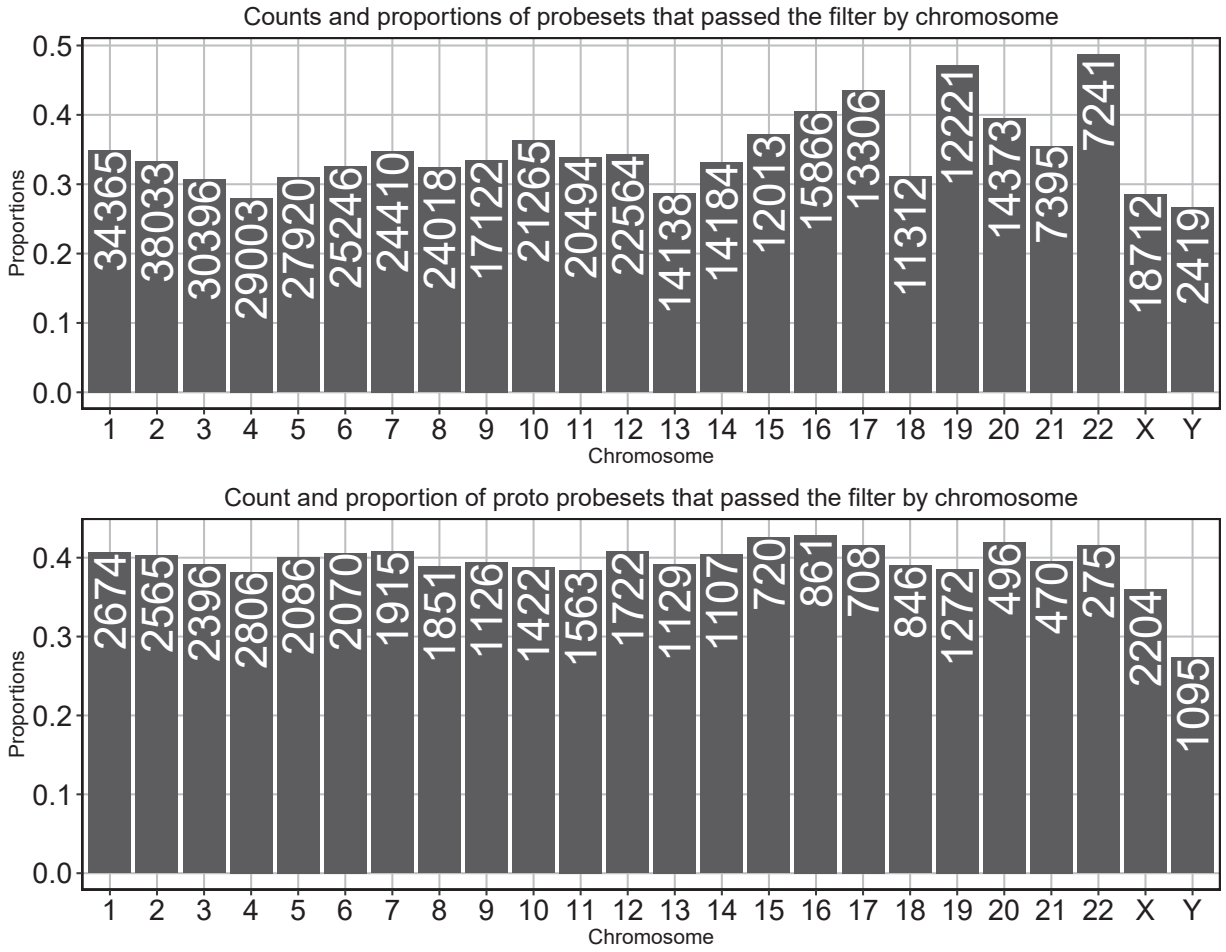


Figure 17: **Counts and proportion over intensity thresholds on chromosomes**.

On the figures, the proportion of probesets that passed the intensity filter was arround 0.3 - 0.5 . We could

not detect a chromosome where the probesets are clearly more or less expressed than the others, on all dataset or on protos repertoire only. Hervs protos are globally more expressed than the total dataset, mostly represented by HERVs and MALRs Dfam repertoires.

Supervised analysis

Differential expression

In this section, we wanted to find all probesets for which expression is different between conditions. We were interested in several endpoints:

- HAI
- Ratio D3/D1 of cd74
- Mortality

For all the endpoints, we choose a log Fold Change threshold equal to 1 and an adjusted p.value (by Benjamini-Hochberg method) threshold equal to 0.05.

HAI

For HAI endpoint, we wanted to study the differential expression between:

- HAI vs NO HAI (HAI_NOHAI)
- HAI vs NO HAI at D1 (HAID1_NOHAID1)
- HAI vs NO HAI at D3 (HAID3_NOHAID3)
- HAI vs NO HAI at D6 (HAID6_NOHAID6)

```
f=factor(covdesc$haiday)
design=model.matrix(~0+f)
colnames(design)=levels(f)
load('../data/cdf1/corfit.RData') #duplicateCorrelation, made on bmx cluster
fit <- lmFit(fltr_RMA,design,block=covdesc$patients,
             correlation=corfit$consensus)

#model 1
cont.wt=makeContrasts(
  HAI_NOHAI = (Yes_D1 + Yes_D3 + Yes_D6) - (No_D1 + No_D3 + No_D6),
  HAID1_NOHAID1 = Yes_D1 - No_D1,
  HAID3_NOHAID3 = Yes_D3 - No_D3,
  HAID6_NOHAID6 = Yes_D6 - No_D6,
  levels=design
)

fit2=contrasts.fit(fit,cont.wt)
fit2=eBayes(fit2)
```

```
fc_thresh = 1
pval_thresh=0.05

results = decideTests(fit2, p.value=pval_thresh,lfc=fc_thresh, adjust.method="fdr")
```

```
vennDiagram(results, include=c("up","down"),
            circle.col=c("yellow","green", "blue", "red"),
            counts.col=c("darkgreen","darkred"),
             cex=c(1,0.7,0.5))
```

