Long non-coding RNAs in cancer : the role of HOTAIR in Epithelial-to-Mesenchymal Transition

Claire Bertrand

Université Pierre et Marie Curie

Ecole doctorale « Complexité du Vivant »

Long non-coding RNAs in Cancer

The role of HOTAIR in Epithelial-to-Mesenchymal Transition

par Claire BERTRAND

Thèse de Doctorat de Biologie Moléculaire

Dirigée par Antonin MORILLON

Présentée et Soutenue publiquement le 27 octobre 2014

Devant un jury composé de :

Pr. Frédéric DEVAUX	Président
Pr. Torben Heick JENSEN	Rapporteur
Dr. Sven DIEDERICHS	Rapporteur
Dr. Jean-Christophe ANDRAU	Examinateur
Dr. Marina PINSKAYA	Examinateur
Dr. Antonin MORILLON	Directeur de thèse

ABSTRACT

In the recent years, high-throughput sequencing studies have shown that the human genome is pervasively transcribed into thousands of processed non-coding transcripts, including a large class of long non-coding RNAs (lncRNAs). An increasing number of studies is underlining the importance and the diversity of lncRNA roles in genome regulation, as well as their impact on development and diseases, particularly in cancer. The processes leading to cancer initiation and progression in humans are extensively studied, in particular the Epithelial-to-Mesenchymal Transition (EMT) which enables epithelial cancer cells to migrate and invade other tissues to form metastases. If several lncRNAs have been associated with this process, their molecular function in EMT is not clearly defined.

Using a specific and well-established *in vitro* cell model of EMT, derived from primary Human Epithelial Kidney cells, we aimed to isolate EMT-associated lncRNAs, and to investigate their functional role in this process. Using high-throughput RNA sequencing approaches, we defined a catalogue of previously annotated, but also novel lncRNAs significantly deregulated between epithelial and mesenchymal states of HEK cells. Among them, we identified HOTAIR, already linked to cancer metastasis, and well described as a scaffold RNA guiding chromatin-modifying complexes PRC2 and LSD1/CoREST/REST to repress gene transcription at the epigenetic level. Using loss- and gain-of-function approaches, we showed that HOTAIR is not an inducer of the EMT process *per se*, but is a major regulator of cell proliferation rate as well as migratory and invasive capacities. Furthermore, we generated stable cell-lines over expressing HOTAIR transcripts missing PRC2- or LSD1-interacting domains, and performed global transcriptome analysis as well as phenotypic studies. Our results showed that LSD1-interacting domain is crucial for HOTAIR-mediated gene regulation.

Altogether, our results give new insights into lncRNAs potential role in EMT, a crucial process in cancer development and metastasis, especially with a better understanding of HOTAIR-mediated gene regulation mechanism and its role in the acquisition of a metastatic phenotype by cancer cells. Further studies will be performed to deeper investigate lncRNAs role in EMT, particularly for previously unannotated lncRNAs.

RÉSUMÉ

De nombreuses études ont montré récemment que le génome humain est très largement transcrit en milliers d'ARN non traduits en protéines. Parmi ces ARN, les longs ARN non-codants (ARNlnc) ont été très étudiés, pour leur rôle majeur dans la régulation du génome, ainsi que leur impact majeur au cours du développement et de la progression de nombreuses maladies telles que le cancer. Comprendre les mécanismes permettant l'initiation et la progression des cancers est aujourd'hui un enjeu majeur. La transition épithélio-mésenchymateuse (TEM), donnant à une cellule la capacité de migrer et donc de former des métastases, semble être l'un des processus cruciaux transformant une tumeur bénigne en maladie mortelle. Certains ARNlnc ont été associés à ce phénomène, mais leur fonction reste à définir.

Un modèle *in vitro* de cellules immortalisées dérivées de cellules épithéliales rénales (HEK) primaires, et des approches de séquençage d'ARN à très haut débit, nous ont permis de définir un catalogue d'ARN, déjà annotés ou nouveaux, dérégulés entre cellules épithéliales et mésenchymateuses. Parmi eux, nous avons identifié HOTAIR, largement étudié pour son expression augmentée dans les tumeurs métastasées, et pour son interaction avec les complexes protéiques PRC2 et LSD1/CoREST/REST, induisant la répression de la transcription de nombreux gènes via des modifications épigénétiques de la chromatine. Par des approches de perte et de gain de fonction, nous avons montré que HOTAIR n'est pas impliqué dans l'initiation de la TEM, mais est un régulateur majeur de la prolifération cellulaire, ainsi que des capacités de migration et d'invasion des cellules. Nous avons par la suite générés des lignées cellulaires sur-exprimant HOTAIR privé de son domaine d'interaction avec PRC2 ou LSD1. L'étude du phénotype de ces cellules, ainsi que l'établissement de leur transcriptome par séquençage haut débit, a permis de montrer que le domaine d'interaction avec le complexe LSD1/CoREST/REST est crucial pour la régulation de nombreux gènes par HOTAIR.

Ces résultats permettent une meilleure compréhension du rôle des ARNlnc dans la TEM, et notamment de la fonction cruciale de HOTAIR dans l'acquisition d'un phénotype métastatique par des cellules épithéliales cancéreuses.

TABLE OF CONTENTS

ABBREVIATIONS

As-lncRNA: antisense long non-coding RNA

ASO: Antisense Oligonucleotide

EMT: Epithelial-to-Mesenchymal Transition

ENCODE: Encyclopedia of DNA Elements

HEKs: Human Epihelial Kidney cells

HEK-Epi: Epithelial HEK cells

HEK-Mes: Mesenchymal HEK cells

ChIP: Chromatin Immunoprecipitation

ChIP-seq: ChIP sequencing

lncRNA: long non-coding RNA

MET: Mesenchymal-to-Epithelial Transition

miRNA: micro RNA

mRNA: messenger RNA

ncRNA: non-coding RNA

nt: nucleotide

PCR: Polymerase Chain Reaction

PDs: Population Doublings

qPCR: quantitative PCR

RNA-seq: RNA sequencing

RT: Reverse Transcription

siRNA: small interfering RNA

TGF-β: Transforming Growth Factor beta

INTRODUCTION

The principal RNA participants in gene expression, messenger RNAs (mRNAs), and their central role as intermediates between DNA sequence and protein synthesis have been discovered in the 1950's, establishing the central dogma of molecular biology. As a consequence, intergenic regions have been considered as non-functional, while mRNAs and encoded proteins have been extensively studied as the only actors of all cellular processes and diseases. Over the past years, however, this dogma has been largely questioned. In the early 1980's, the first transcripts without coding capability (non coding (nc)RNAs), the small nuclear (sn)RNAs, emerged, as well as other abundant classes such as small nucleolar (sno)RNAs. The first long non-coding (lnc)RNA gene reported was the imprinted H19 gene, in 1990, quickly followed by the discovery of the silencing X-inactive-specific transcript (Xist) lncRNA gene, in 1991. Micro (mi)RNAs and their many relatives were discovered in the early 2000's, revealing the importance of post transcriptional events in gene expression, particularly in eukaryotic organisms. This "ncRNA revolution" attracted increasing attention with the recent discovery of numerous other lncRNAs, and the rapid evolution of RNA and DNA high-throughput sequencing technologies.

In 2012, the Encyclopedia of DNA Elements (ENCODE) Consortium established by RNA-sequencing approaches a genome-wide catalogue of human transcripts, comprising annotated and thousands of previously unannotated ncRNAs. They reported evidence that either processed or primary transcripts cover a total of 62.1% and 74.7% of the human genome, respectively, while no more than 2% of the whole genome is protein coding (Djebali et al. 2012). Moreover, the proportion of genome that is transcribed into ncRNAs correlates with the degree of organism complexity among species, whereas it is not the case for the number of protein-coding genes even taking alternative splicing and post-translational regulation into account (Prasanth & Spector 2007). Expression of ncRNAs is very tissue and developmental stage specific, and is highly misregulated during pathological processes (Derrien et al. 2012; Esteller 2011). Several examples of ncRNAs have been shown to play key roles in the regulation of epigenetic and transcriptomic landscapes of the cell, hence emphasizing their putative importance as markers and driving forces of development and diseases.

In addition to well-known housekeeping ncRNAs (ribosomal (r)RNAs, transfer (t)RNAs, snRNAs, snoRNAs) involved in mRNA processing and translation, there is a distinct class of ncRNAs with proved or putative regulatory role. Traditionally, they are divided in function of their size, into small ncRNAs that are shorter than 200 nucleotides (nt), and lncRNAs of at

least 200 nt in length, on the basis of RNA purification protocols (Kapranov et al. 2007). Small ncRNAs have been extensively studied during the past decade, and have been shown to play a role in gene regulation at transcriptional and post-transcriptional levels, in majority, through specific base pairing with their RNA/DNA target (Huang et al. 2013). LncRNAs are less characterized, with exact structure and function known just for several of them (Fatica & Bozzoni 2014). However, an increasing number of studies revealed that they can be involved in the regulation of every stage of gene expression, underlining the importance of deeper insights into lncRNAs and their functional relevance.

In this introduction, I will first review the current knowledge on human ncRNAs, with a short description of small ncRNAs, and then focusing on lncRNAs: genomic location, biogenesis, expression and molecular mechanisms of action. In a second part, I will focus on a role of ncRNAs in diseases and cancer, and particularly in a specific biological process associated with carcinogenesis, the Epithelial-to-Mesenchymal Transition (EMT).

A. Human non-coding RNAs

1. Small non-coding RNAs

Since the first miRNAs and RNA interference (RNAi) were discovered in 1993 and 1998, respectively, numerous studies highlighted the importance of these small ncRNAs in regulation of diverse processes, such as development, apoptosis, stem cell self renewal, differentiation and maintenance of cell integrity by gene silencing pathways directing translational repression of mRNA degradation. Three main classes of small ncRNAs have been well studied: miRNAs, small interfering (si)RNAs and Piwi-interacting (pi)RNAs (Kugel & Goodrich 2012; Stefani & Slack 2008; Mattick 2009; Huang et al. 2011).

a. miRNAs

The miRNAs are endogenous non-coding RNA molecules of 21-24 nucleotides. In a majority, they are transcribed by RNA polymerase II (RNAPII) as primary transcripts (pri-miRNAs) that are 2-4 kb long single stranded RNAs, and further processed in the nucleus by Drosha in 70-100 nt long hairpin shaped precursor miRNA (pre-miRNA). Pre-miRNAs are actively exported to the cytoplasm, where Dicer further processes them into ≈ 22 nt mature

double stranded miRNAs. Subsequently, one strand of the miRNA duplex is incorporated into the multiprotein Argonaute-containing RNA-silencing complex (RISC), guides it to target complementary mRNAs and induces its degradation or inhibition of translation initiation (Bartel et al. 2004; Bartel 2005). Highly conserved in evolution, miRNAs are major regulators of many fundamental biological processes, such as cell proliferation, metabolism, embryogenesis, aging and cell death (He & Hannon 2004; Thomson & Lin 2009). Furthermore, miRNAs expression has been shown to be highly misregulated in various cancers (Brase et al. 2011; Hassan et al. 2012), Alzheimer and Parkinson's diseases (Junn & Mouradian 2010; Gehrke et al. 2010).

b. siRNAs

The siRNAs are derived from endogenous double stranded (ds)RNA precursors, resulting from processes such as pseudogenes hybridized to mRNAs, inverted repeats or bidirectional transcription, and are further processed by Dicer enzymes. SiRNAs are then loaded into specific Argonaute proteins, which guide RNAi at both RNA and DNA levels via RISC, in a process similar to miRNAs (Okamura & Lai 2008). SiRNAs have been shown to be involved in regulation of long-term gene expression, through RNA-directed DNA methylation, but also in chromatin modification events and in repression of retrotransposons (D. H. Kim et al. 2006; Ting et al. 2005; Watanabe et al. 2008; Tang 2010).

c. piRNAs

Associated with Piwi-family proteins, these small ncRNAs are 24-30 nt in length and arise from repetitive elements including transposable elements. Once loaded into Piwi proteins, they can target and induce cleavage of RNA molecules post-transcriptionally (Grivna et al. 2006; Aravin et al. 2006; Girard et al. 2006; Ishizu et al. 2011). Their function in human cells is still obscure, but their relationship with carcinogenesis has been recently reported. Indeed, two piRNAs were found aberrantly expressed between gastric cancer and non-cancerous tissues (Cheng et al. 2011; Cheng et al. 2012).

2. Long non-coding RNAs

If the exact number of human lncRNAs remains to be defined, the rapid evolution and development of high-throughput sequencing technologies allowed the establishment of detailed lncRNAs catalogues from numerous cell lines and cancer tissues.

After the development of microarrays, and later RNA-sequencing, strand-specific RNA-sequencing protocols were released in 2008 to deal with the complexity of eukaryotic "pervasive transcription" and the fact that many genes produces antisense transcripts (Wang et al. 2009; Jensen et al. 2013; van Dijk et al. 2014). In parallel, numerous sequencing protocols were developed. As an example, single cell transcriptomics provide a much more detailed view of transcription dynamics among seemingly identical cells. Fluorescent *in situ* RNA-sequencing (FISSEQ) enables single cell transcriptome study, but also determination of the precise location of each transcript within the cell (Lee et al. 2014). The fact that RNA-sequencing measures RNA steady-state levels, which do not directly reflect transcriptional activity or protein synthesis, is a major limitation of these techniques. Therefore, new methods were developed, such as global run-on-sequencing (GRO-seq) assay, which map and quantify transcriptionally engaged polymerase density genome wide (Core et al. 2008). Native elongating transcript sequencing (NET-seq) is based on the immunoprecipitation of RNA polymerase followed by deep sequencing of 3' ends of co-precipitated nascent RNAs, providing higher resolution than RNA polymerase ChIP-seq and keeping RNA strand information (Churchman & Weissman 2011).

Altogether, these technologies allowed identification and classification of growing numbers of lncRNAs, numbers that will certainly continue to increase in the next few years. The absence of described function for the majority of them, and their apparent low level of expression, led some authors to consider this "dark matter" of the genome to be only transcriptional noise, but many independent studies and techniques support the reality of pervasive transcription (van Bakel et al. 2010; Clark et al. 2011).

a. Diversity and genomic origin of human lncRNAs

The most complete human lncRNAs catalogue released to date by the ENCODE project defines 14 880 lncRNAs in the human genome (Harrow et al. 2012). If complete annotation

of the genome is still in progress, and the lncRNAs repertoire continues growing, classes of lncRNAs have already been established, based on their genomic location. Indeed, lncRNAs can be transcribed from protein-coding genes, both in intronic and/or exonic regions and in sense or antisense orientation, but also from intergenic regions including regulatory elements such as promoters, enhancers, centromeres, telomeres and repetitive sequences (Figure 1) (Derrien et al. 2012). They are in a majority generated by RNAPII machinery, vary a lot in length, have mono- or multi-exonic organisation, and are transcriptionally processed (Gibb et al. 2011).