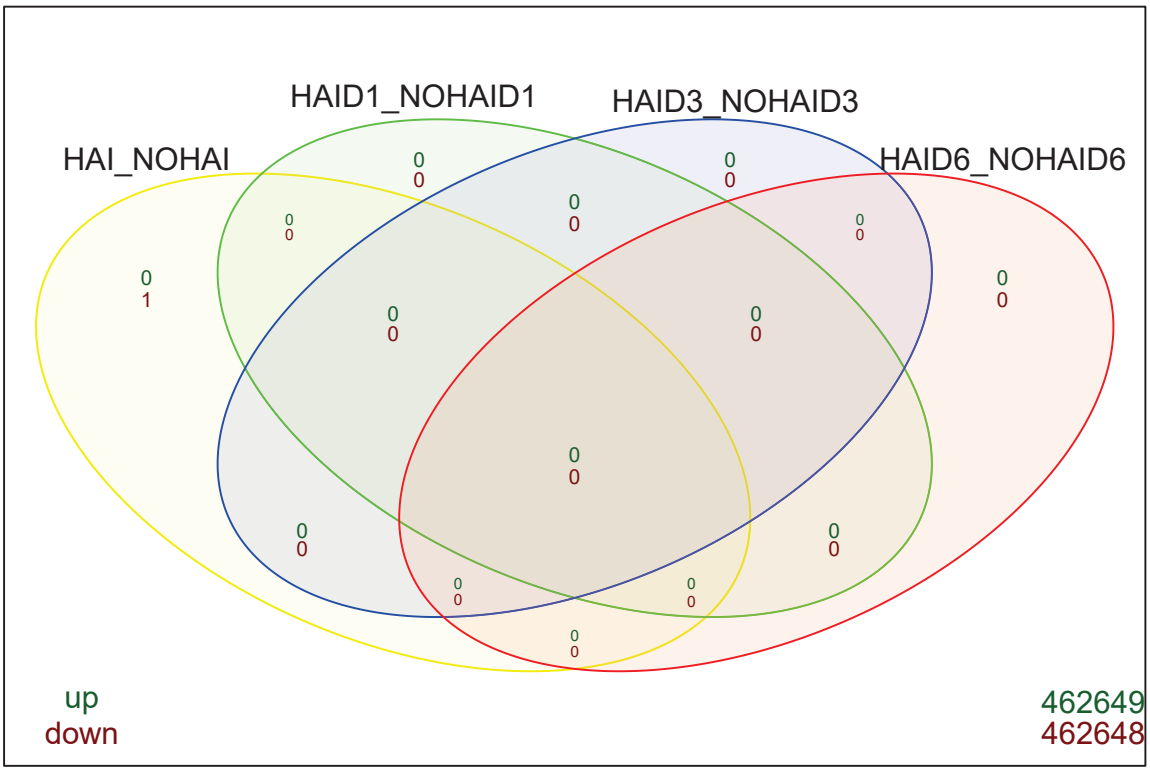


Figure 18: **Venn diagram for HAI endpoint**.

```
top_all=topTable(fit2,adjust="fdr",number=nrow(fltr_RMA))

top1=topTable(fit2,adjust="fdr",number=nrow(fltr_RMA), coef=1)
#coef is for choosing p_value of corresponding contrast
#the consequence is that we don't have logfc for all questions, so i create all data
top2=topTable(fit2,adjust="fdr",number=nrow(fltr_RMA), coef=2)
top3=topTable(fit2,adjust="fdr",number=nrow(fltr_RMA), coef=3)
top4=topTable(fit2,adjust="fdr",number=nrow(fltr_RMA), coef=4)


# regrouping of each adjusted pvalue for each contrast in the same data
i=match(rownames(top_all), rownames(top1))
top_all$adj.P.coef1=top1$adj.P.Val[i]
i=match(rownames(top_all), rownames(top2))
top_all$adj.P.coef2=top2$adj.P.Val[i]
i=match(rownames(top_all), rownames(top3))
top_all$adj.P.coef3=top3$adj.P.Val[i]
i=match(rownames(top_all), rownames(top4))
top_all$adj.P.coef4=top4$adj.P.Val[i]
```

```
i=getAnnotMatches(rownames(top_all))
aliases=ann$ALIAS[i]
top_all$rep=ann$repertoire[i]
top_all$alias=aliases
```

We could not detect differentially expressed probesets with HAI endpoint (The only detected probeset was a long non codant RNA and found for the question HAI vs NOHAI regrouping all timepoints, which is meaningless compared to other questions. So we ignired it).

D3/D1 CD74 ratio

We selected Youden value corresponding to the ratio D3/D1 of CD74 on RTqPCR made on the MIP rea cohort in previous analysis (ratio = 1.23). The population has been splitted into 2 categories. Individuals with cd74 ratio D3/D1 over 1.23 are "High" (proxy of non immunosuppressed patients). Those with ratio D3/D1 under 1.23 are FALSE (or ratioinf, proxy of immunosuppressed patients).

For cd74 ratio D3/D1 endpoint, we want to study the differential expression between:

- Low ratio vs High ratio
- Low ratio vs High ratio at D1
- Low ratio vs High ratio at D3
- Low ratio vs High ratio at D6

```
f=factor(paste(covdesc$youden_cd74, covdesc$Day2, sep='_'))
design=model.matrix(~0+f)
colnames(design)=levels(f)
load(paste0('../data/cdf',icdf,'/corfit_youden_cd74.RData'))
fit <- lmFit(fltr_RMA,design,block=covdesc$patients, correlation=corfit$consensus)

cont.wt=makeContrasts(
  LowD1_HighD1 = FALSE_D1 - TRUE_D1,
  LowD3_HighD3 = FALSE_D3 - TRUE_D3,
  LowD6_HighD6 = FALSE_D6 - TRUE_D6,
  Low_High = (FALSE_D1+FALSE_D3+FALSE_D6) - (TRUE_D1+TRUE_D3+TRUE_D6),
  levels=design
)
fit2=contrasts.fit(fit,cont.wt)
fit2=eBayes(fit2)
```

```
results = decideTests(fit2, p.value=0.05,lfc=1, adjust.method="fdr")
```

```
vennDiagram(results, include=c("up","down"),
            circle.col=c("yellow","green", "blue", "red"),
            counts.col=c("darkgreen","darkred"),
             cex=c(1,0.7,0.5))
```

For this endpoint, and with these thresholds, we detected 12 and 103 up-regulated probesets between patients with ratio <1.23 and patients with ratio > 1.23 at D1 and D3 respectively. It means that these probesets are more expressed in immunosuppressed patients than the others. We did not detect any down regulated probesets (Figure 19).

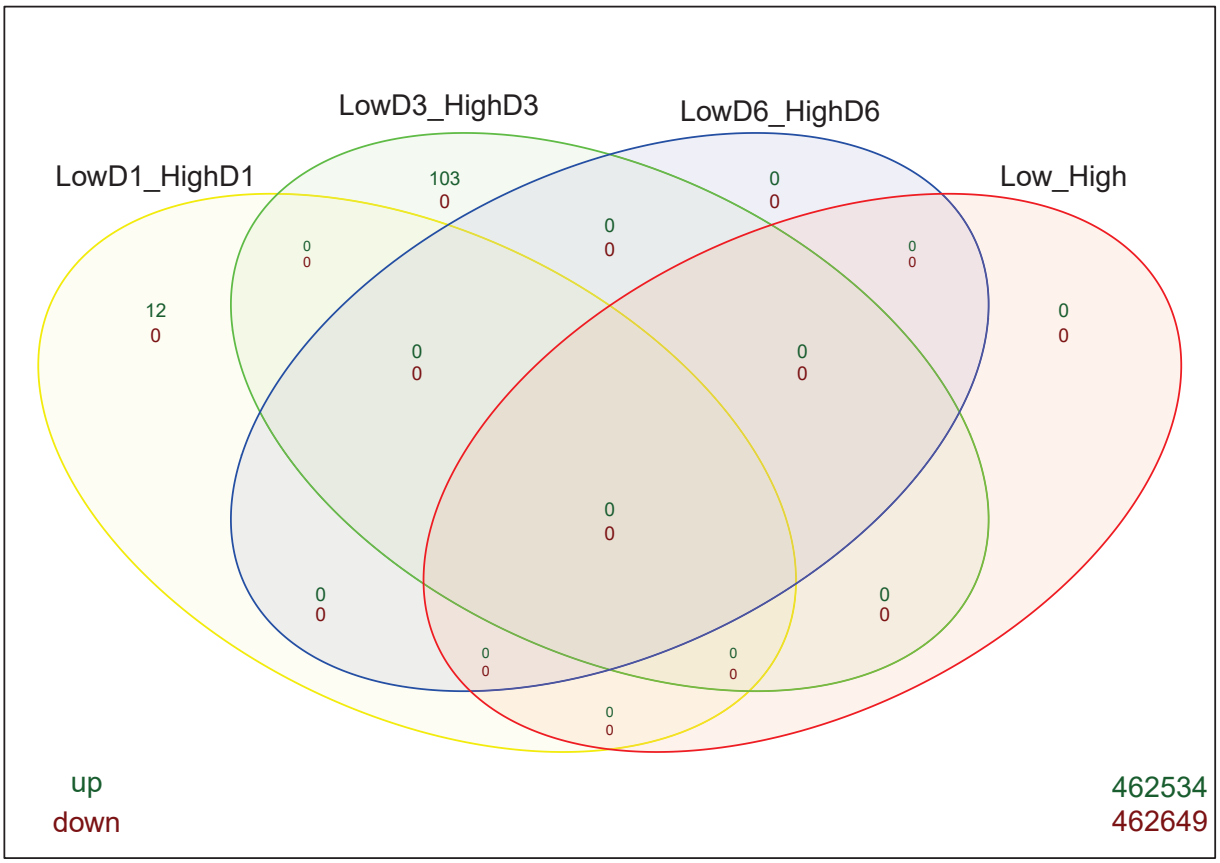


Figure 19: **Venn diagram for CD74 endpoint**. The numbers in dark green represent the probesets which are up-regulated for the corresponding question. In dark red the down-regulated probesets.

```
top_all=topTable(fit2,adjust="fdr",number=nrow(fltr_RMA))
top1=topTable(fit2,adjust="fdr",number=nrow(fltr_RMA), coef=1)
#coef is for choosing p_value of corresponding contrast
#the consequence is that we don't have logfc for all questions, so i create all data
top2=topTable(fit2,adjust="fdr",number=nrow(fltr_RMA), coef=2)
top3=topTable(fit2,adjust="fdr",number=nrow(fltr_RMA), coef=3)
top4=topTable(fit2,adjust="fdr",number=nrow(fltr_RMA), coef=4)

# regrouping of each adjusted pvalue for each contrast in the same data
i=match(rownames(top_all), rownames(top1))
top_all$adj.P.coef1=top1$adj.P.Val[i]
i=match(rownames(top_all), rownames(top2))
top_all$adj.P.coef2=top2$adj.P.Val[i]
i=match(rownames(top_all), rownames(top3))
top_all$adj.P.coef3=top3$adj.P.Val[i]
i=match(rownames(top_all), rownames(top4))
top_all$adj.P.coef4=top4$adj.P.Val[i]

i=getAnnotMatches(rownames(top_all))
aliases=ann$ALIAS[i]
top_all$rep=ann$repertoire[i]
top_all$alias=aliases
chr=ann$CHR[i]
strt=ann$CHRLOC[i]
end=ann$CHRLOCEND[i]
top_all$coords_19=paste0("chr",chr, ':', strt, '-', end)
```

```
pval_thresh=0.05
lfc_thresh=1
top_signif=top_all[which((abs(top_all$LowD1_HighD1) > lfc_thresh &
                                        top_all$adj.P.coef1 < pval_thresh) |
                                      (abs(top_all$LowD3_HighD3) > lfc_thresh &
                                        top_all$adj.P.coef2 < pval_thresh) |
                                      (abs(top_all$LowD6_HighD6) > lfc_thresh &
                                        top_all$adj.P.coef3 < pval_thresh) |
                                      (abs(top_all$Low_High) > lfc_thresh &
                                        top_all$adj.P.coef4 < pval_thresh)
                            ),]
top_signif_herv=top_signif[which(top_signif$rep!="HTA" &
                                   top_signif$rep!="U133" & top_signif$rep!="Opti"),]
```

Description of the 115 differentially expressed probesets:

Number of probesets targetting genes: 61

```
#Number of probesets targetting genes
nrow(top_signif[which(top_signif$rep=="U133" | top_signif$rep=="Opti" |
                        top_signif$rep=="HTA" ),])
```

Number and list of Genes (not probesets): 10

```
# length(unique(top_signif$alias[which(top_signif$rep=="U133" | top_signif$rep=="Opti" |
```

```
##  [1] "IL18R1"  "MOV10"   "IL1R2"   "IL18RAP" "STAT4"   "OAS2"    "OAS3"
##  [8] "CYP1B1"  "IFIH1"   "IL1R1"
```

Number of HERVs/MALRs probesets and elements respectively : 54 & 31

```
#Number and list of hervs/malr probesets and elements
nrow(top_signif_herv)
length(unique(substr(rownames(top_signif_herv), 1,7)))
```

There are 10 distinct genes and 31 distinct HERVs or MALRs elements among the 115 probesets.

To have a better idea of eventual colocalizations between hervs and genes, we would like to regroup the closest ones. The output shows the HERVs or MALRs which are close (<10kb) to a gene.

```
## 021460102-HERV0599uL_st 021460302-MALR1004uL_at 021460605-HERV0882cL_at
##                "IL18R1"                 "IL18R1"               "IL18RAP"
## 021460101-HERV0599uL_st 021455702-MALR1018uL_at 021455602-MALR1012uL_at
##                "IL18R1"                  "IL1R2"                  "IL1R2"
## 021460608-HERV0882cL_at 021455602-MALR1012uL_st 021460608-HERV0882cL_st
##               "IL18RAP"                  "IL1R2"               "IL18RAP"
## 021455501-MALR1020cL_st 021456001-MALR1017uL_at 021455601-MALR1012uL_at
##                 "IL1R2"                  "IL1R2"                  "IL1R2"
## 021456001-MALR1017uL_st 021460802-MALR1041uL_at 021460502-HERV0566uL_at
##                 "IL1R2"                "IL18RAP"                 "IL18R1"
## 021460702-MALR1014uL_at 021455601-MALR1012uL_st 021460601-HERV0882cI_st
##               "IL18RAP"                  "IL1R2"               "IL18RAP"
## 021460802-MALR1041uL_st 021460502-HERV0566uL_st 021460605-HERV0882cL_st
##               "IL18RAP"                 "IL18R1"               "IL18RAP"
## 021460601-HERV0882cI_at 021460607-HERV0882cL_st
##               "IL18RAP"               "IL18RAP"
```

Among the HERV/MALR elements which are differentially expressed, some of it are within a 10kb window from differentially expressed genes. Specifically the region containing IL18R1, IL18RAP and IL1R2also contains many HERVs/MALRs.

top 20 of all differentially expressed probesets

The columns ratioinfD1_ratiosupD1, ratioinfD3_ratiosupD3, ratioinfD6_ratiosupD6 correspond to the log fold changes for the corresponding questions, adj.P.Val for the adjusted p value (by Benjamini-Hochberg) of the whole model, rep for the corresponding repertoire and alias for the alias of genes.

```
#top 20 of all probesets
#head(top_signif[,c(1,2,8,14,15)], n=20)
knitr::include_graphics("../youden_top_signif_nocode.png")
```

	X	LowD1_HighD1	LowD3_HighD3	LowD6_HighD6	AveExpr	adj.P.Val	rep	alias
PSR02009932_at		-0.3685800	1.635551	0.71779888	5.754239	0.0002475751	HTA	IL18R1
PSR02009928_st		-0.3072105	1.365185	0.48892554	7.176583	0.0002475751	HTA	IL18R1
PSR02009933_at		-0.2620939	1.550406	0.46680459	6.533891	0.0002475751	HTA	IL18R1
PSR02009933_st		-0.2873342	1.580797	0.49528535	6.225051	0.0002475751	HTA	IL18R1
IL18R1-opti_st		-0.2800335	1.401550	0.37614464	5.923488	0.0002475751	Opti	IL18R1
PSR02009932_st		-0.2512269	1.509631	0.52475507	6.021597	0.0002475751	HTA	IL18R1
PSR02009929_at		-0.3494956	1.397809	0.56320002	5.768926	0.0002475751	HTA	IL18R1
PSR01015801_at		1.0721514	-0.178140	-0.09800167	5.847369	0.0002475751	HTA	MOV10
PSR02009912_st		-0.2385982	1.071377	0.42102374	9.090205	0.0002475751	HTA	IL18R1
PSR02009927_st		-0.2668339	1.238713	0.39714904	8.701146	0.0002848419	HTA	IL18R1
100223002-HERV0378uL_st		-0.3996421	1.333662	0.18265451	4.946994	0.0002848419	HERV_Dfam	LTR103b_Mam
PSR02009929_st		-0.2792397	1.184171	0.35869126	8.043761	0.0002848419	HTA	IL18R1
PSR02009912_at		-0.2295794	1.022783	0.43778837	9.097963	0.0002848419	HTA	IL18R1
PSR02009917_at		-0.2123254	1.234174	0.47050287	7.223407	0.0002848419	HTA	IL18R1
206618_st		-0.3129585	1.394988	0.27678226	6.980964	0.0002848419	U133	IL18R1
100222901HFRDsLU3_at		-0.2286827	1.323496	0.42022236	2.817106	0.0003223114	HERV_proto	HERV-FRD_10v1
100223001-HERV0378uL_at		-0.2665462	1.392995	0.26820003	2.256451	0.0003551593	HERV_Dfam	LTR103b_Mam
021460102-HERV0599uL_st		-0.1678879	1.016879	0.29314858	8.562564	0.0003806177	HERV_Dfam	LTR82B
PSR02009805_st		-0.3159810	1.097431	0.28932772	6.713285	0.0003910529	HTA	IL1R2
PSR02009917_st		-0.2203316	1.217135	0.45871883	7.827840	0.0003910529	HTA	IL18R1

top 10 of differentially expressed probesets targetting HERVs or MALRs

```
#top 10 hervs
#head(top_signif_herv[,c(1,2,8,14,15)], n=10)
knitr::include_graphics("../youden_top_hervs_nocode.png")
```