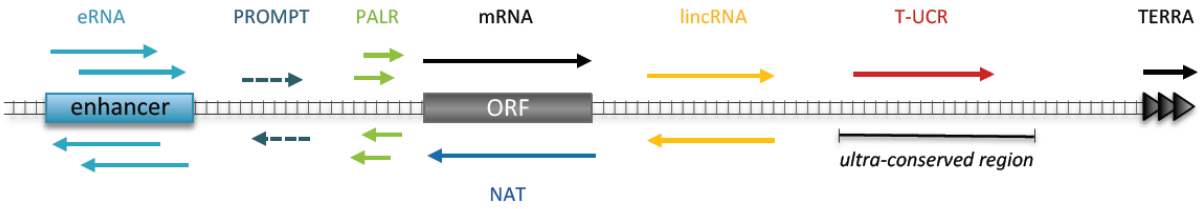

Figure 1 | Diversity and genomic origin of human lncRNAs. eRNA, enhancer-associated and enhancer-like RNA. PROMPT, promoter upstream transcript. PALR, promoter-associated lncRNA. mRNA, messenger RNA. NAT, natural antisense transcript. LincRNA, long-intervening/intergenic ncRNA. T-UCR, transcribed ultra-conserved region. TERRA, telomeric repeat-containing RNA.

i. PALRs (Promoter-Associated Long Non-Coding RNAs)

PALRs are lncRNAs transcribed from promoters of protein-coding genes, and their presence positively correlates with promoter activity. They have been shown to be involved in the modulation of associated coding-genes expression. Several examples have been described, such as CCND1 or Six3OS (Wang et al. 2008; Rapicavoli et al. 2011; Tisseur et al. 2011).

ii. PROMPTs (PROMoter Upstream Transcripts)

PROMoter Upstream Transcripts are a class of relatively short (200-600 nt), polyadenylated transcripts, produced generally in antisense direction, 0.5 to 2.5 kilobases (kb) upstream of active transcription start sites of their associated genes. Highly unstable, they have been identified following the depletion of human exosome components (Preker et al. 2008; Preker et al. 2011). Polyadenylation-like signals have been found around PROMPT 3' ends. Functional but linked to RNA degradation, and mostly represented in promoter-

upstream regions, these asymmetrically distributed signals allows more efficient RNAPII progress in the sense direction from gene promoters, providing directional RNA output from a bidirectional transcription process (Ntini et al. 2013).

iii. eRNAs (enhancer-associated ncRNAs, enhancer-like ncRNAs)

Enhancer-associated ncRNAs, or ncRNAs having enhancer-like function, have been discovered recently. Mostly bidirectional and non-polyadenylated, these transcripts are expressed at a level positively correlated with the level of messenger RNA synthesis at nearby genes. This observation suggests that eRNAs have a direct role in the recruitment of the RNA Polymerase II (RNAPII) machinery, or in the configuration of promoter-proximal DNA for transcription activation (Kim et al. 2010; Natoli & Andrau 2012; Koch et al. 2008). The abundance of eRNAs and enhancer-ncRNAs in human and mouse (3000 and 2000, respectively) suggests that their function might be conserved between organisms. It has been reported that 17b-oestradiol (E2) binding to oestrogen receptor a causes a global increase in eRNA transcription on enhancers adjacent to E2-upregulated coding genes. These induced eRNAs seem to have a major role in the ligand-dependent induction of target coding genes, thus underlining the important function of eRNAs in regulated programs of gene transcription (Li et al. 2013).

iv. LincRNAs (long intervening or intergenic ncRNAs)

LincRNAs are the most studied human ncRNAs, because their intergenic localization simplifies their analysis by avoiding the complications arising from overlap with other types of genes. Preferentially found within 10 kb of protein-coding genes, lincRNA genes exhibit similar transcription regulation, chromatin-modification patterns and splicing signals, but are generally shorter with fewer exons than protein-coding genes (Guttman et al. 2011; Guttman et al. 2010). An integrative approach defined a reference set of 4 662 lincRNAs that unifies existing annotation sources with transcripts reconstructed from >4 billion RNA-seq reads collected across 24 human tissues and cell types. This study showed that lincRNAs are expressed in a highly tissue-specific manner, much more than protein-coding genes. The median lincRNA level is only about a tenth that of the median mRNA level. No significant enrichment of correlated co-expression between lincRNAs and their neighbouring genes beyond that expected for any two neighbouring protein-coding genes was found. It has also

been shown that an additional set of 2 305 exhibit high evolutionary conservation and ambiguous coding potential, suggesting that they could function as ncRNA or as small peptides. These small-translated ORFs might prevent ribosome scanning or translation in downstream regions of the transcripts, thereby enabling the lincRNAs to perform noncoding functions in the cytoplasm without interference from the ribosome. They might also tether factors to ribosomes or modulate the stability of the lincRNA by influencing RNA decay pathways, some of which depend on translation (Cabili et al. 2011; Ulitsky & Bartel 2013). Less than 1% of these lincRNAs have been characterized, but several well-known examples, such as HOTAIR, MALAT1 or Xist, already show their importance in cell processes, during development and carcinogenesis.

v. VlincRNAs (very long intergenic ncRNAs)

Recently, a new class of 580 ncRNAs has been identified. Called very long intergenic ncRNAs (vlincRNAs) for their length comprised between 50 and 700 kb, these transcripts arise from intergenic regions of the human genome, but can partially overlap protein-coding genes. Identified in two types of tumors for the moment, they significantly overlap with lincRNAs found in normal human embryonic and stem cells, suggesting they might have a function in early development. Readily identifiable in tumors, particularly "stem cell like" tumors, they are less detectable in normal tissues, suggesting in addition a function in oncogenesis (Kapranov et al. 2010).

vi. NATs (Natural Antisense Transcripts)

Natural Antisense Transcripts (NATs) are multi-exonic, capped and polyadenylated transcripts, that share complementary exons with sense-paired genes (Khorkova et al. 2014; Su et al. 2010). Several examples of NATs, such as ANRIL, have been described as involved in the control of gene expression in human cells, through the recruitment of the PRC2 chromatin-modifying complex (Yap et al. 2010). Antisense transcription can also perturb sense gene expression through a transcriptional interference mechanism independent of a lncRNA itself, as exemplified by the ncRNA *Airn* (Latos et al. 2012). In addition, it has been shown in mouse that they can form RNA/RNA hybrids, trigger RNAi machinery and production of endo-siRNA (Carlile et al. 2009; Tam et al. 2008; Watanabe et al. 2008).

vii. T-UCRs (Transcribed Ultra-Conserved Regions)

Recently discovered, these ncRNAs are transcribed from ultra-conserved regions (UCRs) that have been identified by bioinformatics comparison between mouse, rat and human genomes. Among these 481 UCRs longer than 200 nt, that can be found in gene deserts or overlapping with protein-coding genes, a large fraction are transcribed in a tissue-specific pattern (Scaruffi 2011). It has been shown that malignant cells have a unique spectrum of expressed UCRs when compared with the corresponding normal cells, suggesting that T-UCRs might be involved in carcinogenesis. Distinct T-UCR expression signatures were also found between leukemias and carcinomas (Calin et al. 2007).

viii. LncRNAs containing repeated sequences: examples of TERRAs and SINEUPs

In a large variety of Eukaryotes, telomeres are transcribed from C-rich strand into 300 bp up to 100 kb lncRNAs named Telomeric Repeat-containing RNAs (TERRAs). These transcripts control telomeres length and chromatin structure, and are regulated by the developmental and physiological state of the cell (Azzalin et al. 2007; Schoeftner & Blasco 2010). Another type of lncRNAs containing repeated sequences has been identified recently. Termed SINEUPs, they contain a region antisense to the 5' untranslated region (UTR) of a target gene and short interspersed nuclear elements (SINEs). They do not affect the sense mRNA expression level, but enhance the translation of the encoded protein. One example is the lncRNA antisense to Uchl1, which transcription is under the control of stress signalling pathways, and which increases the synthesis of UCHL1 protein, involved in brain function and neurodegenerative disease (Carrieri et al. 2012).

b. Molecular functions of human long non-coding RNAs

Despite the high number of lncRNAs annotated in the past few years, the molecular function of a majority of them remains to be defined. On the other hand, several examples have been well studied, and show the importance of lncRNAs in the regulation of gene expression, at various levels. Two types of regulation can be distinguished: one that does not imply a lncRNA *per se* but its transcription, and one involving a lncRNA that can act as

molecular scaffold for proteins, pair to complementary RNA or DNA, or be a source of miRNAs (Figure 2) (Wilusz et al. 2009).

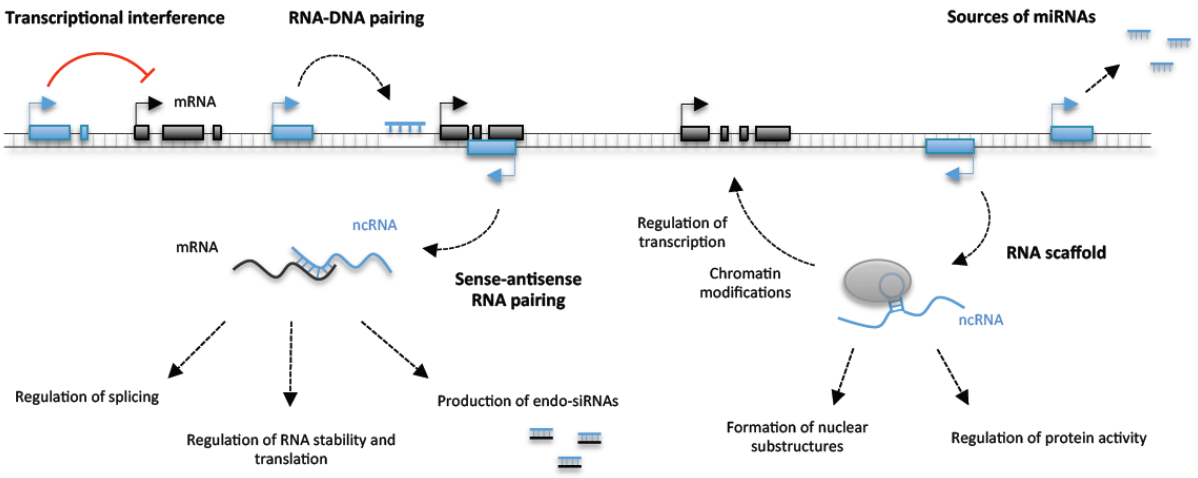

Figure 2 | Regulatory functions of lncRNAs. Transcription of a lncRNA *per se* can regulate the expression of adjacent or overlapping genes via transcriptional interference. LncRNAs can also influence transcription by RNA-DNA pairing, produce endo-siRNAs and regulate mRNA splicing, stability, translation by RNA-RNA pairing. LncRNAs can act as molecular scaffold for protein binding, regulating protein activity, transcription, chromatin modification and formation of nuclear substructures. LncRNAs can finally be sources of miRNAs. Adapated from Wilusz et al. 2009.

i. Transcriptional interference

The act of lncRNA transcription itself can influence transcription of adjacent or overlapping genes. For example, transcription through a regulatory element such as a promoter can block its function. First identified in yeast, this mechanism has been termed transcriptional interference (Martens et al. 2004). LncRNA transcription can prevent the binding of transcription factors or the RNA polymerase machinery, inhibiting transcription initiation or elongation. In this case, the lncRNA promoter is finely tuned to exert regulatory function, and the lncRNA product is a marker of transcriptional interference in action but is not required for its success (Batista & Chang 2013). *Airn* (antisense *Igf2r* RNA non-coding) lncRNA, paternally expressed from the Igf2r imprinted cluster, silences *in cis* the paternal alleles of *Igf2r*, *Slc22a3* and *Slc22a2* (Sleutels et al. 2002). Different mechanisms are involved, since Igf2r is silenced in all embryonic, extra embryonic and adult tissues that express *Airn*, whereas *Slc22a3* and *Slc22a2* are only silenced in some extra embryonic

lineages. In particular, *Igf2r* silencing occurs through silencing of its promoter by *Airn* transcription, but not by its spliced or unspliced lncRNA product (Latos et al. 2012).

ii. LncRNAs as molecular scaffolds for protein binding

LncRNAs exhibit the capacity to fold into very stable secondary and higher order structures, rendering them able to interact with proteins. Even with differences in their primary sequences, lncRNAs can fold into same types of secondary structure, thus allowing their interactions with the same protein (Pang et al. 2006; Mercer & Mattick 2013). On the other hand, multiple proteins can possess the same types of RNA-binding motifs, organized in combinatorial way, to allow association with very different RNA molecules (Lunde et al. 2007). Scaffold lncRNAs can target proteins to specific chromatin regions, regulating the establishment or the maintenance of particular chromosomal domains. They can also modulate proteins activity, or be involved in the formation of specific nuclear substructures.

- LncRNAs in regulation of chromatin organization

In Eukaryotes, nucleosomes are the fundamental unit of chromatin, constituted of 147 DNA base pairs wrapped around a histone octamer. All DNA-templated processes such as transcription, replication, repair and recombination, are regulated by chromatin highly dynamic structure and compaction. Two different chromatin states have been defined: euchromatin, open and easily transcribed, and heterochromatin, compact and inaccessible for transcription. Switch from one chromatin state to another is ensured by the epigenetic status of a locus, through histone modifications and DNA methylation, tightly regulated by chromatin modifying complexes. Euchromatin is characterized by enrichment in acetylated and histone H3 lysine 4 (H3K4) trimethylated nucleosomes, while heterochromatin is enriched in H3K27 and H3K9 methylated and deacetylated nucleosomes. H3K4 methylation is ensured by the MLL family of methyltransferases, and reverted by the LSD1 histone demethylase (Kouzarides 2007). Polycomb Repressive Complex 2 (PRC2) and G9a histone methyltransferase regulate H3K27 and H3K9 methylation, respectively (Zaidi et al. 2010). High cytosine methylation by DNA methyltransferases in promoter proximal CpG islands is also a characteristic of repressed genes (Zentner & Henikoff 2013). Numerous lncRNAs have been shown to play a fundamental role in the regulation of epigenetic landscape through

binding and *cis*- or *trans*-targeting of histone modifying complexes to specific regions of the genome (Figure 3).

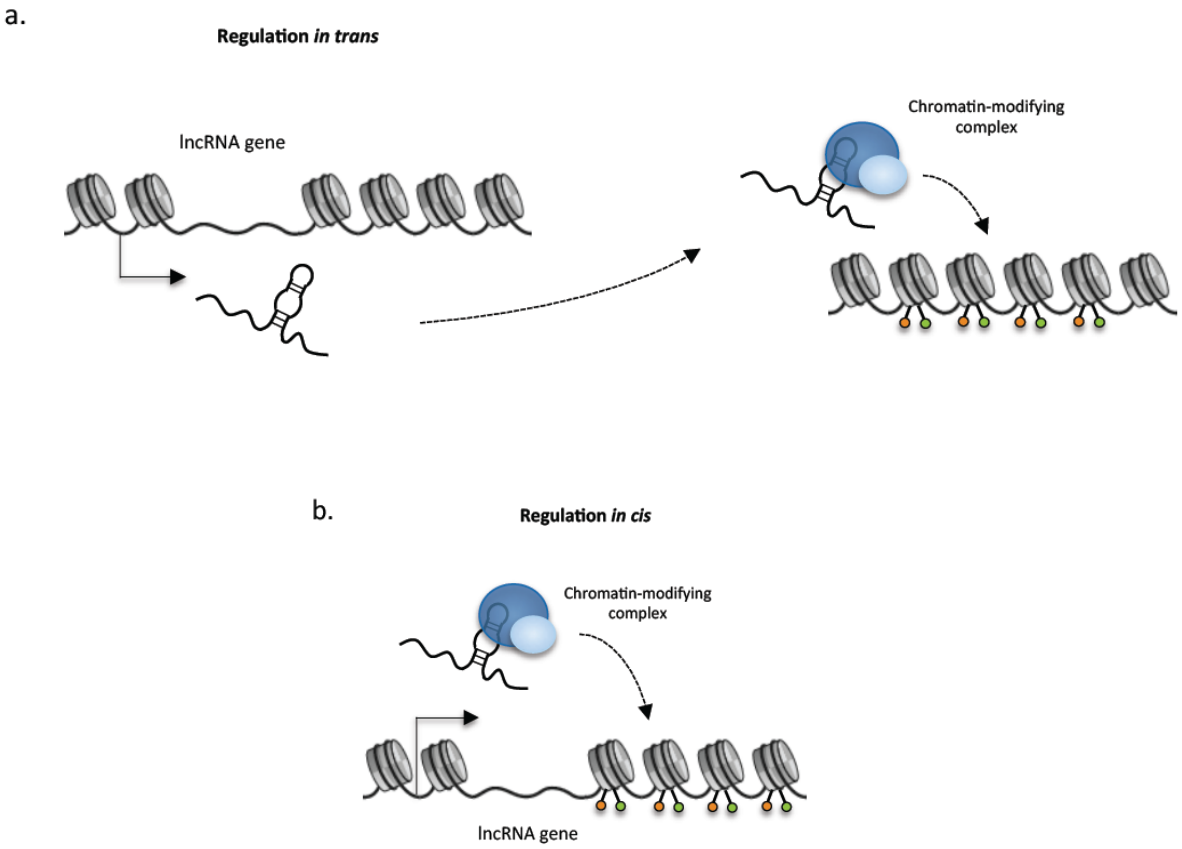

Figure 3 | *Cis*- and *trans*-regulation of chromatin organization by lncRNAs. LncRNAs can regulate epigenetic landscape **a.** on other chromosomes (*in trans*), or **b.** in close spatial proximity (*in cis*) or through binding and targeting of chromatin-modifying complexes to specific regions of the genome. Chromatin-modifying complexes can promote either activation of transcription (activating histone mark, in green), or gene silencing (repressive histone mark, in orange).