	X	LowD1_HighD1	LowD3_HighD3	LowD6_HighD6	AveExpr	adj.P.Val	rep	alias
100223002-HERV0378uL_st		-0.3996421	1.333662	0.18265451	4.946994	0.0002848419	HERV_Dfam	LTR103b_Mam
100222901HFRDsLU3_at		-0.2286827	1.323496	0.42022236	2.817106	0.0003223114	HERV_proto	HERV-FRD_10v1
100223001-HERV0378uL_at		-0.2665462	1.392995	0.26820003	2.256451	0.0003551593	HERV_Dfam	LTR103b_Mam
021460102-HERV0599uL_st		-0.1678879	1.016879	0.29314858	8.562564	0.0003806177	HERV_Dfam	LTR82B
021460302-MALR1004uL_at		-0.1538368	1.060042	0.46150059	7.467887	0.0004947587	MALR_Dfam	MLT1D
021460605-HERV0882cL_at		-0.1440405	1.557205	0.30966895	2.960710	0.0008493322	HERV_Dfam	MER61-int
021460101-HERV0599uL_st		-0.2805083	1.012695	0.28276023	5.118808	0.0010143706	HERV_Dfam	LTR82B
021455702-MALR1018uL_at		-0.4603911	1.095959	0.43252591	7.976117	0.0011277884	MALR_Dfam	MLT1J
021455602-MALR1012uL_at		-0.3652058	1.218490	0.36388324	8.048459	0.0012509121	MALR_Dfam	MLT1G1
060531302-HERV0492cL_at		-0.6983111	1.444861	0.08345865	4.202371	0.0012876890	HERV_Dfam	LTR33B

By analysing the top10 of probesets targetting HERVs (or MALR), the first 3 HERVs (1 dfam (2 probesets: 100223002-HERV0378uL_st and 100223001-HERV0378uL_at) and 1 proto: 100222901HFRDsLU3_at) are both very close or within the gene OLAH, involved in fatty acid biosynthesis. No obvious link could be found between this gene and "sepsis" or "immunodeficiency" in the litterature.

IL18R1 was the most present differentially expressed gene as 14 probesets targeted IL18R1 on top 20 of all probesets. Interestingly, the 4th (021460102-HERV0599uL_st), 5th(021460302-MALR1004uL_at), 6th (021460605-HERV0882cL_at) and 7th (021460101-HERV0599uL_st) herv/malr are located within this gene (or in the region). IL18R1 is a cytokine receptor that belongs to the interleukin 1 receptor family. This receptor specifically binds interleukin 18 (IL18), and is essential for IL18 mediated signal transduction. IFN-alpha and IL12 are reported to induce the expression of this receptor in NK and T cells. Alternatively spliced transcript variants encoding different isoforms have been found for this gene [provided by RefSeq, Sep 2013].

Here is a picture of IL18R1 in its genomic context taken from Ensembl browser:

```
knitr::include_graphics("../IL18R1.png")
```

On figure 20 the dotted red lines show the position of the 4 HERVs or MALRs found differentially expressed in the region. Interestingly, the one on the left is in a region marked as promoter flank and corresponds to a conserved region accross mammals species. It suggested that this HERV could have a role on the expression of IL18R1 (as alternative promoter or enhancer)

We also represented examples of probesets targeting IL18R1 and HERVs in the region.

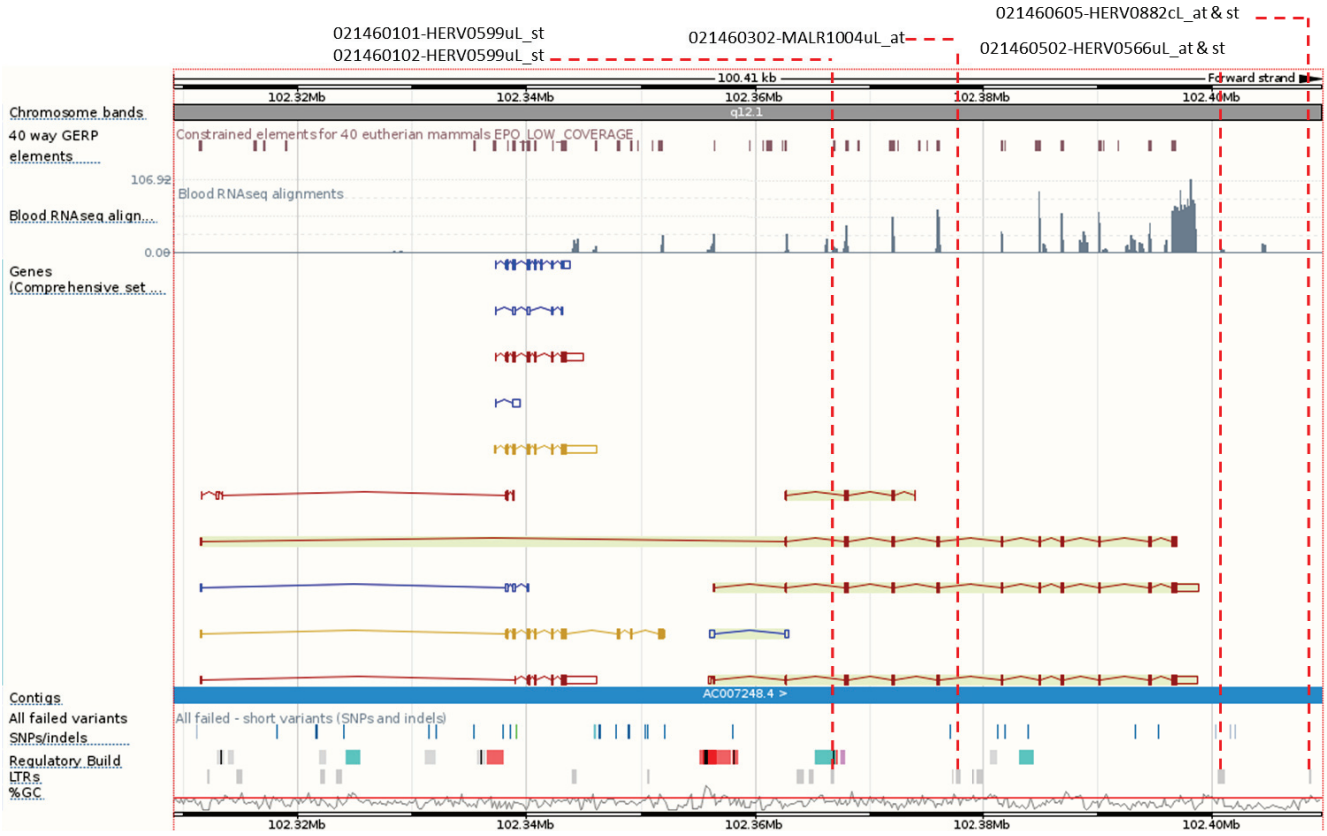


Figure 20: **IL18R1 genomic context**. Figure taken from Ensembl website. The dotted red lines show the position of the HERVs and MALRs found differentially expressed in the region.

```
covdesc$youdenday=as.factor(paste(covdesc$youden_cd74, covdesc$Day2, sep='_'))

# probeset example for IL18R1 and probesets of hervs in the region
probeset_list=c("PSR02009932_at", "021460102-HERV0599uL_st", "021460101-HERV0599uL_st",
                "021460302-MALR1004uL_at", "021460605-HERV0882cL_at")
plots=list()
for(i in c(1:length(probeset_list)))
{
    if (i==1){alias="IL18R1"}
    else {alias=''}
    name=probeset_list[i]
   plots[[i]]=boxMyProbe2(ps_name=name, fltr_RMA=fltr_RMA ,
                          covdesc=covdesc, question=alias,
                          detail=alias, xgroup="youdenday",
                          sample_id="Identifiant_patients", plot=FALSE)
}
```

```
do.call("grid.arrange", c(plots, ncol=3))
```