Cis-acting lncRNAs target genomic loci in close spatial proximity to a lncRNA gene. Nascent ANRIL (antisense non-coding RNA in the INK4 locus) has been shown to recruit Polycomb Repressive Complex 1 (PRC1) and PRC2, via binding to their subunits CBX7 and SUZ12, respectively, inducing H3K27 methylation and silencing of INK4b/ARF/INK4a locus (Figure 4a) (Yap et al. 2010). Kcnq1ot1 is a 91 kb-long antisense transcript, which promoter is located in intron 10 of the Kcnq1 gene. This promoter is methylated on the maternal chromosome, and unmethylated on the paternal chromosome. Production of Kcnq1ot1 RNA on the paternal chromosome induces silencing of eight to ten protein-coding genes, spread

over a 1 mega base (Mb) region, through interaction with EZH2 and Suz12 subunits of PRC2 but also with G9a histone methyltransferase (Pandey et al. 2008). Regulation by *cis*-acting lncRNAs can be spread over longer distance and even to a whole chromosome, as demonstrated by Xist lncRNA function. LncRNA Xist (X-inactive specific transcript) play a central role in X-chromosome inactivation (XCI), process in mammals that ensures equal transcripts levels between males and females by genetic inactivation of one of the two X chromosomes in females. The processed Xist transcript coats and silences the entire X-inactive chromosome, by recruiting chromatin remodelling complexes including PRC2 which induces H3K27 trimethylation (Pontier & Gribnau 2011).

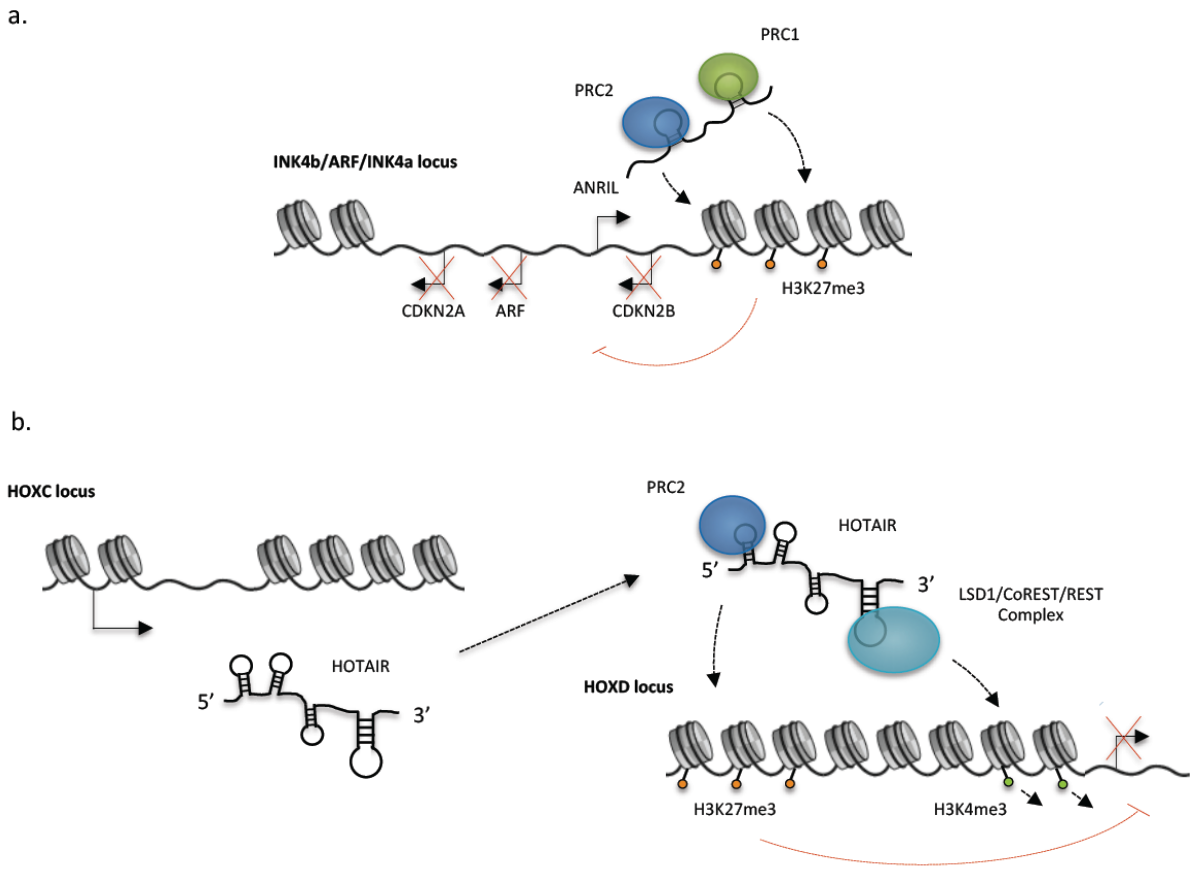

Figure 4 | ANRIL and HOTAIR examples of cis- and trans-regulation of chromatin organization. **a.** ANRIL is transcribed from the INK4b/ARF/ARF4a locus, and recruits PRC1 and PRC2 complexes to promote H3K27 trimethylation and induce gene silencing on the same locus. **b.** HOTAIR is transcribed from the HOXC locus, and silences *in trans* the HOXD locus, via recruitment of PRC2 and LSD1 complexes, which promotes trimethylation of histone H3K27 and demethylation of H3K4, respectively.

Trans-acting lncRNAs affect the transcription of genes located on other chromosomes. One of the most studied lncRNAs is HOTAIR (*HOX* antisense intergenic RNA). Transcribed from the *HOXC* locus on chromosome 12, spliced and polyadenylated, this 2.2 kb lncRNA has been shown to induce transcriptional silencing of the *HOXD* locus localized on chromosome 2, among multiple genes across the genome. Silencing is achieved by HOTAIR binding to PRC2 via EZH2 subunit, but also HOTAIR interaction with CoREST/REST repressor complex via its subunit LSD1, promoting both H3K27 trimethylation and H3K4 demethylation (Figure 4b) (Rinn et al. 2007; Tsai et al. 2010).

How lncRNAs target proteins to specific genomic locations is still poorly described. Some lncRNAs have been shown to associate with DNA binding proteins at their target genes. As an example, Xist lncRNA binds to YY1, a bivalent protein capable of binding both RNA and DNA through different sequence motifs, allowing its loading onto X chromosome (Jeon & Lee 2011). Other lncRNAs, such as HOTAIR, can interact directly with chromatin through specific DNA-binding motifs (Chu et al. 2011).

- Ribo–repressors and –activators of transcription

Several examples of lncRNAs have been shown to exert regulatory functions through binding with transcription activators or repressors, or RNAPII, promoting or preventing their binding to promoters. For example, the lincRNA-p21 is transcribed next to p21 (Cdkn1a) gene. LincRNA-p21 binds hnRNP K, a RNA-binding protein well known as a transcriptional repressor, and mediates its binding to target genes, ultimately leading to gene silencing (Huarte et al. 2010). Evf-2 lncRNA, transcribed from the Dlx-5/6 ultraconserved region, forms a stable complex with the Dlx-2 transcription factor, promoting its binding to the Dlx-5/6 enhancer to increase its transcriptional activity (Feng et al. 2006). Gas5 (Growth arrest-specific 5) lncRNA binds to the DNA-binding domain of the glucocorticoid receptor (GR), competing with DNA glucocorticoid response element for binding to the GR, and preventing GR-mediated transcriptional activation (Kino et al. 2010). PANDA (P21 associated ncRNA DNA damage activated) is one of the five lncRNAs transcribed from the p21 promoter (CDKN1A) upon DNA damage. This non-spliced 1.5 kb lncRNA inhibits the expression of genes by sequestering the transcription factor NF-YA from occupying target gene promoters (Hung et al. 2011). Enhancer-like RNAs interacts with the Mediator complex, forming

chromatin loops at gene promoters and favouring long-range transcriptional activation (Lai et al. 2013).

- Scaffold RNAs as regulators of proteins' activity

LncRNAs binding to proteins can affect their activities. For example, TERRAs are transcribed from telomeric and subtelomeric regions. In vitro, they have been shown to bind the reverse transcriptase subunit of telomerase (TERT), competing with telomeric DNA and inhibiting telomerase function (Redon et al. 2010). Gastric carcinoma high expressed transcript 1 (GHET1) lncRNA physically associates with insulin-like growth factor 2 mRNA binding protein (IGF2BP1), enhancing its physical interaction with c-Myc mRNA, thus resulting in increased stability of c-Myc mRNA and expression (Yang et al. 2014). In addition to previously described interactions with PRC2 and LSD1 complexes, HOTAIR has been recently shown to associate with E3 ubiquitin ligases Dzip3 and Mex3b, bearing RNA-binding domains. Dzip3 and Mex3b are involved in the ubiquitination of Ataxin-1 and Snurportin-1, respectively. HOTAIR facilitates this process, accelerating degradation of these proteins (Yoon et al. 2013).

- Formation of nuclear substructures

Several lncRNAs are involved in the formation of specific substructures, as exemplified by well-known NEAT1 and NEAT2/MALAT-1 ncRNAs. NEAT1 is a 4 kb, unspliced, polyadenylated, nuclear-restricted ncRNA. It has been shown to be necessary for the formation of specific nuclear structures, the paraspeckles. Like most nuclear structures, the specific function and formation of paraspeckles is not fully understood, but paraspeckles-associated proteins PSP1 and p54 are implicated in pre-mRNA splicing, transcription regulation and nuclear retention of RNA. NEAT1 not only binds PSP1 and p54, but also is required for their specific localization (Clemson et al. 2010). MALAT-1 (metastasis-associated lung adenocarcinoma transcript 1) is a 8 kb, highly conserved and nuclear-restricted lncRNA. In the nucleus, nuclear speckles are highly dynamic subnuclear domains enriched with pre-mRNA splicing and processing factors. They are thought to be involved in the assembly, modification and storage of the pre-mRNA splicing machinery. MALAT-1 interacts with serine/arginine (SR) splicing factors, modulating their distribution to nuclear speckles. It has also been shown that MALAT-1 regulates alternative splicing of pre-mRNAs

by controlling the functional levels of SR splicing factors (Figure 5) (Tripathi et al. 2010; Bernard et al. 2010).

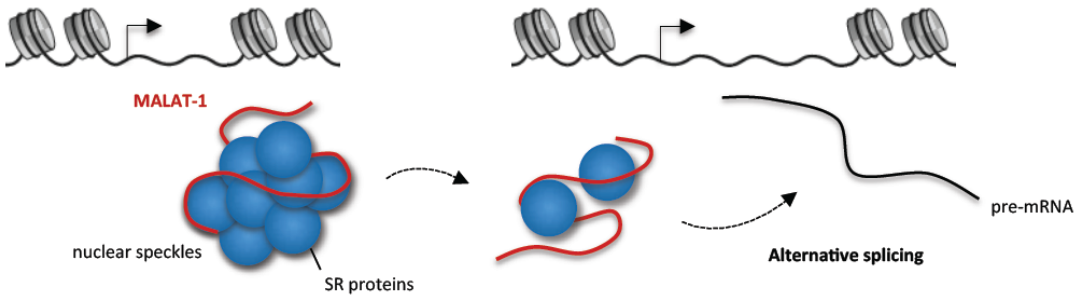

Figure 5 | MALAT-1 lncRNA regulatory role. MALAT-1 interacts with and regulates the distribution of serine/arginine splicing factors to nuclear speckles. MALAT-1 is also involved in the regulation of pre-mRNAs alternative splicing, by controlling the functional levels of SR splicing factors.

iii. LncRNA-RNA pairing: regulation of RNA processing, stability and translation

Based on Watson-Crick complementarity, lncRNAs can pair with other RNA molecules, with various consequences on processing, stability and translational potential of a complementary RNA target. This process is often observed for lncRNAs antisense to protein-coding genes. For example, ZEB2 NAT pairs with ZEB2 mRNA, promoting the retention of the IRES-containing first intron within the ZEB2 pre-mRNA. This pairing results in enhanced ZEB2 mRNA translation, and increased ZEB2 protein levels within the cell (Beltran et al. 2008). Cytoplasmic, polyadenylated Alu repeats-containing lncRNAs have been shown to be involved in a particular mechanism. Imperfect base pairing between Alu element in a lncRNA and the 3' UTR of a translationally active mRNA induce the formation of a Staufen 1 (STAU1)-binding site, thus resulting in mRNA degradation by this protein. With this process, an individual lncRNA can down regulate a subset of STAU1-mediated messenger RNA decay (SMD), and distinct lncRNA can down regulate the same SMD target (Gong & Maquat 2011).

Two major roles of lncRNA-RNA pairing have also been described in the context of miRNA pathway. LncRNAs can block miRNA-induced mRNAs degradation or translation

arrest, by masking miRNA-binding sites within the target mRNA. They can also sequester miRNAs, acting as "sponges" to prevent miRNA binding to their target mRNAs. The most known example is a lncRNA transcribed from PTENP1 pseudogene, that exhibits the same sequence as the 3' UTR of PTEN protein-coding gene. Thus, PTENP1 lncRNA binds miRNAs targeting PTEN mRNA, resulting in its stabilization (Poliseno et al. 2010). Another well-known transcript in this category is the highly up-regulated in liver cancer (HULC) spliced, polyadenylated and cytoplasmic lncRNA. It has been demonstrated that HULC act as an endogenous "sponge", which down regulates a series of miRNAs activities. For example, inhibition or miR-372 leads to reduced translational repression of its target gene, PRKACB (Figure 6) (Wang et al. 2010). β-secretase-1 (BACE1)-AS antisense lncRNA form a RNA-RNA duplex with the BACE1 mRNA in the same region as miR-485-5p, therefore preventing miRNA-mediated BACE1 decay (Faghihi et al. 2008).

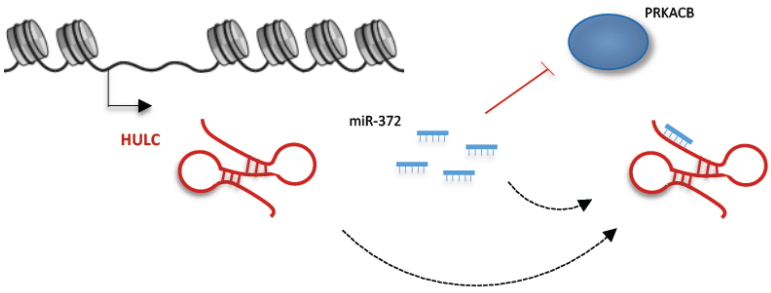

Figure 6 | HULC as a molecular "sponge" for miRNAs. HULC lncRNA binds miRNAs, including miR-372. HULC reduces the availability of this miRNA for its natural target gene, PRKACB. HULC overexpression thus leads to reduced translational repression of PRKACB.

iv. LncRNA-DNA pairing: regulation of transcription and DNA recombination

LncRNAs are able to pair with double-stranded or single-stranded DNA, forming specific structures involved in gene regulation. For example, a lncRNA is transcribed from the minor upstream promoter of the gene encoding dihydrofolate reductase (DHFR). This lncRNA pairs to the DNA sequence to form a stable complex with the major promoter of DHFR and concurrently interacts with the general transcription factor IIB, impeding the formation of the

pre initiation complex from the major promoter and inducing transcriptional repression of DHFR (Martianov et al. 2007).