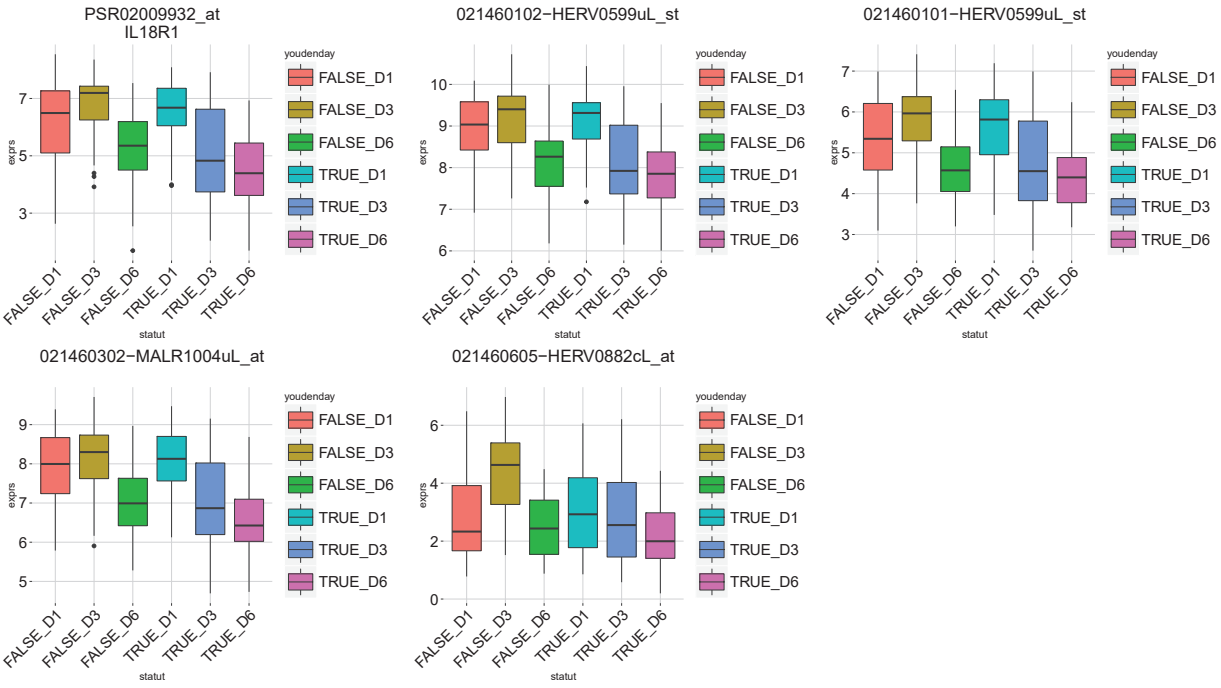


Figure 21: **IL18R1 expression and associated HERV-MALRs**. Each column represents a group of samples based on the day of sample and the D3/D1 ratio of CD74 of patients. FALSE is equal to a Low ratio, and TRUE a high D3/D1 ratio of CD74.

We confirmed from the boxplots what we found in differentially expressed analysis, that the group FALSE_D3 (Low ratio D3/D1 of CD74 at D3) is higher expressed than TRUE_D3 group (High ratio D3/D1 of CD74 at D3). We also saw that the HERV/MALRs elements which are co-localized with IL18R1 had the same expression profile as IL18R1, except for 021460605-HERV0882cL_at probeset which was higher expressed for the immunosuppressed group at D3 compared to the other groups.

There were also HERV/MALRs (021455702-MALR1018uL_at, 021455602-MALR1012uL_at) within IL1R2, which were also differentially expressed, and in the same region as IL18R1.

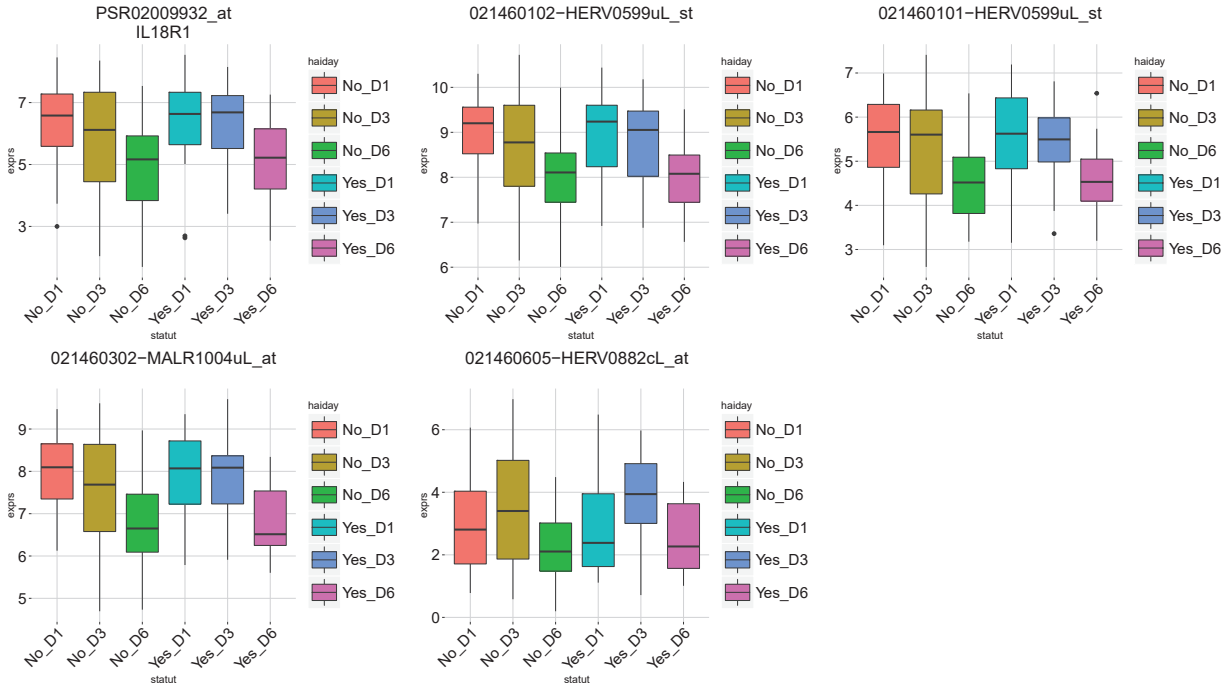


Figure 22: **IL18R1 expression and associated HERV-MALRs by HAI-day groups**. Each column represents a group of samples based on the day of sample and the HAI status of patients. No is equal to No HAI, and Yes to HAI.

```
# probeset example for IL18R1 and probesets of hervs in the region
probeset_list=c("PSR02009932_at", "021460102-HERV0599uL_st", "021460101-HERV0599uL_st",
                "021460302-MALR1004uL_at", "021460605-HERV0882cL_at")
plots2=list()
for(i in c(1:length(probeset_list)))
{
    if (i==1){alias="IL18R1"}
    else {alias=''}
    name=probeset_list[i]
   plots2[[i]]=boxMyProbe2(ps_name=name, fltr_RMA=fltr_RMA ,
                            covdesc=covdesc, question=alias,
                           detail=alias, xgroup="haiday",
                           sample_id="Identifiant_patients", plot=FALSE)

}
```

Here are shown the same probesets but this time as a function of HAI and day groups of patients. Yes correspond to HAI and NO to No HAI (figure 22).

```
do.call("grid.arrange", c(plots2, ncol=3))
```

We see that IL18R1 and associated HERVs have lower expression at D6 compared to other days, no matter the HAI status. The probeset 021460605-HERV0882cL_at has different profile with low expression at D1 and D6 compared to D3, no matter the group.

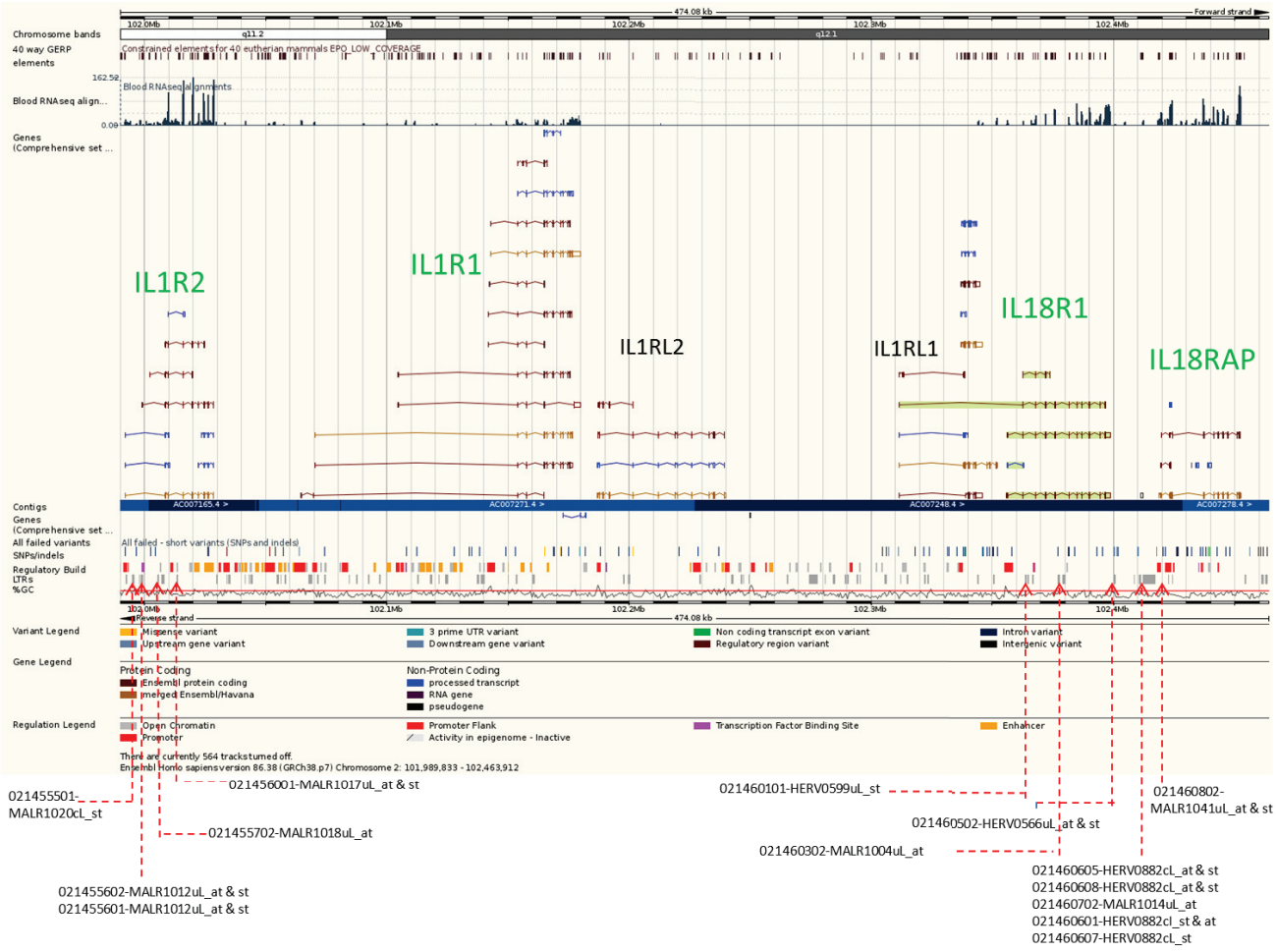


Figure 23: **IL1 receptors family genomic context**. Figure taken from Ensembl website. In green are the gene names which are differentially expressed. The dotted red lines represent the differentially expressed HERVs/MALRs in the region.

```
knitr::include_graphics("../data/Human_IL18R1_large_region_modified.png")
```

This figure 23 shows the genomic region of IL1 receptors family. The IL1 superfamily is a group of 11 cytokines, which the majority are proinflammatory cytokines, including IL18, IL1α et IL1β.

In green are the genes found differentially expressed between ratio CD74 groups of patients at D3. We observed that IL18R1 was not the only differentially expressed gene in this region, but there were also IL1R2, IL1R1 and IL18RA, mebers of IL1 receptors family. Are also represented differentially expressed HERVs/MALRs. Interestingly we saw that those elements are localized on the borders of the region, there was no differentially expressed HERV/MALR in the middle. We don't really know what it means but it suggests at least that the HERV expression was not random. In summary, all this region was highly expressed in septic shock patients at D3. IL1 receptors family and associated HERVs could be markers of this state. And it would be interesting to understand the mechanisms explaining this co expression HERVs / genes in all the region, for exemple by removing the co-expressed HERVs and see if the gene is still over expressed in this context.

We also have on top 10 differentially expressed HERV, on chromosome 6, a HERV (060531302-HERV0492cL_at), near FKBP5 (involved in immunity, but not present on the chip), potentially

interesting too. The protein encoded by FKBP5 gene is a member of the immunophilin protein family, which play a role in immunoregulation and basic cellular processes involving protein folding and trafficking. This encoded protein is a cis-trans prolyl isomerase that binds to the immunosuppressants FK506 and rapamycin. It is thought to mediate calcineurin inhibition. It also interacts functionally with mature hetero-oligomeric progesterone receptor complexes along with the 90 kDa heat shock protein and P23 protein. This gene has been found to have multiple polyadenylation sites. Alternative splicing results in multiple transcript variants.[provided by RefSeq, Mar 2009]

Because they are co-expressed and co-localized HERVs/MALRs with known genes involved in immunity, these HERVs are potential biomarkers of immunosuppression state and would deserve to be validated in PCR.

Mortality

For the mortality at D28 endpoint, we wanted to study the differential expression between:

- Survivor vs No survivor
- Survivor D1 vs No survivor D1
- Survivor D3 vs No survivor D3
- Survivor D6 vs No survivor D6

We did not find any differentially expressed probesets. (Scripts and figures not shown).

Longitudinal analysis

For longitudinal analysis, we wanted to study the differential expression between:

- D3 vs D1
- D6 vs D1
- D6 vs D3

```
f=factor(covdesc$Day2)
design=model.matrix(~0+f)
colnames(design)=levels(f)

load(paste0('../data/cdf',icdf,'/corfit_day.RData'))
fit <- lmFit(fltr_RMA,design,block=covdesc$patients,
             correlation=corfit$consensus)

cont.wt=makeContrasts(
  D3_D1 = D3 - D1,
  D6_D1 = D6 - D1,
  D6_D3 = D6 - D3,
  levels=design
)
fit2=contrasts.fit(fit,cont.wt)
fit2=eBayes(fit2)
```

```
results = decideTests(fit2, p.value=0.05,lfc=1, adjust.method="fdr")
vennDiagram(results, include=c("up","down"),
            circle.col=c("yellow","green", "blue", "red"),
            counts.col=c("darkgreen","darkred"),
            cex=c(1,0.7,0.5))
```

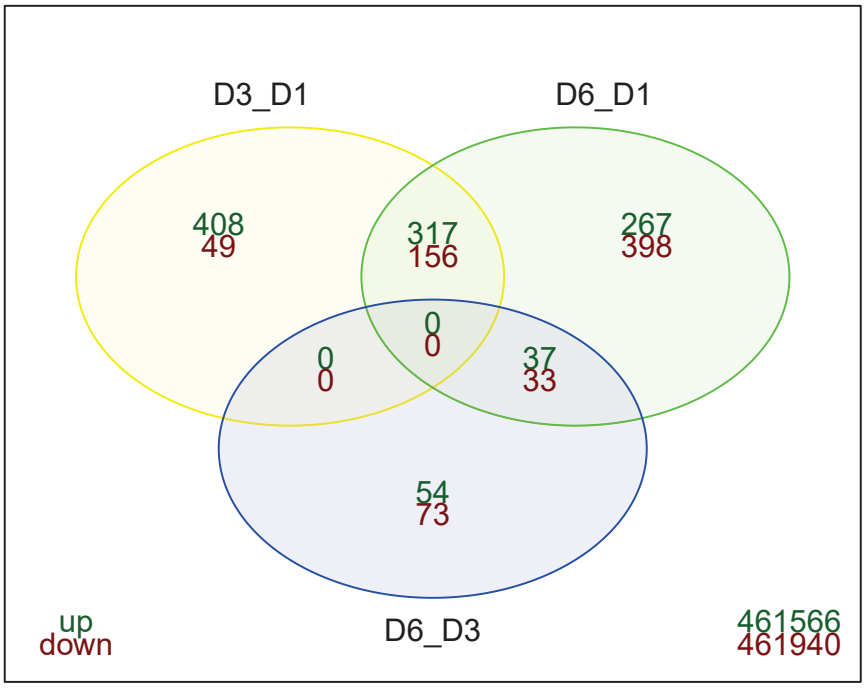


Figure 24: **Venn diagram for longitudinal model**. The numbers in dark green represent the probesets which are up-regulated for the corresponding question. In dark red the down-regulated probesets.

```
top_all=topTable(fit2,adjust="fdr",number=nrow(fltr_RMA))

top1=topTable(fit2,adjust="fdr",number=nrow(fltr_RMA), coef=1)
#coef is for choosing p_value of corresponding contrast
#the consequence is that we don't have logfc for all questions, so i create all data
top2=topTable(fit2,adjust="fdr",number=nrow(fltr_RMA), coef=2)
top3=topTable(fit2,adjust="fdr",number=nrow(fltr_RMA), coef=3)

# regrouping of each adjusted pvalue for each contrast in the same data
i=match(rownames(top_all), rownames(top1))
top_all$adj.P.coef1=top1$adj.P.Val[i]
i=match(rownames(top_all), rownames(top2))
top_all$adj.P.coef2=top2$adj.P.Val[i]
i=match(rownames(top_all), rownames(top3))
top_all$adj.P.coef3=top3$adj.P.Val[i]

i=getAnnotMatches(rownames(top_all))
aliases=ann$ALIAS[i]
top_all$rep=ann$repertoire[i]
top_all$alias=aliases
chr=ann$CHR[i]
strt=ann$CHRLOC[i]
end=ann$CHRLOCEND[i]
top_all$coords_19=paste0("chr",chr, ':', strt, '-', end)
```