Formation of RNA-DNA duplex hybrids, or R-loop structures, in chromatin complex can cause the displacement of a one DNA strand. This mechanism have been observed for lncRNAs identified as major actors of specific immunological phenomena. All vertebrates throughout their lifetimes encounter innumerable pathogenic and environmental insults. A recognition system capable of identifying these infinitely diverse particles as nonself is required, with a demand for diversity that exceeds the amount of information that can be encoded in one specific locus. LncRNAs regulation is one of the mechanisms that allow this diversity to be created, in particular in the context of the immunoglobulin Class-Switch Recombination (CSR). This mechanism changes B cells production of antibodies from one isotype to another, swapping the initially expressed constant region for an alternate downstream region to change the effector function associated with a particular immunoglobulin. R-loop, formed by a G-rich lncRNA hybridized to its cognate DNA, facilitates the Activation-Induced Deaminase (AID) targeting to the locus during CSR initiation (Teng & Papavasiliou 2009; Matthews et al. 2014).

v. LncRNAs as a source of miRNAs

Several examples of lncRNAs are processed into miRNAs. Integrative transcriptome analysis revealed that conservation of lncRNAs across vertebrates is generally restricted to patches, suggesting that discrete functional domains are present into one lncRNA molecule. Indeed, analysis of genome-wide transcriptome databases compared to small RNA deep sequencing datasets revealed that lncRNAs harbour miRNA clusters, suggesting that they could be processed into small RNAs. In the same study, previously mentioned PTENP1 lncRNA has been described harbouring five small RNA clusters, which function is still unknown (Jalali et al. 2012). The H19 locus produces a 2.3 kb lncRNA, which transcription is regulated by a complex interplay of factors. In addition to its contribution to the strong oncogenic behaviour of c-Myc, H19 is involved in a molecular mechanism integrating p53 and hypoxia-inducible factor 1-α with the hypoxic stress response. Furthermore, H19 is a precursor of miR-675, which is excised from its exon one, and targets specific mRNAs (Tsang et al. 2010; Matouk et al. 2007; Matouk et al. 2010; Matouk et al. 2014).

B. Carcinogenesis and Epithelial-Mesenchymal Transition

1. Cancer: a general overview

The term "cancer" encompasses more than 100 different diseases, all characterized by an uncontrolled and unrestrained proliferation of cells. These cells have escaped to the normal controls of cell division (neoplastic), and have acquired the capacity to invade and colonize other tissues, with the potential to kill the organism.

The cell type and the embryologic origin of the tissue from which a cancer arise allow its classification. A tumor arising from epithelial cells, which derive from ectoderm or endoderm, is termed carcinoma, whereas a tumor from connective tissue or muscle cells, which derive from mesoderm, is termed sarcoma. These categories do not include leukemias and lymphomas, which derive from hemopoietic tissues that grow as individual cells in the blood, or as a solid mass, respectively, and cancers derived from the nervous system, such as glioma. Many subdivisions have been defined for each type of cancer, according to specific cell type and location in organisms.

Cells growing as a self-contained mass form a tumor said to be benign. As the tumor grows in size, it acquires the capacity to stimulate formation of new blood vessels, in a process called angiogenesis. In addition, acquisition of migratory properties allows the cells to invade the surrounding tissue, spread by lymphatic and circulatory systems, and establish distant secondary areas of growth. This process is called metastasis, and the tumor become malignant. Aside from the consequences arising from the expansion of a cell mass in a tissue, the metastatic process is the most important factor determining cancer as a life-threatening disease. Indeed, the more widely a cancer spreads, the harder it becomes to eradicate, and an estimation revealed that more than 90% of all cancer deaths are associated with metastasis (Spano et al. 2012).

Analyses of cancer incidence and mortality revealed that about 90% of human cancers arise from epithelial tissues. It could be due to the fact that most of the cell proliferation in the body occurs in epithelia, or that epithelial tissues are frequently exposed to physical and chemical damages that favour the development of cancer. Leukemias, as well as brain and central nervous system cancers, have the highest incidence and death rate found among

children. Among adult men, the most frequent is prostate cancer, whereas adult women are affected mostly by breast cancer. However, lung and bronchus cancers are the leading causes of death for both sexes. The incidence of these tumors increases exponentially with aging (Figure 7) (Institut National du Cancer, American Cancer Society) (Depinho 2000; Jemal et al. 2008).

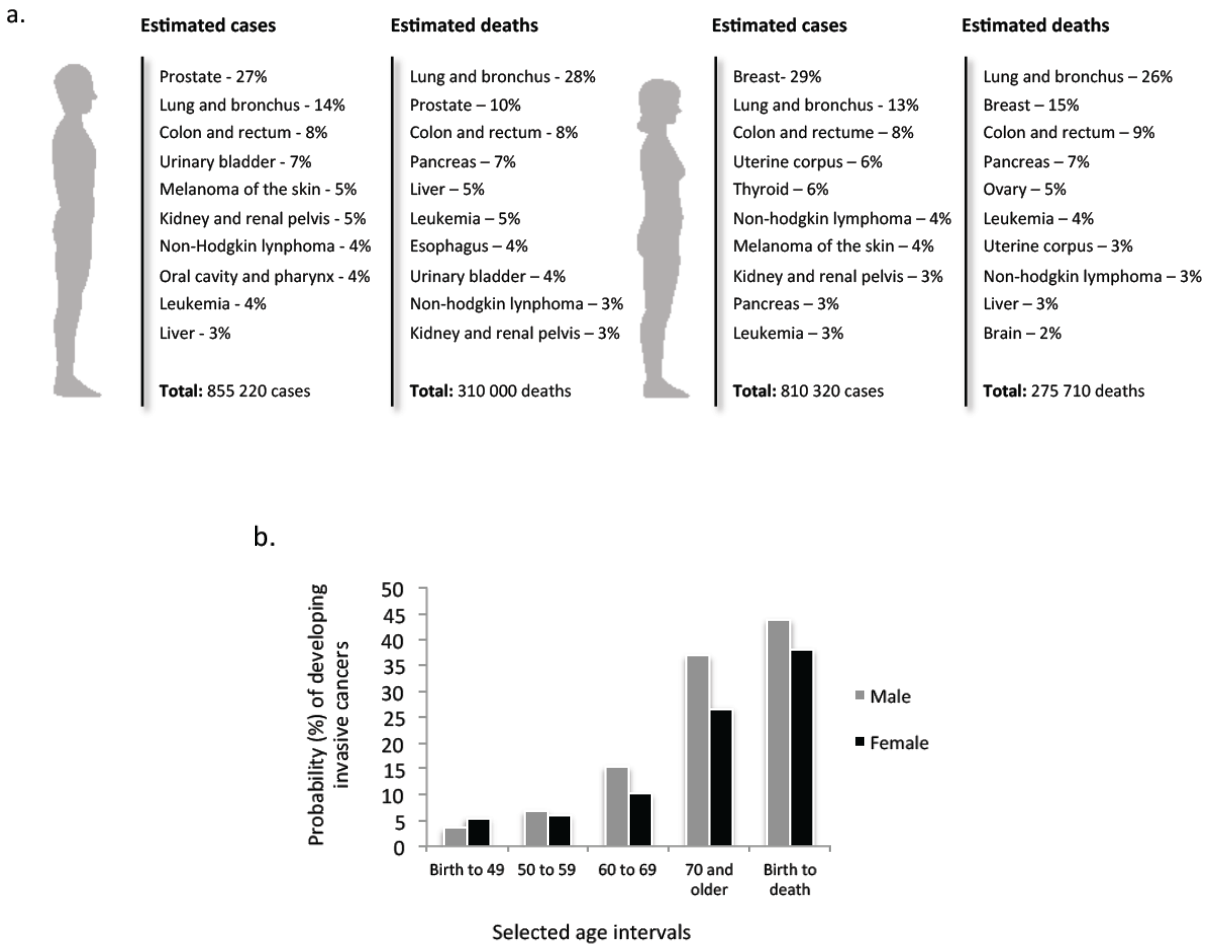

Figure 7 | Estimated cancer incidence in the US in 2014. **a.** Estimated numbers of new cancer cases and deaths, in male and female, with percentages of each cancer types. **b.** Probability of developing invasive cancers during selected age intervals (for those who are cancer-free at the beginning of each age interval), for male and female, and based on cancer statistics obtained between 2008 and 2010. Source: American Cancer Society, Surveillance Research, 2014.

a. The multistep progression model

The evolution of a normal cell into a cancer cell requires multiple heritable changes within the cell, defining carcinogenesis as a multistep process involving multiple genes. This process

is often separated into three stages, initiation, promotion and progression, each divided into several steps involving numerous genetic changes, which are still not fully characterized. Initiation involves the induction of irreversible alterations in a cell, which are frequently mutational events. Promotion is the process by which initiated cell clonally expands into a visible tumor, often a benign lesion at this stage. Progression is the evolution of a benign tumor into a malignant cancer (Figure 8) (Fearon & Vogelstein 1990; Barrett 1993).

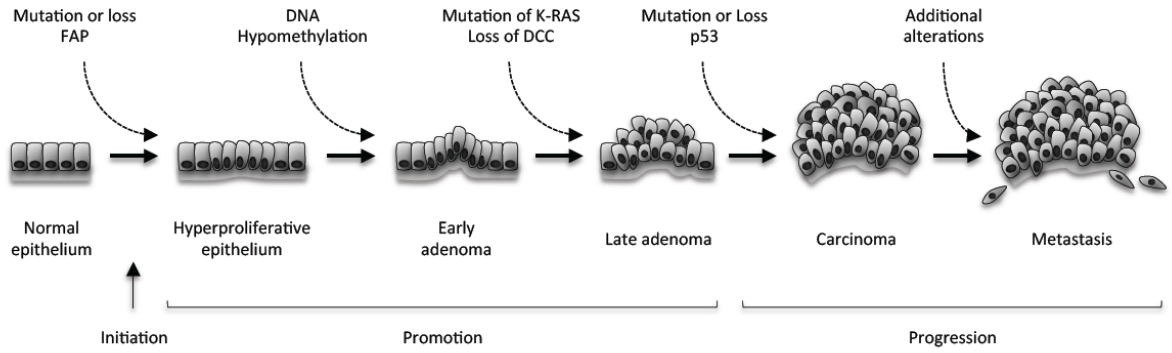

Figure 8 | The multistep progression model of colorectal cancer. The initiation step is a mutation or a loss of FAP gene, promoting the hyper proliferation of epithelial cells. A series of genomic alterations then induce the formation of an adenoma, and mutation or loss of p53 tumor-suppressor gene promotes tumor progression. Additional genomic alterations can lately initiate the metastatic process. Adapted from Fearon & Vogelstein, 1990.

Among the numerous genes that are affected at each of these stages, three distinct classes have been defined: oncogenes, tumor-suppressor genes, and "caretaker" genes. Oncogenes are genes which expression promotes cancer evolution. Activation of these genes often results from chromosomal translocations, gene amplifications or subtle intragenic mutations affecting crucial residues that regulate the activity of the gene product. An activating somatic mutation in one allele of an oncogene is generally sufficient to confer a selective growth advantage on the cell. On the contrary, tumor-suppressor genes show reduced activity after genetic alterations such as missense mutations at essential residues of the gene product, mutations that result in a truncated protein, deletions or insertions of various sizes, or epigenetic silencing. Mutations in both maternal and paternal alleles of these genes are generally required to confer a selective advantage to the cell, this situation arising commonly from deletion of one allele via a major chromosomal event coupled with an intragenic

mutation on the other allele. Caretaker genes promote carcinogenesis in a completely different way. Indeed, this class includes genes involved in repair of mistakes made during DNA replication, and involved in control processes such as mitotic recombination and chromosomal segregation. These genes keep genetic alterations to a minimum. When they are inactivated, mutations in other genes occur at a higher rate (Vogelstein & Kinzler 2004).

The progressive accumulation of genetic alterations over a long period of time gives cells selective advantages. Indeed, it has been shown by epidemiological studies that the incidence of sporadic epithelial cancers increases as a function of the sixth power of age (Armitage & Doll 1954). Both oncogenes and tumor-suppressor genes, when altered by genome instability, drive the neoplastic process by increasing tumor cell number through stimulation of cell growth and division, or inhibition of cell death or cell cycle arrest. It has been suggested that five or six alteration events are enough to drive the carcinogenesis process, each event increasing the probability of the next one upon rounds of clonal expansion. Three key elements have been identified in the conversion of a normal cell into a cancer cell: inactivation of p53/Rb pathways, immortalization through expression of telomerase (hTERT) and expression of an oncogenic form of RAS (Hahn et al. 1999).

b. The hallmarks of cancer

If different genetic alterations are involved in cancer initiation and development, it has been shown that all cells with fully malignant potential share common acquired physiological capacities, termed hallmarks of cancer (Figure 9) (Hanahan et al. 2000; Hanahan & Weinberg 2011).

i. Self-Sufficiency in growth signals

Growth signals (GS) are required for normal cells to move from a quiescent state into an active proliferative state. The transmission of these signals is achieved via the binding of distinct classes of signalling molecules by transmembrane receptors of the cell: diffusible growth factors, cell-to-cell adhesion/interaction molecules, and extracellular matrix components. Several oncogenes act by mimicking normal growth signalling, leading cells to generate their own GS and actively proliferate (Hanahan et al. 2000).

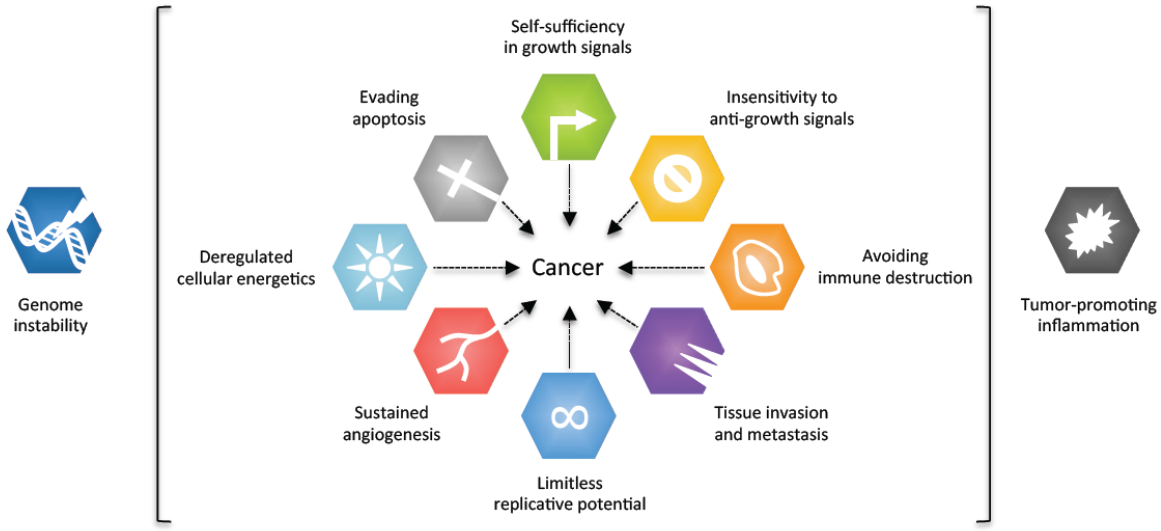

Figure 9 | The hallmarks of cancer. Cancer cells share acquired characteristics, giving them fully malignant capacities. Originally, six hallmarks have been defined: self-sufficiency in growth signals, insensitivity to anti-growth signals, tissue invasion and metastasis, sustained angiogenesis, limitless replicative potential, and resistance to apoptosis. Two additional hallmarks have been described more recently: the deregulation of cellular energetics, and the resistance to immune destruction. Two processes are described as responsible for tumor initiation: genome instability, and tumor-promoting inflammation. Adapted from Hanahan et al. 2000, Hanahan & Weinberg 2011.

ii. Insensitivity to antigrowth signals

Multiple antiproliferative signals are necessary for the maintenance of cellular quiescence and tissue homeostasis, including soluble growth inhibitors and immobilized inhibitors, in the extracellular matrix and on the surface of nearby cells. Cancer cell evade these antiproliferative signals, for example by a disruption of the retinoblastoma (Rb) pathway (Weinberg 1995; Burkhart & Sage 2008).

iii. Evading apoptosis

Apoptosis is a programmed cell death, orchestrated by two classes of components, sensors and effectors. Sensors detect abnormalities in extra and intracellular environment, such as DNA damages or hypoxia. Effectors induce in response the death pathway. Resistance to apoptosis can be acquired by cancer cells through different strategies, but the most common involves the p53 tumor suppressor gene and inactivation of p53 protein, resulting in the

removal of a key component of the DNA damage sensor system (Evan 1998; Adams & Cory 2007; Lowe et al. 2004).

iv. Limitless replicative potential

Growth signal autonomy, insensitivity to antigrowth signals, and resistance to apoptosis lead to a dissociation between a cell's growth program and signals in cell's environment. However, if resulting in the generation of a vast cell population, these mechanisms are not sufficient to ensure expansive tumor growth. Indeed, many, if not all, mammalian cells carry an intrinsic program that limits their multiplication.

Cells in culture can progress through a certain number of population doublings before they stop growing, a process termed senescence (Hayflick 2000; Hayflick 2003). Disabling pRb and p53 tumor-suppressor proteins allow the cells to grow for additional generations, but they quickly enter into crisis, characterized by massive cell death. Occasionally, a cell variant can emerge with the capacity to multiply without limit, a trait called immortalization (Wright et al. 1989). This limitless replicative potential is acquired during tumor progression and is essential for the development of their malignant growth state. Normal cells have the capacity for 60-70 population doublings, this generational limit being a barrier to cancer.

The factor limiting cell proliferation is the existence of specific nucleoprotein structures at the end of chromosomes, termed telomeres. Telomeres are constituted of long arrays of double stranded TTAGGG repeats, G-rich 3' single strand overhang, and a core of six associated telomere-specific proteins termed shelterin or telosome (Figure 10) (Artandi & DePinho 2010; Blasco 2007). The length of telomeres determines the replicative potential of cells, as they progressively shorten with each mitotic division. This shortening is explained by the inability of DNA polymerase to fully replicate the 3' end of chromosomal DNA during each S phase. Telomeres protect the ends of chromosomal DNA, and their progressive erosion eventually causes the lost of this ability. Unprotected chromosomal ends participate in end-to-end chromosomal fusions (Counter et al. 1992). Telomeres length declines with aging in humans, suggesting that telomere dysfunction and telomere-induced senescence might increase with aging (Frenck et al. 1998).

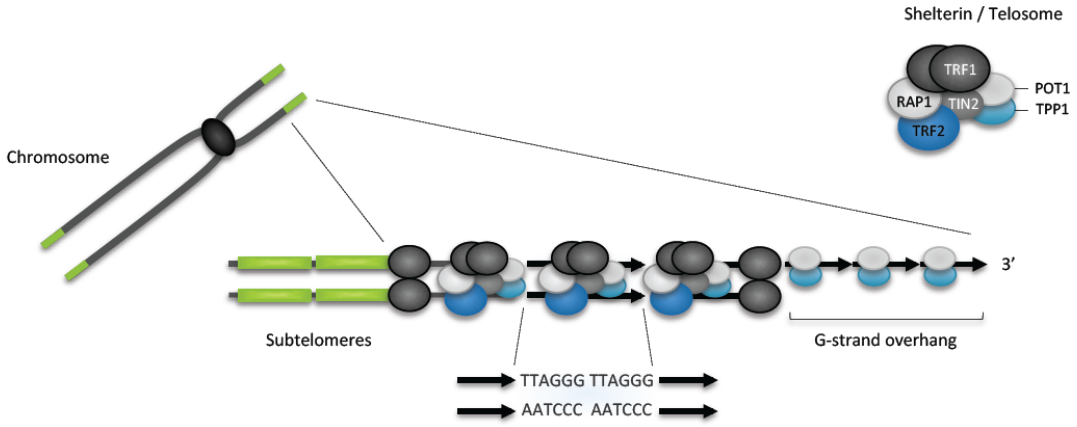

Figure 10 | Telomeres structure. Located at the end of chromosomes, telomeres are constituted of tandem repeats of the TTAGGG sequence, bound by the shelterin complex, and a G-rich 3' single strand overhang. Adapted from Blasco 2007.

The maintenance of telomeres is ensured by telomerase, a ribonucleoprotein consisting of a reverse transcriptase (TERT) and a RNA template (TERC), that mediates the addition of telomere repeat sequences to the 3' overhang (Greider & Blackburn 1989). Ectopic expression of telomerase is used to immortalize cultured cells, whereas endogenous telomerase up regulation is a frequent event in human cancers (Shay & Bacchetti 1997; Bodnar 1998).

v. Sustained angiogenesis

Oxygen and nutrients are crucial for cell survival and functions, as well as the ability to evacuate metabolic wastes and carbon dioxide, forcing all cells in a tissue to grow close to a capillary blood vessel. During embryogenesis, this closeness is ensure by a tightly regulated birth of new endothelial cells and their assembly into tubes. In the adult, angiogenesis occurs transiently, for specific processes such as wound healing. In contrast, during tumor progression, angiogenesis is always activated, allowing formation of new vessels that sustain expanding neoplastic growth (Hanahan & Folkman 1996).

vi. Tissue Invasion and Metastasis

Carcinomas arising from epithelial tissues can progress to higher pathological grades of malignancy, by local invasion and distant metastasis. Associated cells develop alteration in their shape, but also in their attachment to other cells and to the extracellular matrix (ECM). One key cell-to-cell adhesion molecule altered in this process is E-cadherin. This protein is involved in the formation of adherens junctions with adjacent epithelial cells, and allows the assembly and stability of quiescent epithelial cell sheets. E-cadherin is frequently down regulated, and can be inactivated by occasional mutations, in human carcinomas (Berx & van Roy 2009; Cavallaro & Christofori 2004). In addition, expression of multiple genes encoding cell-to-cell and cell-to-ECM adhesion proteins is altered in highly aggressive carcinomas, whereas adhesion molecules normally associated with the cell migrations that occur during embryogenesis and inflammation are up regulated (Cavallaro & Christofori 2004).

The process of invasion and metastasis has been schematized as a sequence of discrete steps, termed the invasion-metastasis cascade (Fidler 2003; Talmadge & Fidler 2010). This cascade involves a succession of cell changes, first with a local invasion, followed by intravasation by cancer cells into nearby blood and lymphatic vessels. Cancer cells then transit through the lymphatic and haematogenous systems, and escape from the lumina of the vessels into the parenchyma of distant tissues, a process termed extravasation. At this step, cells form small nodules, termed micro metastases, which can grow into macroscopic tumors, in a process termed colonization. The mean by which transformed epithelial cells acquire the abilities to invade, resist apoptosis and disseminate is not fully understood. However, a developmental regulatory program, the Epithelial-to-Mesenchymal Transition (EMT), has become prominently implicated (Barrallo-Gimeno & Nieto 2005; Yilmaz & Christofori 2009; Polyak & Weinberg 2009; Klymkowsky & Savagner 2009).

vii. Emerging hallmarks

In addition to the six previously described and well-known hallmarks of human cancers, two particular attributes of cancer cells might be functionally important (Colotta et al. 2009; Luo et al. 2009; Negrini et al. 2010). The first one is the reprogramming of energy metabolism within the cell, in order to fuel cell growth and division (DeBerardinis et al. 2008; Jones & Thompson 2009; Hsu & Sabatini 2008). Secondly, cancer cells avoid recognition and

destruction by the immune system (Kim et al. 2007; Teng et al. 2008). If the mechanism is not well described, studies have shown that cancer cells can disable several components such as CD8+ cytotoxic T lymphocytes (CTLs) or natural killer (NK) cells by secretion of immunosuppressive factors (Yang et al. 2010; Shields et al. 2010). Another observed mechanism is the recruitment of inflammatory cells that are actively immunosuppressive (Mougiakakos et al. 2010; Ostrand-Rosenberg & Sinha 2009).

c. Acquisition of cancer hallmarks

Cancer cells can survive, disseminate and proliferate thanks to the acquired capabilities previously defined. The acquisition of these hallmarks is not completely understood, however two mechanisms have already been implicated: genomic instability and inflammatory state of premalignant and malignant lesions.

i. Genome instability

As previously mentioned, succession of genome alterations increases the probability to develop a malignant tumor from normal cells, by conferring selective advantages on subclones of cells through oncogenes expression and tumor-suppressor genes down regulation, ensuring their dominance in a local tissue environment. Thus, genome instability is a central player in the acquisition of cancer hallmarks.

If the genomic maintenance machinery normally ensures a very low rate of spontaneous mutations, loss of one or several of its components in cancer cells considerably increases genome mutability (Negrini et al. 2010; Salk et al. 2010; Gisselsson et al. 2001). These components include genes whose products are involved in the detection and repair of DNA damages, or in the inactivation and interception of mutagenic molecules before they damage the DNA (Ciccia & Elledge 2011; Jackson & Bartek 2010; Harper & Elledge 2007; Kastan 2008). Alterations in these genes further accelerate the accumulation of mutations. In addition, loss of telomeric DNA generates karyotypic instability, allowing amplification, deletion and fusion of chromosomal segments (Artandi & DePinho 2010).

Development of advanced technologies, such as comparative genomic hybridization (CGH) and high-throughput DNA sequencing, allowed to identify the most frequent genomic

aberrations and mutations in cancer cells. Importantly, specific amplifications, deletions and mutations are recurrent at particular sites in the genome, indicating genes whose alteration favour neoplastic progression in different tumor types (Korkola & Gray 2010).

ii. Tumor-promoting inflammation

Cells from the immune system can densely infiltrate tumors and favour their development. This idea emerged from the observation that inflammatory disorders induce an increased risk for cancer development, as exemplified by colon carcinoma, associated with inflammatory bowel disease (Triantafillidis et al. 2009), stomach cancer after *Helicobacter pylori* infection (Correa 2003), and hepatocellular carcinomas after hepatitis C infection (El-Serag 2002).

Inflammatory cells, including granulocytes, dendritic cells, macrophages, natural killer cells and mast cells are present in the carcinoma microenvironment, at densities ranging from subtle infiltrations to gross inflammations. If it was previously believed that the presence of these cells reflected an attempt by the immune system to eradicate tumors, it is now clear that this tumor-associated inflammatory response enhances the acquisition of cancer capabilities by neoplastic cells. This support can be achieved by supplying specific molecules to the tumor environment, such as growth factors sustaining proliferative signalling, survival factors limiting cell death, proangiogenic factors, and extracellular matrix-modifying enzymes facilitating angiogenesis, invasion and metastasis (Coussens & Werb 2002; DeNardo et al. 2010; Grivennikov et al. 2010; Karnoub & Weinberg 2007).

2. The Epithelial-to-Mesenchymal Transition

a. What is the EMT?

The epithelial-to-Mesenchymal transition is a highly conserved fundamental process by which an epithelial cell undergoes multiple biochemical changes to assume a mesenchymal cell phenotype (Kalluri & Weinberg 2009). Epithelial cells are characterized by cohesive interactions between cells, existence of three membrane domains (apical, lateral and basal), presence of tight junctions between apical and lateral domains, apicobasal-polarized distribution of organelles and cytoskeleton components, and lack of mobility of individual epithelial cells with respect to their local environment. On the contrary, mesenchymal cells

have loose interaction among cells so that no continuous cell layer is formed, have no clear apical and lateral membranes, and have no apical-polarized distribution of organelles and cytoskeleton components. In addition, mesenchymal phenotype includes enhanced migratory capacities, invasiveness, elevated resistance to apoptosis and increased production of ECM components (Larue & Bellacosa 2005). Once EMT is completed, the underlying basement membrane is degraded, and mesenchymal cells can migrate away from the epithelial layer in which they originated.

Several distinct molecular regulatory networks are necessary to initiate and complete an EMT, including activation of transcription factors, expression of specific cell-surface proteins, cytoskeletal proteins activation and reorganisation, production of ECM-degrading enzymes, and changes in the expression of specific miRNAs. In some cases, like specific stages of embryogenesis, EMT is reversible and cells undergo the reciprocal mesenchymal-to-epithelial transition (MET). Activation of EMT and MET programs are involved during development for dispersing cells in embryos, during adulthood for tissue repair, but also for initiation of invasive and metastatic behaviour of epithelial cancers (Kalluri & Weinberg 2009).

b. The three subtypes of EMT

Three types of EMTs have been defined, according to the biological context in which they occur and their functional consequences (Figure 11). However, the specific signals initiating each type of EMT are not well defined.

i. Type 1: EMT during implantation, embryogenesis and organ development

A first type of EMTs is associated with implantation, embryo formation and organ development. These EMTs are components of the complex mechanisms allowing the development of a complete organism from a single cell, the fertilized egg. Indeed, during specific steps of embryogenesis, cells within certain epithelia are plastic and able to move back and forth between epithelial and mesenchymal states via EMT and MET, allowing completion of secondary epithelial tissues development (Lee et al. 2006). In this case, epithelial cells exert tissue-specific functions, whereas mesenchymal cells play a supporting

role. As an example, an EMT involving epithelial cells of the neuroectoderm gives rise to migratory neural crest cells (Sauka-Spengler & Bronner-Fraser 2008).

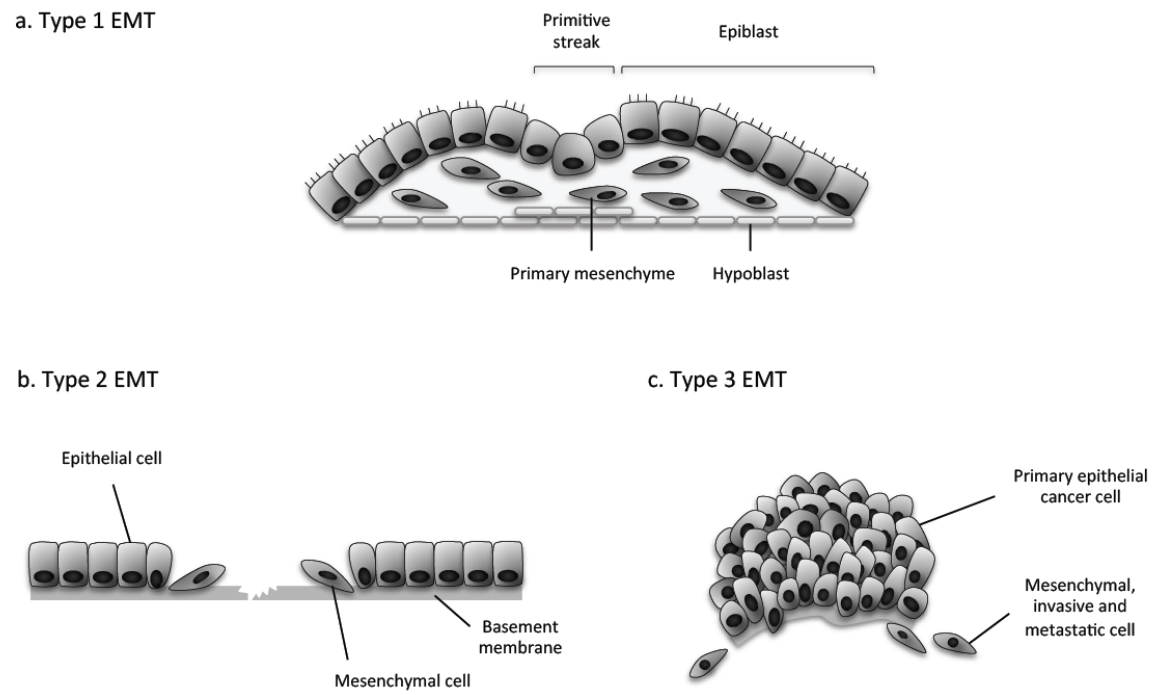

Figure 11 | The three types of EMT. **a.** Type 1 EMT is associated with implantation and embryonic gastrulation, and gives rise to the mesoderm and endoderm and to mobile neural crest cells. A primitive epithelium, the epiblast, gives rise to the primary mesenchyme via an EMT. Primary mesenchyme can form a secondary epithelia by a MET, and secondary epithelia can further differentiate through subsequent EMT to generate other types of epithelial tissues. **b.** EMTs are involved in tissue regeneration, in the context of inflammation and fibrosis. Type 2 EMT is expressed extended periods of time and can result in the destruction of an affected organ if the primary inflammatory insult is not removed or attenuated. **c.** Type 3 EMTs are involved in transformation of epithelial cancer cells into mesenchymal cells, enabling invasion and metastasis. Adapted from Kalluri & Weinberg, 2009.