Description of all differentially expressed probesets

As we can see on the Venn diagram (Figure 24), there are many differentially expressed probesets across D1, D3 and D6 after septic shock.

Total number of differentially expressed probesets: 1749

```
pval_thresh=0.05
lfc_thresh=1

top_signif=top_all[which((abs(top_all$D3_D1) > lfc_thresh &
                            top_all$adj.P.coef1 < pval_thresh) |
                           (abs(top_all$D6_D1) > lfc_thresh &
                              top_all$adj.P.coef2 < pval_thresh) |
                           (abs(top_all$D6_D3) > lfc_thresh &
                              top_all$adj.P.coef3 < pval_thresh)
                         ),]
top_signif_herv=top_signif[which(top_signif$rep!="HTA" &
                                   top_signif$rep!="U133" & top_signif$rep!="Opti"),]
#nrow(top_signif)
```

Number of probesets targetting genes: 609

```
#Number of probesets targetting genes
# nrow(top_signif[which(top_signif$rep=="U133" | top_signif$rep=="Opti" |
#                         top_signif$rep=="HTA" ),])
```

Number of Genes (not probesets): 52

```
#Number of Genes (not probesets)
# length(unique(top_signif$alias[which(top_signif$rep=="U133" | top_signif$rep=="Opti" |
#                                        top_signif$rep=="HTA" )]))
```

Number of hervs/malr probesets and elements: 1140 and 690 respectively

```
#Number and list of hervs/malr probesets and elements
# nrow(top_signif_herv)
# length(unique(substr(rownames(top_signif_herv), 1,7)))
```

Differentially expressed probesets between D1 and late time points

For this analysis, we thoughtk it was more interesting to focus on the ones moving between D1 and D3/D6. As we don't know the exact first day of the shocks (D1 is the day of arrival at the hospital, not necessarily the first day of shock), we want to keep probesets moving early and remaining at the same level of expression in late time points.

Total number of differentially expressed probesets between D1 and late time points: 473

```
top_signif=top_all[which((abs(top_all$D3_D1) > lfc_thresh &
                            top_all$adj.P.coef1 < pval_thresh) &
                           (abs(top_all$D6_D1) > lfc_thresh &
                              top_all$adj.P.coef2 < pval_thresh)
),]
top_signif_herv=top_signif[which(top_signif$rep!="HTA" &
                                   top_signif$rep!="U133" & top_signif$rep!="Opti"),]
#nrow(top_signif)
```

Number of probesets targetting genes: 191

```
#Number of probesets targetting genes
# nrow(top_signif[which(top_signif$rep=="U133" | top_signif$rep=="Opti" |
#                         top_signif$rep=="HTA" ),])
```

Number and list of Genes (not probesets): 16

```
#Number and list of Genes (not probesets)
# length(unique(top_signif$alias[which(top_signif$rep=="U133" | top_signif$rep=="Opti" |
#                                        top_signif$rep=="HTA" )]))
unique(top_signif$alias[which(top_signif$rep=="U133" | top_signif$rep=="Opti" |
                                top_signif$rep=="HTA" )])
```

```
##  [1] "GOS2"    "MME"     "OSM"     "BCL2A1"  "IL10"    "S1PR1"   "SOCS3"
##  [8] "MERTK"   "HGF"     "MBP"     "GADD45A" "HPSE"    "MYH11"   "CLEC7A"
## [15] "SLPI"    "FLT1"
```

Number of hervs/malr probesets and elements: 1282, 166 respectively

```
#Number and list of hervs/malr probesets and elements
# nrow(top_signif_herv)
# length(unique(substr(rownames(top_signif_herv), 1,7)))
```

16 genes were differentially expressed between D1 and the two late time points. Are there co-localized HERVs/MALR arround these genes (<10kb) ?

```
ts=top_signif[which(top_signif$rep=="U133" |
                            top_signif$rep=="Opti" |
                            top_signif$rep=="HTA" ),]
genes_coord=ts[!duplicated(ts$alias),]

range_genes=sapply(genes_coord$coords_19, function(c)
{
  chr=strsplit(c, ":")[[1]][1]
  st=strsplit(c, ":")[[1]][2]
  start=as.numeric(strsplit(st, '-')[[1]][1])
  stop=as.numeric(strsplit(st, '-')[[1]][2])
  new_start=start#-10000 #10 kb is enough ?
  new_stop= stop#+10000
  #print(chr)
```

```
  #range=paste0(chr,':',new_start,'-', new_stop)
  return(c(chr, new_start, new_stop))
})
#Problem on coordinates, remove it
colnames(range_genes)=genes_coord$alias
range_genes=range_genes[,c(-3,-9)]
#print(range_genes)

#for each herv/malr probeset, are they in a range of a gene
t=sapply(seq(1, length(top_signif_herv$coords_19)), function(j)
{
  c=top_signif_herv$coords_19[j]
  if(!is.na(c))
  {
    chr=strsplit(c, ":")[[1]][1]
    st=strsplit(c, ":")[[1]][2]
    start=as.numeric(strsplit(st, '-')[[1]][1])
    stop=as.numeric(strsplit(st, '-')[[1]][2])
    rg=lapply(seq(1,ncol(range_genes)),function(i)
    {
      x=range_genes[,i]
      #print(x)
      if(!(any(is.na(x))))
      {
        if(chr==x[1])
        {
          #print(start - as.numeric(x[2]))
           if(abs(start-as.numeric(x[2]))< 10000 || abs(stop-as.numeric(x[3]))< 10000)
          {
            return(paste(rownames(top_signif_herv)[j],colnames(range_genes)[i], sep='__'))
          }
          else if (start>x[2] && stop <x[3])
          {
            return(paste(rownames(top_signif_herv)[j],colnames(range_genes)[i], sep='__'))
          }
        }
      }
    return(NA)
    })
    return(rg)
  }
  else return(NA)
})
print(unlist(t[!is.na(t)]))
```

```
## [1] "032409902-MALR1003uL_at__MME"   "032409902-MALR1003uL_st__MME"
## [3] "032409701-MALR1018uL_at__MME"   "160238102-MALR1026cL_st__MYH11"
## [5] "041525401-MALR1002uL_at__HPSE"  "021612603ERVFsLU3p_at__MERTK"
## [7] "012468903HK01uLRp_st__IL10"
```

Among co-localized and differentially expressed HERV/MALR and genes, we identified a HERV proto (012468903HK01uLRp_st) close (6kb) to IL10 gene. The HERV was part of HML-1 family.