ii. Type 2: EMT associated with tissue regeneration and organ fibrosis

Type 2 EMTs are involved in wound healing, tissue regeneration and organ fibrosis. In this case, EMTs are associated with repair events, that generates fibroblasts and other related cells to reconstruct tissues following trauma and inflammatory injury. Contrary to type 1 EMTs, type 2 EMTs are linked to inflammation, and stop once inflammation is attenuated. In organ fibrosis, persistent inflammation ensures a continuous EMT, resulting eventually in organ

destruction (K. K. Kim et al. 2006; M. Zeisberg et al. 2007; E. M. Zeisberg, Tarnavski, et al. 2007; E. M. Zeisberg, Potenta, et al. 2007).

iii. Type 3: EMT associated with cancer progression and metastasis

The significance of EMT in cancer progression has been controversial for a long period of time, mostly because EMT is difficult to follow in time and space in human tumors, and because the great diversity observed for cellular organization in cancer makes it impossible to recognize EMT without ambiguity (Tarin 2005). However, in an increasing number of studies, EMT is proposed to be the critical mechanism for the acquisition of malignant phenotypes by epithelial cancer cells (Thiery 2002). Cancer cells can pass through EMTs to differing extents, retaining epithelial traits while acquiring some mesenchymal ones, or becoming fully mesenchymal. These cells eventually enter into the invasion-metastasis cascade. After intravasation, transport through the circulation, extravasation, the cells form micro metastases, that can grow into macroscopic metastases. Interestingly, secondary colonies established by EMT-derived migratory cells resemble, at the histopathological level, the primary tumor from which they arose, without longer exhibiting mesenchymal traits. It has been proposed that an MET occurs during the course of secondary tumor formation (Figure 12) (Zeisberg et al. 2005; Thiery 2002). The fact that disseminated cancer cells undergo MET suggests that after extravasation into the parenchyma of a distant organ, they encounter a microenvironment without signals that were responsible for induction of the EMT in the primary tumor (Thiery 2002; Bissell et al. 2002).

iv. EMT markers

The signals initiating each type of EMT are not well defined. However, epithelial and mesenchymal states of cells that passed through an EMT are characterized by the expression of several specific markers (Figure 13) (Kalluri & Weinberg 2009).

c. Mechanism of EMT in cancer

The signalling components that contribute to EMT induction in primary tumors remain to be fully identified. One hypothesis is that the genetic and epigenetic alterations that occur in cancer cells render them particularly responsive to EMT-inducing heterotypic signals.

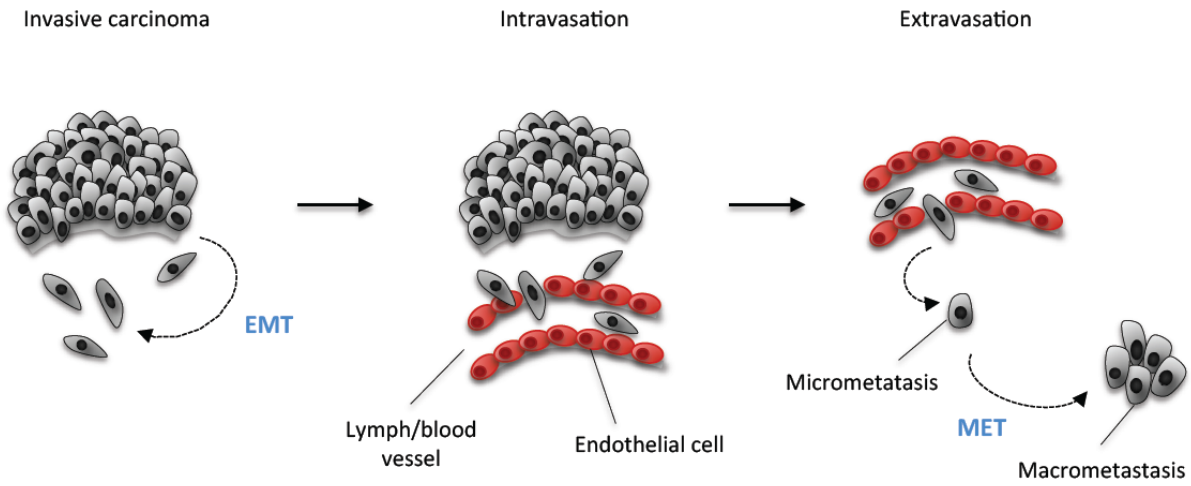

Figure 12 | EMT-MET interconversions in metastasis. Genomic alterations in a carcinoma can lead cells to acquire a migratory phenotype, possibly through epithelial-to-mesenchymal transition (EMT). Migratory cells can intravasate into blood or lymph vessels, allowing their passive transport to distant organs. At secondary sites, solitary cancer cells can extravasate and either remain solitary, forming micrometastasis, or form a new carcinoma, possibly through a mesenchymal-to-epithelial transition (MET).

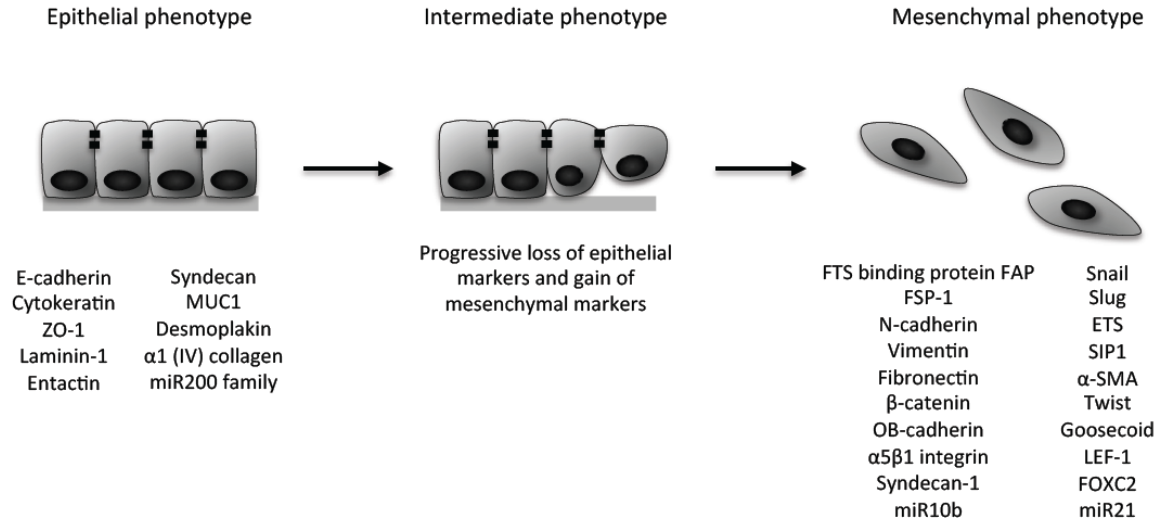

Figure 13 | Common cell markers of epithelial and mesenchymal cells. Colocalization of these two sets of distinct markers defines an intermediate phenotype of EMT, indicating that cells have passed only partly through an EMT. ZO-1, zona occludens 1; MUC1, mucin 1, cell surface associated; SIP1, survival of motor neuron protein interacting protein 1; FOXC2, forkhead box C2. Adapted from Kalluri & Weinberg, 2009.

These signals originate in the tumor-associated stroma (Weinberg 2008; Ansieau et al. 2008; Smit & Peeper 2008). Among them are found hepatocyte growth factor (HGF), epidermal growth factor (EGF), platelet-derived growth factor (PDGF) and transforming growth factor beta (TGF-β), altogether responsible for induction or functional activation of EMT-inducing transcription factors such as Snail, Slug, zinc finger E-box binding homeobox 1 (ZEB1) and 2 (ZEB2), Twist, Goosecoid and FOXC2 (Thiery 2002; Shi & Massagué 2003; Niessen et al. 2008; Medici et al. 2008; Kokudo et al. 2008). Intracellular signalling networks are implicated, among other signal-transducing proteins, such as extracellular signal-regulated kinases (ERK), mitogen-activated protein kinases (MAPK), Ras, lymphoid enhancer binding factor (LEF), β-catenin, as well as cell surface proteins (Kalluri & Weinberg 2009). EMT activation is also facilitated by the disruption of cell-cell adherens junctions and cell-ECM adhesions (Weinberg 2008; Gupta et al. 2005; J. Yang et al. 2006; Yang & Weinberg 2008) (Figure 14).

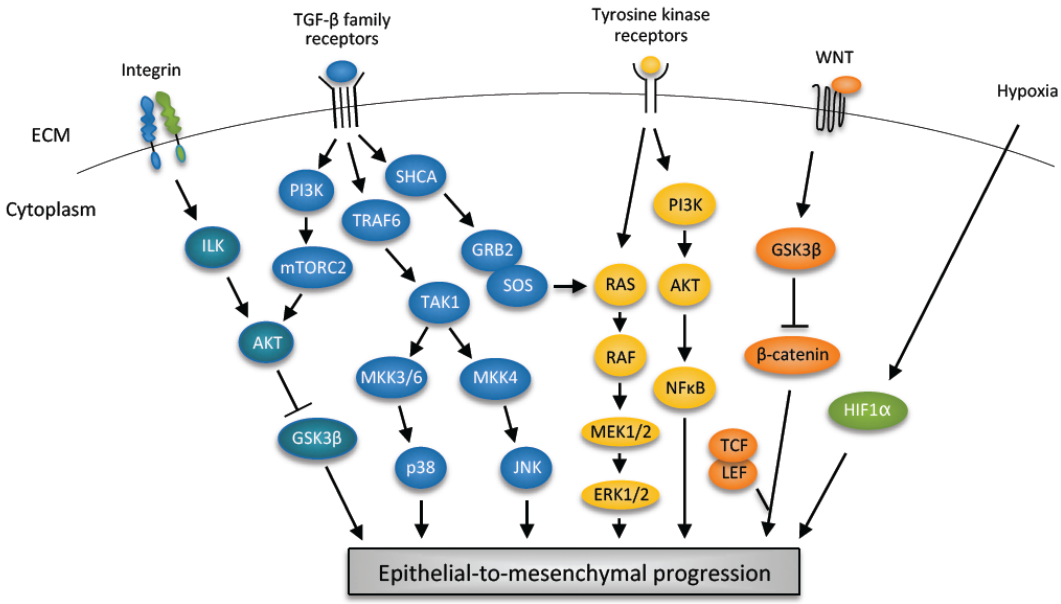

Figure 14 | Major pathways involved in EMT promotion. Transforming growth factor-β (TGFβ) can promote EMT through SMAD proteins, but also via activation of the PI3K–AKT, ERK MAPK, p38 MAPK and JUN N-terminal kinase (JNK) pathways. TβRI phosphorylates the adaptor protein SRC homology 2 domain-containing-transforming A (SHCA), creating a docking site for growth factor receptor-bound protein 2 (GRB2) and son of sevenless (SOS) and initiating the RAS–RAF–MEK–ERK MAPK pathway. Association of TNF receptor-associated factor 6 (TRAF6) with the TGFβ receptor complex activates TGFβ-activated kinase 1 (TAK1), and p38 MAPK and JNK. Several growth factors that act through receptor tyrosine kinases (RTKs), including epidermal growth factor (EGF), fibroblast growth factor (FGF), hepatocyte growth factor (HGF) and vascular endothelial growth factor (VEGF), can induce EMT, via activation of the RAS–RAF–MEK–ERK MAPK signalling cascade. Once activated ERK1 and

ERK2 MAPK can facilitate EMT by increasing the expression of EMT transcription factors and regulators of cell motility and invasion. Other signalling pathways are involved, such as WNT, promoting EMT by inhibiting glycogen synthase kinase-3β (GSK3β) to stabilize β-catenin. β-catenin translocates to the nucleus to engage the transcription factors lymphoid enhancer-binding factor 1 (LEF) and T cell factor (TCF) and promote a gene expression programme that favours EMT. Hypoxia in the tumour environment can promote EMT through hypoxia-inducible factor 1α (HIF1α), which activates the expression of TWIST. EMT responses can be increased through crosstalk and cooperation between distinct pathways. For example, RTK- or integrin-induced AKT activation can induce SNAIL expression through nuclear factor-κB (NF-κB) and stabilize SNAIL and β-catenin by inhibiting GSK3β, thus cooperating with WNT signalling. TGFβ signalling can also increase EMT responses initiated by growth factors such as FGF or EGF. ECM, extracellular matrix.

Among the factors involved in EMT activation and completion, two have been extensively studied. TGF-β is a positive regulator of tumor progression and metastasis, and has been shown to induce EMT in certain types of cancer cells (Bierie & Moses 2006; Oft et al. 1998; Song 2007). Two signalling pathways have been identified as mediating TGF-β-induced EMT: a first one involving Smad proteins (Miyazawa et al. 2002; Derynck et al. 2001; Heldin et al. 1997), LEF and β-catenin (L. Yang et al. 2006), and a second one implicating p38 MAPK, Ras homolog gene family member A (RhoA), and integrin β1-mediated signalling (Bhowmick, Zent, et al. 2001; Bhowmick, Ghiassi, et al. 2001). Another important feature in the induction of EMT is the expression of E-cadherin, for which a loss has been observed in a majority of human cancers (Tepass et al. 2000). Cytoplasmic sequestration of β-catenin is important for the preservation of epithelial features of cancer cells, and β-catenin movement to the nucleus correlates with loss of E-cadherin expression, as well as acquisition of a mesenchymal phenotype and invasive properties (Stockinger et al. 2001; Gottardi et al. 2001). Thus, a loss of cell surface E-cadherin induces an increase in cancer cells susceptibility to enter into an EMT, and an inverse relationship between levels of E-cadherin and patients survival has been found (Hirohashi 1998). The central role of E-cadherin loss in the EMT program is further demonstrated by the actions of EMT-inducing transcription factors such as Snail and Slug, which expression are induced by TGF- β, and that repress E-cadherin expression (Medici et al. 2008).

In addition, miRNAs are also components of the cellular signalling circuitry that regulates the EMT program, as exemplified by the role of miR200 and miR205, which inhibit the repressors of E-cadherin (ZEB1 and ZEB2) and help in the maintenance of the epithelial cell phenotype (Korpal et al. 2008; Park et al. 2008; Gregory et al. 2008). Interestingly, a lncRNA,

which expression is activated upon TGF-β treatment and therefore termed lncRNA-activated by TGF-β (lncRNA-ATB), is involved in the regulation of ZEB1 and ZEB2 by competitively binding miR-200 family members (Yuan et al. 2014). In several cancers, a loss of miR200 correlates with a decrease in E-cadherin levels (Kalluri & Weinberg 2009).

3. Long non-coding RNAs in cancer

Small non-coding RNAs have been extensively studied in the past decade as important players in a wide range of cellular processes and highly deregulated in the majority of human carcinomas (Oom et al. 2014). However, the emergence of lncRNAs as major players in human diseases is more recent. Numerous lncRNAs have been identified as deregulated or exerting a regulatory function in different types of disorders, such as autism, Alzheimer's and Parkinson's diseases, schizophrenia, diabetes and cardiovascular diseases (Sánchez & Huarte 2013). Aberrant lncRNAs expression has also been measured in a wide variety of cancers (Recapitulated in Table 1). LncRNAs expression can be highly specific to one cancer type, such as PCA3 in prostate cancer, or detected in almost all cancer types. In addition, many lncRNAs have been identified as important players in the acquisition of cancer hallmarks (Gutschner et al. 2012). This observation is not surprising, considering the fundamental role of lncRNAs in gene regulation, and the diversity of mechanism that have been described. Specific and aberrant expression of lncRNAs in diverse cancer types suggests that they could be established as biomarkers for cancer diagnosis and classification, but also as novel therapeutic targets.

a. Regulatory roles of lncRNAs in cancer

i. H19 and Xist, imprinted lncRNA genes

Imprinting, the process whereby the copy of a gene inherited from one patient is epigenetically silenced, is a key feature of cancer, as its loss results in altered gene expression. Interestingly, multiple maternally and paternally expressed genes with a high frequency of ncRNA genes are found in imprinted regions (Lim & Maher 2010). Two of the best-known imprinted genes are H19 and XIST, two lncRNAs.

Table 1 | Cancer-associated lncRNAs. These transcripts have been identified as aberrantly expressed, and/or playing a major function in tumor progression, in cell lines and cancer tissues.

LncRNA	Cytoband	Cancer Types	References
HOTAIR	12q13.13	Breast, brain, stomach, bladder, lung, colon, uterus, cervix, liver, ovary, oesophagus, skin (melanoma), pharynx, pancreas	Chakravadhanula, Ozols, Hampton, Zhou, & Bhardwaj, 2014; Ding et al., 2014; Endo et al., 2013; X.-S. Ge et al., 2013; Gupta et al., 2010; L. Huang et al., 2014; Kim et al., 2013; D. Li et al., 2013; Qiu et al., 2014; Sorensen et al., 2013; Svoboda et al., 2014; Tang, Zhang, Su, & Yu, 2013; W. Zhao, An, Liang, & Xie, 2014
MALAT1/NEAT2	11q13.1	Colon, cervix, stomach, pancreas, bladder, lung, prostate	Fan, Shen, & Tan, 2014; Gutschner et al., 2013; Jiang et al., 2014; J. Liu, Chen, Dang, Li, & Luo, 2014; J. Wang et al., 2014; Zheng et al., 2014
HULC	6p24.3	Liver, stomach, colon, liver	Gibb, Brown, & Lam, 2011; Xie, Ma, & Zhou, 2013; Y. Zhao et al., 2014
H19	11p15.5	Bladder, lung, liver, breast, endometrial, cervix, oesophagus, ovary, prostate colon	Fellig et al., 2005; Hibi et al., 1996; Tsang et al., 2010
ANRIL	9p21.3	Stomach, oesophagus, breast, prostate, blood (leukemias)	Chen et al., 2014; C. H. Li & Chen, 2013; Pasmant et al., 2007
MEG3	14q32.2	Stomach, lung, cervix, bladder, liver, pituitary gland	C Braconi et al., 2011; K. Lu et al., 2013; Yan et al., 2014; Zhou et al., 2007
GAS5	1q25.1	Liver, colon, stomach, kidney, breast, lung, bladder, pancreas, prostate	Krell et al., 2014; Z. Liu et al., 2013; X. Lu et al., 2013; Mourtada-Maarabouni, Pickard, Hedge, Farzaneh, & Williams, 2009; M R Pickard, Mourtada-Maarabouni, & Williams, 2013; Mark R Pickard & Williams, 2014; M. Sun et al., 2014; Tu, Li, Mei, & Li, 2014; Zhou et al., 2007
PRNCR1	8q24.2	Prostate, colon	Chung et al., 2011; L. Li et al., 2013
LOC285194	3q13.31	Bone, colon, oesophagus	Pasic et al., 2010; Qi et al., 2013
PCGEM1	2q32.2	Prostate	Petrovics et al., 2004
UCA1/CUDR	19p13.12	Bladder, tongue, breast, colon	Fang et al., 2014; J. Huang et al., 2014; X.-S. Wang et al., 2006
DD3/PCA3	9q21.22	Prostate	Kok et al., 2002
Anti-NOS2A	17q23.2	Brain	Korneev, Korneeva, Lagarkova, & Kiselev, 2008
TUC338	12q13.13	Liver	Chiara Braconi et al., 2011
PTENP1	9p13.3	Lung, skin (melanoma), uterus, blood (leukemias)	Ioffe, Chiappinelli, Mutch, Zighelboim, & Goodfellow, 2012; Marsit et al., 2005; Poliseno et al., 2011
SRA-1/SRA	5q31.3	Breast	Leygue, Dotzlaw, Watson, & Murphy, 1999
NcRAN	17q25.1	Colon, bladder, brain	Yu et al., 2009; Zhu et al., 2011
Xist	Xq13.2	Ovary, breast, kidney, colon	Lassmann et al., 2007; Weakley, Wang, Yao, & Chen, 2012
PCAT-1	8q24.21	Prostate, colon	X. Ge et al., 2013; Prensner et al., 2011
HIF1A-AS2	14q23.2	Breast, kidney	Cayre, Rossignol, Clottes, & Penault-Llorca, 2003; Thrash-Bingham & Tartof, 1999
GHET1	7q36.1	Stomach	Yang et al., 2014
CAI2	9p21	Brain	Barnhill et al., 2014
lncRNA-EBIC	16q	Cervix	N. Sun et al., 2014
Sox2ot	3q26.33	Lung	Hou et al., 2014
PCNA-AS1	20p12.3	Liver	Yuan et al., 2014

H19 lncRNA is expressed exclusively from the maternal allele, with its reciprocally imprinted protein-coding gene IGF2. Loss of imprinting at this locus results in high H19 expression in several cancers, including oesophagus, breast, colon, liver, bladder and hepatic metastases (Matouk et al. 2007; Fellig et al. 2005; Hibi et al. 1996). H19 has been described as having both oncogenic and tumor-suppressor roles. Indeed, it expression is activated by the oncogenic transcription factor c-Myc in colon cancer (Barsyte-Lovejoy et al. 2006), and down regulated by the tumor suppressor gene and transcriptional activator p53 (Dugimont et al. 1998; Farnebo et al. 2010). In addition, H19 processed product, miR-675, down regulates the tumor suppressor gene retinoblastoma (RB1) in colorectal cancer (Tsang et al. 2010). On the other hand, it has been shown in mouse that a lack of H19 induces larger and earlier tumor growth compared to wild-type situation (Leighton et al. 1995).

Xist, in mouse, is transcribed from a paternal allele and induces random X inactivation. Additional lncRNAs, such as Tsix, are involved in its regulation, and both Xist and Tsix can be processed in small RNAs (Ogawa et al. 2008). It is not clear whether this mechanism is conserved in humans, but Xist expression levels have been correlated with outcome in some cancers, such as the therapeutic response in ovarian cancers (Huang et al. 2002). Loss of Xist does not result in reactivation of the X chromosome. In breast cancer, two active X chromosome are frequently observed, mostly consequences of the loss of the inactive X chromosome, and duplication of the active X. Heterogeneous Xist expression levels are detected in this case (Sirchia et al. 2005; Sirchia et al. 2009; Vincent-Salomon et al. 2007).

ii. HOTAIR, MALAT-1 and HULC in cancer metastasis

Several lncRNAs have been implicated in metastasis, but the first and the most famous one is HOTAIR. This lncRNA is highly up regulated in a wide variety of cancers, including both primary and metastatic breast tumors with a 2000-fold increased transcription over normal breast tissues (Gupta et al. 2010). HOTAIR level was found to be correlated with metastasis and poor survival rates, establishing a link between lncRNA and patient prognosis. Numerous lncRNAs are transcribed from the HOX locus, suggesting that HOTAIR could be one example of a global regulatory phenomenon (Khalil et al. 2009). HOTAIR role as a molecular scaffold ensures the transcriptional silencing of a region of the HOXD locus, among many other genes, and remodels the gene expression pattern of breast epithelial cells to more

closely resemble that of embryonic fibroblasts, promoting cancer metastasis (Gupta et al. 2010).

Another well-studied lncRNA, MALAT-1, has been associated with high metastatic potential and poor patient prognosis during a comparative screen of non-small cell lung cancer patients with and without metastatic tumors (Ji et al. 2003). MALAT-1 is highly expressed in normal tissues, and up regulated in breast, prostate, colon, liver and uterus cancers (Guffanti et al. 2009; Gibb et al. 2011). MALAT-1 locus has further been identified to harbour chromosomal translocation breakpoints associated with cancer (Kuiper 2003; Davis et al. 2003). MALAT-1 silencing impairs the *in vitro* migration of lung adenocarcinoma cells (Tano et al. 2010), and reduces cell proliferation and invasive potential of a cervical cancer cell line (Guo et al. 2010), suggesting that MALAT-1 regulates the invasive potential of metastatic tumor cells. Furthermore, MALAT-1 has been shown to control the level of EMT-associated ZEB1, ZEB2 and SNAI1 transcription factors, implicated in the regulation of E-cadherin transcription (Ying et al. 2012).

HULC, highly up regulated in hepatocellular carcinomas, was also found as highly expressed in hepatic colorectal cancer metastasis. HULC has mainly been described as a "molecular decoy", sequestering miR-372 (Wang et al. 2010). Its over expression correlates with lymph node metastasis, distant metastasis and advanced tumor node metastasis stages, and gain-of-function studies showed that HULC promotes proliferation, invasion and inhibited cell apoptosis in human gastric cancer cell line. On the contrary, silencing of HULC effectively reversed the EMT phenotype (Zhao et al. 2014).

iii. MEG3 and linc-p21 in tumor suppression by p53 stimulation

The first lncRNA proposed to function as a tumor suppressor was the maternally expressed gene 3 (MEG3). If highly expressed in normal human tissues, especially in the brain and the pituitary gland (REF 156, 157), MEG3 is not detectable in a variety of human cancer cell lines, thus suggesting its role in the suppression of tumor development. Moreover, its ectopic expression suppresses the growth of cancer cells in culture, further supporting its role as a tumor-suppressor (Zhang et al. 2003). 12 MEG3 isoforms have been detected, all keeping the last exon that encodes an evolutionarily conserved miRNA, miR-770, with a high number of putative RNA targets (Hagan et al. 2009). However, MEG3 major role is its ability to

stimulate p53-dependent pathways (Zhou et al. 2007). Conversely, the lincRNA-p21 has been identified as a downstream repressor of p53 transcriptional response, suggesting that p53 transcriptional network includes numerous regulatory lncRNAs (Huarte et al. 2010).

iv. ANRIL, a regulator of tumor-suppressor genes

ANRIL is transcribed from and regulates *in cis* the expression of the INK4b-ARF-INK4a locus, a well-characterized tumor suppressor locus involved in cell cycle control, cell senescence, stem cells renewal and apoptosis (Kamijo et al. 1997; Kamb et al. 1994; Serrano et al. 1993). Aberrant expression and single nucleotide polymorphisms (SNPs) within ANRIL have been identified, and associated with susceptibility to a range of human diseases including cancer (Popov & Gil 2010; Cunnington et al. 2010).

v. T-UCR aberrant expression in carcinoma

The expression of many T-UCRs is significantly altered in cancer, notably adult chronic lymphocytic leukemias, colorectal and hepatocellular carcinomas, and neuroblastomas (Scaruffi et al. 2009; Calin et al. 2007). Their aberrant expression profiles can differentiate types of human cancers, and have been linked to patient outcome (Scaruffi et al. 2009). Some T-UCRs can be found in genomic regions specifically associated with cancer. One example is the T-UCR uc.73A, one of the most highly up regulated T-UCRs in colon cancer, which shows oncogenic properties by proliferation assays (Calin et al. 2007). Similarly, T-UCR uc.338 is up regulated in human hepatocarcinoma tumor and cell lines, and is a part of a larger transcription unit involved in cell growth (Braconi et al. 2011).

b. Diagnostic and therapeutic potential of lncRNAs

The discovery of lncRNAs as key regulators in cancer transformation and progression, expressed in a tissue- and cancer type-specific manner, leads to intriguing possibilities of application for diagnostics and therapeutics. As an example, analysis of HOTAIR expression in several types of cancer already showed the importance of this lncRNA as a marker of tumor progression and metastasis state. Indeed, originally described as involved in primary and metastasized breast tumors (Gupta et al. 2010), HOTAIR over expression is described as an excellent biomarker of tumor recurrence and poor prognosis in hepatocellular carcinoma,

cervical cancer, non-small lung cancer, among many others (Liu et al. 2013; Huang et al. 2014; Yang et al. 2011).

Compared to mRNAs, lncRNAs could be more useful diagnostic tools, as measurement of their expression represents directly the level of the active, mature product, whereas measured mRNA levels does not reflect the level of the functional product, the encoded protein (Cheetham et al. 2013). Another advantage is their stability, which allows non-invasive techniques to measure their level of expression, for example in human serum or urine (Tong & Lo 2006). This type of test is already performed to evaluate prostate cancer risk, by measuring the level of the lncRNA PCA3 in urine samples, avoiding unnecessary prostate biopsies (Sartori & Chan 2014). Given the evolution of transcriptomics technologies, it is now imaginable that analysis of tumor transcriptome will allow more accurate prognostic predictions, but also monitoring of tumor progression, recurrence and metastasis, by determining the expression of molecular markers such as HOTAIR. LncRNAs, typically expressed in a more cell-type specific manner than mRNAs, may also allow the estimation of the tumor cellular composition (Cabili et al. 2011; Chan et al. 2013). Targeted therapies are also now possible, as for H19-driven cancer types. Indeed, a plasmid carrying diphtheria toxin under the control of the H19 regulatory sequence has been developed to target cells over expressing H19. Intra-tumoral injection of the plasmid in patients with bladder, ovarian and pancreatic cancer successfully reduced tumor size (Cheetham et al. 2013). Reduction of MALAT-1 expression level by siRNAs can influence the migratory and proliferative potential of lung adenocarcinoma and cervical cancer cells in culture (Guo et al. 2010; Tano et al. 2010). Expression of molecules targeting these lncRNAs in a tumor-restricted manner will further allow precise targeting of cancer cells, without excessively damaging healthy tissues.

AIMS AND EXPERIMENTAL DESIGN

In the recent years, and in addition to small ncRNAs and housekeeping ncRNAs, numerous novel lncRNAs have been identified and annotated, arising from diverse regions of the human genome. With the evolution of high-throughput sequencing technologies, this lncRNAs catalogue will certainly continue to increase. The question that remains to be answered is whether all of these non-coding transcripts have a function within the cell. Several examples of lncRNAs, such as HOTAIR, MALAT-1, Xist and ANRIL, have already been characterized, showing the importance of long non-coding transcripts in major cellular processes. Indeed, lncRNAs can exert diverse regulatory functions: transcriptional interference, platforms for protein binding and targeting, regulation of RNA processing, stability and translation, or regulation of transcription and DNA recombination. In particular, lncRNAs have attracted increasing attention because of their tissue- and cell-type expression, but also because of their aberrant expression in a wide variety of human diseases, including cancers.

If the term cancer encompasses a high number of diseases, the mechanisms allowing a normal epithelial cell to form a tumor, and the acquired hallmarks that lead a tumor to become malignant, seems to be common to all tumor types. In particular, the acquisition of an invasive phenotype, allowing metastases formation at distant sites from the primary tumor, is the critical characteristic that transforms a benign tumor into a life-threatening disease. EMT is now considered as the cellular process inducing metastasis formation, through transformation of epithelial, non-motile cells, into mesenchymal, invasive cells. If proteins and small ncRNAs implicated in EMT signalling pathways have been analysed, the role of lncRNAs still remains unclear. Some lncRNAs such as MALAT-1, HULC and lncRNA-ATB, have been shown as involved in initiation and maintenance of the EMT, but their exact roles need further investigations. As EMT is a complex cellular process, it is likely that numerous long non-coding transcripts play an important role in its initiation, evolution and completion.

Given these observations, we addressed the following questions:

- Which annotated lncRNAs are specific to epithelial and mesenchymal states?
- Do novel lncRNAs specifically deregulated in EMT exist within the cell?
- Do these lncRNAs play a role in the induction or maintenance of the EMT?

To answer these questions, we used high-throughput RNA sequencing approaches to establish coding and non-coding transcriptome of cells from a specific *in vitro* model. In this model, established from primary Human Epithelial Kidney (HEK) cells, cells have been immortalized at epithelial and mesenchymal states, before and after a naturally occurring EMT. We established a transcriptome signature for both cell types, comprising annotated and novel long non-coding RNAs significantly deregulated between epithelial and mesenchymal cells. Among them, we selected one well-known lncRNA, HOTAIR, as a promising candidate for its significant up regulation after EMT, and for its already described involvement in the metastatic process. We examined whether HOTAIR could play a role in the EMT process, performing loss- and gain-of-function studies. We observed that HOTAIR doesn't seem to be involved in the initiation or maintenance of the EMT process *per se*, as its depletion or over expression does not affect mesenchymal or epithelial cell phenotypes. However, HOTAIR seems to play a major role in the regulation of cell proliferation, migratory and invasive capacities. These experiments and observations will be presented in a first part of the "results" section, and will be used for a publication *(In preparation)*.