IL10 probesets: D3_D1, D6_D1 and D6_D3 columns show the log fold changes for the corresponding questions, adj.P.Val the adjusted pvalue for the whole model and alias the alias of targetted gene.

```
top_signif[grep('IL10', top_signif$alias),c(1,2,3,7,12)]
```

```
##                     D3_D1     D6_D1      D6_D3    adj.P.Val alias
## PSR01057742_at -1.149863 -1.518875 -0.3690115 2.394300e-28  IL10
## PSR01057742_st -1.060116 -1.395435 -0.3353195 8.323247e-26  IL10
## PSR01057739_st -1.000803 -1.283859 -0.2830564 5.896930e-18  IL10
```

IL10 associated HERVs probesets:

```
top_all[grep('0124689', rownames(top_all)),c(1,2,3,7)]
```

```
##                             D3_D1      D6_D1       D6_D3    adj.P.Val
## 012468903HK01uLRp_st  -1.0470464 -1.2619996 -0.21495317 6.445713e-12
## 012468903HK01uLRp_at  -0.7204375 -0.8257294 -0.10529188 7.059310e-11
## 012468904HK01uLU5p_at -0.7916789 -0.8084284 -0.01674953 1.054071e-07
## 012468904HK01uLU5p_st -0.3597350 -0.3272957  0.03243935 2.536879e-02
```

We observed that the gene had lower expression on late time points compared to D1. The HERVs had the same profile of expression. It would also be interesting to look closer on IL10 gene and its associated HERVs as they can represent a marker of evolution of immunosuppression after a septic shock. It would be interesting to dig into the elements associated to MME and MERTK genes.

Conclusion

After a descriptive section of the HERVome in septic shock patients, we tried to find differentially expressed probesets with endpoints of interest. We did not find any differentially expressed probeset for HAI and mortality endpoints. But by taking into account the ratio of CD74 expression from RTqPCR, we were able to detect 115 differentially expressed probesets. Among these 115, there were many probesets targetting the same elements (genes or hervs), and even in the same region (between 102 and 103 Mb on chromosome 2, in the IL1 receptors family). In this region there are IL18R1, IL1R1, IL1R2 and IL18RAP which were differentially expressed. In the same region, there were many HERVs or MALRs elements having the same expression profile. They could make potential good markers of the immunosuppression state. We can hypothesize that the HERV or MALR elements have a role on the neighbor genes by intiating or ending their transcription, as it has already been seen in litterature in another contexts for another genes and HERVs (Lamprecht et al. 2010; Suntsova et al. 2015).

Acknowledgment

The authors would like to thank A. Lepape, F. Venet, E. Peronnet, B. Meunier, E. Cerrato, V. Cheynet, G. Oriol for their contribution to this work.

References

Gimenez, Juliette, Cécile Montgiraud, Jean-Philippe Pichon, Bertrand Bonnaud, Maud Arsac, Karine Ruel, Olivier Bouton, and François Mallet. 2010. “Custom Human Endogenous Retroviruses Dedicated Microarray

Identifies Self-Induced HERV-W Family Elements Reactivated in Testicular Cancer Upon Methylation Control." *Nucleic Acids Research* 38 (7): 2229–46. doi:10.1093/nar/gkp1214.

Johnson, W. Evan, Cheng Li, and Ariel Rabinovic. 2007. "Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods." *Biostatistics (Oxford, England)* 8 (1): 118–27. doi:10.1093/biostatistics/kxj037.

Lamprecht, Björn, Korden Walter, Stephan Kreher, Raman Kumar, Michael Hummel, Dido Lenze, Karl Köchert, et al. 2010. "Derepression of an Endogenous Long Terminal Repeat Activates the CSF1R Proto-Oncogene in Human Lymphoma." *Nature Medicine* 16 (5): 571–79, 1pfollowing 579. doi:10.1038/nm.2129.

Laska, Magdalena Janina, Tomasz Brudek, Kari Konstantin Nissen, Tove Christensen, Anné Møller-Larsen, Thor Petersen, and Bjørn Andersen Nexø. 2012. "Expression of HERV-Fc1, a Human Endogenous Retrovirus, Is Increased in Patients with Active Multiple Sclerosis." *Journal of Virology* 86 (7): 3713–22. doi:10.1128/JVI.06723-11.

Pérot, Philippe, Nathalie Mugnier, Cécile Montgiraud, Juliette Gimenez, Magali Jaillard, Bertrand Bonnaud, and François Mallet. 2012. "Microarray-Based Sketches of the HERV Transcriptome Landscape." *PLOS ONE* 7 (6): e40194. doi:10.1371/journal.pone.0040194.

Pérot, Philippe, Christina Susanne Mullins, Magali Naville, Cédric Bressan, Maja Hühns, Michael Gock, Florian Kühn, et al. 2015. "Expression of Young HERV-H Loci in the Course of Colorectal Carcinoma and Correlation with Molecular Subtypes." *Oncotarget* 6 (37): 40095–40111. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4741882/.

Rhyu, Dong-Won, Yun-Jeong Kang, Mee-Sun Ock, Jung-Woo Eo, Yung-Hyun Choi, Wun-Jae Kim, Sun-Hee Leem, Joo-Mi Yi, Heui-Soo Kim, and Hee-Jae Cha. 2014. "Expression of Human Endogenous Retrovirus Env Genes in the Blood of Breast Cancer Patients." *International Journal of Molecular Sciences* 15 (6): 9173–83. doi:10.3390/ijms15069173.

Suntsova, Maria, Andrew Garazha, Alena Ivanova, Dmitry Kaminsky, Alex Zhavoronkov, and Anton Buzdin. 2015. "Molecular Functions of Human Endogenous Retroviruses in Health and Disease." *Cellular and Molecular Life Sciences: CMLS* 72 (19): 3653–75. doi:10.1007/s00018-015-1947-6.

Rétrovirus endogènes humains et réponse immunitaire de l'hôte suite à une agression inflammatoire.

Suite à une agression inflammatoire, telle que le choc septique, des brulures graves ou un traumatisme sévère, le système immunitaire répond par une modulation massive du transcriptome dans le sang. On propose d'explorer un autre répertoire que l'expression des gènes et de s'intéresser aux éléments répétés du génome, peu étudiés dans ces contextes, et plus particulièrement aux rétrovirus endogènes humains (HERV). Ils représentent plus de 8% du génome chez l'Homme. Certains sont exprimés dans des situations similaires à l'agression inflammatoire (cancer, maladies auto-immunes) et ont un impact sur la réponse immunitaire.

Dans ce travail, nous cherchons à décrire et comprendre la contribution des HERV, au sein de la réponse immunitaire de l'hôte à l'agression inflammatoire. Pour cela, nous avons développé des méthodes et outils spécifiquement dédiés à la description du HERVome, au niveau génomique et transcriptomique. Nous montrons que les HERV sont exprimés dans le sang, modulés chez les patients, et que certains pourraient jouer un rôle sur l'expression de gènes de la réponse immunitaire situés à proximité. Nous évaluons également le polymorphisme de présence des HERV dans le génome de plus de deux mille individus répartis dans les populations humaines. On met en évidence que le polymorphisme HERV est globalement important, qu'il est lié à la population d'appartenance et que certains loci sont absents dans la majorité des génomes étudiés.

Finalement, par différentes approches, nous identifions des associations entre gènes de la réponse immunitaire et HERV, suggérant que ces éléments peuvent jouer un rôle important dans la réponse de l'hôte à l'agression inflammatoire.

Mots clés : Rétrovirus endogènes humains – agression inflammatoire – réponse immunitaire de l'hôte – choc septique – brulés – traumatisés – transcriptome – génome

Human endogenous retroviruses and host immune response following inflammatory aggression

Following inflammatory injury, like a septic shock, severe burn or important trauma, the immune system responds by a massive modulation of its transcriptome in the blood. We propose to explore another repertoire than gene expression and to focus on repeated elements, especially on HERVs. They represent more than 8% of the human genome. HERVs are expressed in similar settings (cancer or auto-immune diseases) and impact immune response.

In this project, we describe and aim to better understand the HERV contribution in host immune response, following inflammatory aggression. To bring elements of response, we developed specifically dedicated tools to describe the HERVome, either at genomic or transcriptomic level. We show HERVs are expressed in blood in these settings, modulated in patients and could play a role on nearby gene expression. We also evaluate the polymorphism of presence of HERV loci on more than two thousands individuals, grouped into human populations. We show an important HERV polymorphism, that it is population-specific, and that some loci are absent in the majority of the analyzed genomes.

Finally, with different approaches, we identify associations between immune-response genes and HERVs, suggesting these elements can play a role in host immune response following inflammatory aggressions.

Keywords: Human endogenous retroviruses – inflammatory aggression – host immune response – septic shock – burn – traumatized – transcriptome – genome

Thèse préparée au sein de l'EA7426 Physiopathologie de l'immunosuppression associée aux réponses inflammatoires systémiques (PI3), 5ème étage Pavillon P, Hôpital Edouard Herriot