A second part of the project started from the observation that few antisense lncRNAs have been annotated and characterized, despite of their major regulatory potential via pairing with a sense RNA target, mostly because at this time technologies allowing strand specific RNA-sequencing were not commonly used. In addition, previous work on yeast *Saccharomyces cerevisiae* performed in our laboratory described a novel class of lncRNAs, sensitive to degradation by XRN1 exoribonuclease in the cytoplasm, by a majority antisense to protein-coding genes, and involved in gene silencing. Interestingly, XRN1 is highly conserved through the Eukaryotic kingdom. Given these observations, we addressed two more questions:

- Is it possible to define a complete catalogue of antisense lncRNAs in human cells?
- Do XRN1-sensitive lncRNAs also exist in human cells?

To answer these questions, we first revisited publicly available RNA-seq datasets from the ENCODE project, using an original bioinformatics pipeline dedicated to novel lncRNAs identification, particularly antisense to protein-coding genes. We identified high numbers of previously unannotated antisense non-coding transcripts, among which we defined a novel

class of intronic antisense transcripts, termed INATs. We validated their existence within several human cell lines, and started their characterization. Secondly, and following the same strategy as previously developed in our laboratory with the yeast *Saccharomyces cerevisiae*, we analysed XRN1 expression in human cell lines, performed its depletion by siRNA in HeLa and MCF7 cells, and used high-throughput RNA sequencing approaches to identify XRN1-sensitive ncRNAs. Preliminary results of these studies will be presented in a second part of the "results" section.

RESULTS

A. LncRNAs in Epithelial-to-Mesenchymal Transition

1. Immortalized Human Epithelial kidney cells, a model to study lncRNAs in EMT

EMT can be induced in immortal epithelial cancer cell lines by several commonly used ways, such as induction of stress by UV or nutrient depletions, specific treatments like TGF-β, or ectopic expression of EMT inducers like Twist or SNAI1. However, these treatments induce changes in gene expression that are highly specific to treatment modalities, with relatively few common EMT signatures (Gröger et al. 2012). To understand whether long non-coding RNAs play a role in EMT, without studying indirect effects of such treatments, we have chosen to use a specific *in vitro* model originated from primary Human Epithelial Kidney (HEK) cells.

Initially developed in the laboratory of Arturo Londoño-Vallejo, this model has already been used to show that telomere-driven chromosome instability in human epithelial cells induces widespread changes in microRNA (miR) expression, ultimately leading to major perturbations in cell differentiation program. Notably, miR-200 family was down regulated, inducing directly the activation of the EMT program. The miR deregulation induced recapitulated the most common miR expression changes described in renal cancers and associated with tumor progression (Castro-Vega et al. 2013). Therefore, this system provides a perfect experimental setup to examine lncRNAs expression changes associated with EMT, and to understand whether lncRNAs could be directly involved in cell differentiation.

a. HEK-Epi and HEK-Mes cells are immortalized from HEK Primary cells

Stress-induced premature senescence, related with culture conditions and expression of cyclin-dependant kinase inhibitors p16 and p21, occurs in cultured epithelial cells at early passages. To extend the replicative capacity of cells beyond the entry into senescence, primary HEK cultures have been transfected with the early region of SV40 (ER-SV40). ER-SV40, which drives the expression of large T and small T antigens, has a main role in inactivation of the retinoblastoma (pRB) and p53 pathways.

After introduction of ER-SV40, cells can grow for an additional 60-70 PDs before hitting crisis. Telomeres progressively shorten during this time, leading to telomere instability and chromosome end-to-end fusions that provoke repeated cycles of BFBs in subsequent passages. Previous studies showed that the first telomeric fusions take place around PD 50, invariably (der-Sarkissian et al. 2004). 20 to 30 PDs following the initiation of genome instability, the cells enter into the crisis period, where they can stay for a variable time before dying. To keep the cells in culture even after the beginning of chromosome instability, cells have been infected with retrovirus-based constructs driving the expression of hTERT, the catalytic subunit of telomerase. It has been previously shown that immortalization of human cells at early passages prevents genome instability and maintains the original phenotype with no sign of cell transformation (Jiang et al. 1999; Morales et al. 1999). Following this idea, cells have been transduced at early passages (30 PDs), to be fixed as normal epithelial cells (HA5+hTERT-Early, in this study: HEK-Epi). Cells have also been transduced with the same vector after the initiation of chromosome instability but well before their entry into crisis (60 PDs) (HA5+hTERT-Late, in this study: HEK-Mes) (Figure 15). Maintained in culture for more than 200 PDs, these cell lines showed no changes in their growth pattern (Castro-Vega, Thesis, 2010).

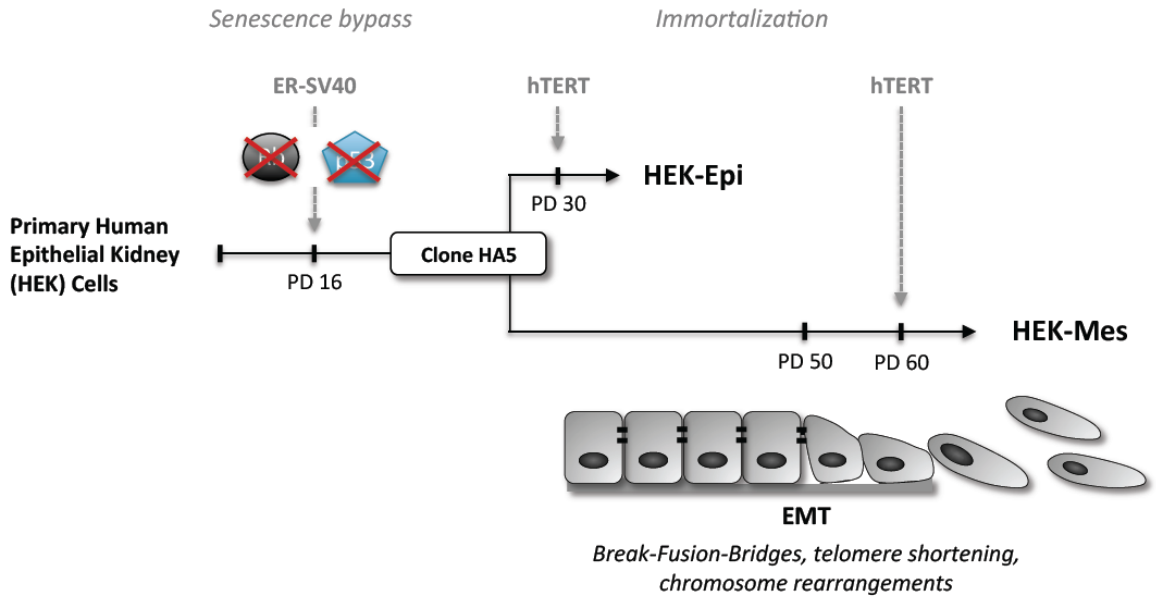

Figure 15 | HEK-Epi and HEK-Mes cells immortalization from HEK primary cells. Primary HEK cells were transfected with ER-SV40 to bypass stress-induced premature senescence, and immortalized by introduction of hTERT at 30 and 60 PDs, to generate HEK-Epi and HEK-Mes cells, respectively. HEK-Epi cells are karyotypically stable, without chromosome instability, whereas HEK-Mes cells underwent telomere shortening, initiated naturally at PD50, and chromosome rearrangements, resulting in karyotypically abnormal cells.

b. Immortalized HEK-Mes cells exhibit an EMT-like phenotype

Previous studies on these two cell lines showed a significant down regulation of the miR-200 family members (Castro-Vega et al. 2013), which have been directly implicated in the induction of EMT (Gregory et al. 2008; Park et al. 2008). These observations led us to analyse phenotypic features of HEK-Epi and -Mes cells.

Under microscopic evaluation, HEK-Epi cells exhibit a round, cobblestone morphology, typical of epithelial cells (Figure R16a), whereas HEK-Mes cells display an elongated, spindle-like shape typical of mesenchymal cells such as fibroblasts (Figure R16b).

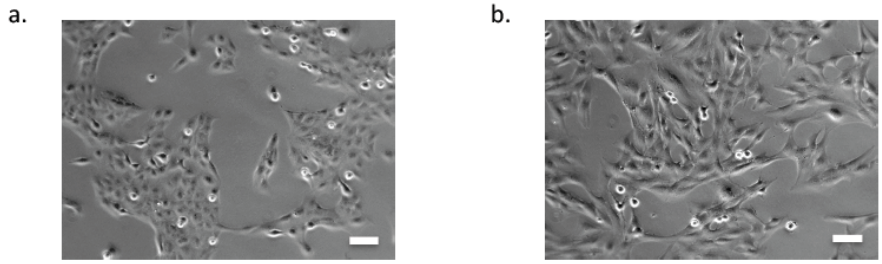

Figure 16 | HEK-Epi and HEK-Mes cells display epithelial and mesechymal morphologies. **a.** HEK-Epi cells show epithelial, cobblestone morphology. **b.** HEK-Mes cells show mesenchymal, spindle morphology. **a**, **b**: magnification 10x, scale bar = 100 µm.

As the migratory phenotype is one of the most striking traits acquired during the EMT process, we attempted to verify whether HEK-Mes cells display such a phenotype. We performed a wound-healing assay (WHA), where an artificial wound was created on confluent cell monolayers by scratching the bottom of the dish with a P200 pipette tip. The rate of cell migration closing the wound was estimated on high power fields (HPF) taken after 12 and 24 hours. Quantification of the wound size was performed using TScratch software (Gebäck et al. 2009). As expected, WHA showed a complete wound recovery for HEK-Mes cells, whereas HEK-Epi cells exhibited a poor migratory capacity, with cells growing on multiple layers instead of invading completely the wound (Figure 17a and b).

We performed a second test, more stringent, to assess the capacity of cells to invade through a barrier of extracellular matrix proteins (Matrigel). HEK-Mes cells showed an

increased invasiveness, with a mean of 100 invading cells per HPF, compared to 25 for HEK-Epi cells (Figure 17c and d).

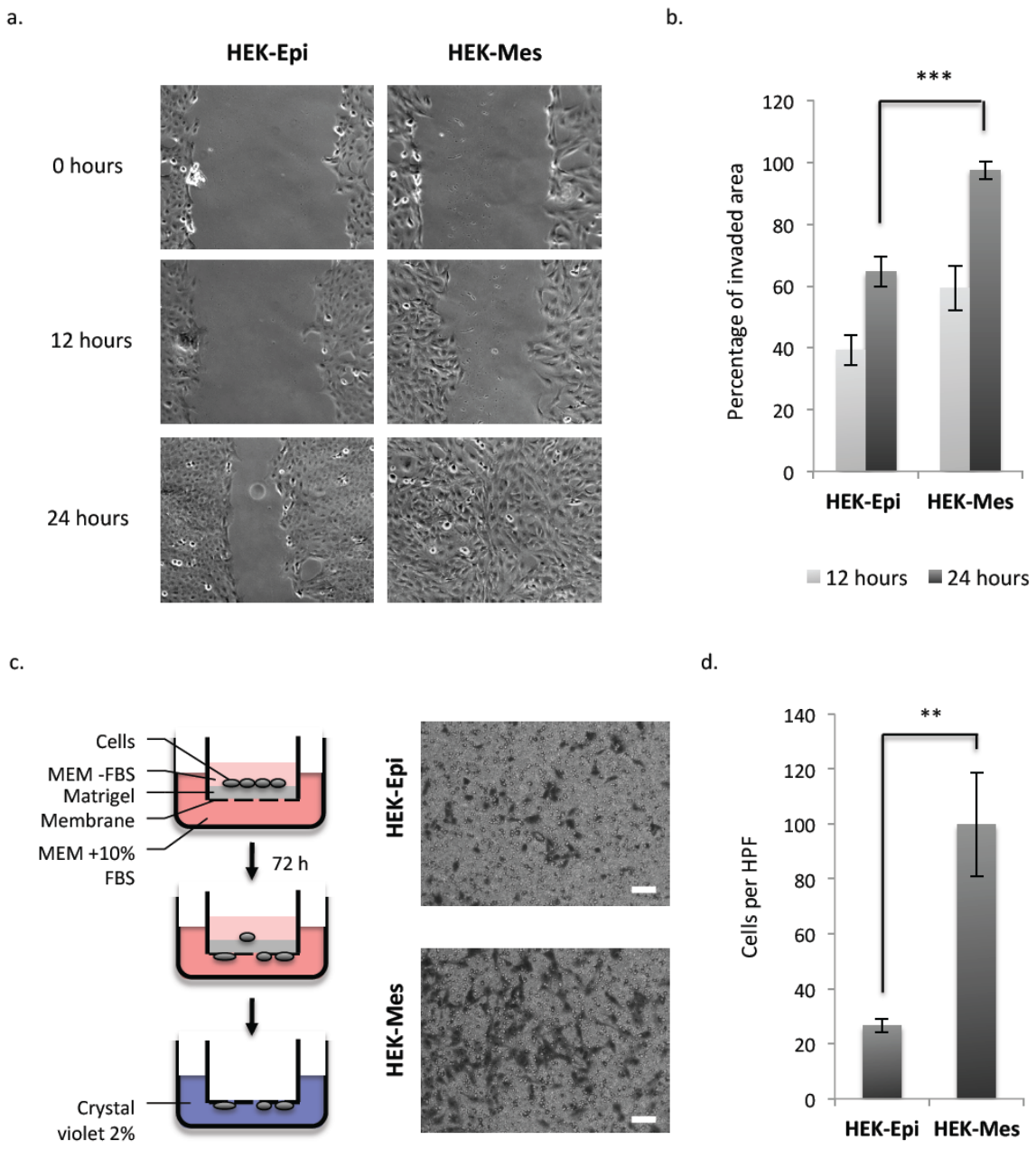

Figure 17 | HEK-Epi and HEK-Mes cells migratory and invasive properties. **a.** Wound healing assay was used to assess cells motility. Phase contrast images show wound recovery at 0, 12 and 24h post-scratch, in 10x magnification. **b.** Wound healing assay quantification. Histograms represent the estimated percentage of the invaded area, taking first picture at 0h as 20% invasion reference, and in a mean of 6 contrast phase images. **c.** Matrigel invasion assay was performed to assess cells invasion capacity. A mean of 10 phase contrast images (HPF) was taken 72h after seeding 200.000 cells on membranes coated with matrigel, in three independent experiments. Scale bar: 100 µm. **d.** Counted number of invading cells per HPF. Error bars indicate standard deviation; Student t-test was used to determine the statistical significance: ** $p<0.01$, *** $p<0.0001$.

We then examined by random-primed RT-qPCR the mRNA abundance of widely used EMT markers, recapitulated in Table 2 (All primer sequences are reported in Supplementary Table S1). To normalize RT-qPCR results, we first selected commonly used housekeeping genes such as glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and beta-2-microglobulin (B2M), but we observed variations in their expression between HEK-Epi and HEK-Mes cells. Therefore, we searched for genes exhibiting stable expression in both cell lines, in microarray datasets previously generated by Arturo Londoño-Vallejo's laboratory. We selected ribosomal protein L11 (RPL11) and RNA polymerase II polypeptide A (POLR2A) for their stable expression in HEK-Epi and HEK-Mes cells, and used them both to normalize all RT-qPCR experiments. Presented results were normalised by RPL11, but normalisation by POLR2A showed equivalent changes in mRNA levels.

Table 2 | Most relevant epithelial (in blue) and mesenchymal (in red) EMT markers.

Cell surface proteins	Cytoskeletal proteins	ECM proteins	Transcription factors
E-cadherin (CDH1)	β-catenin (CTNNB1)	Fibronectin 1 (FN1)	Snail Family Zinc Finger 1 (SNAI1)
Occludin (OCLN)	Smooth muscle actin (ACTA2)	Laminin Gamma 2 (LAMC2)	Zinc Finger E-Box binding Homeobox 2 (ZEB2)
Tight junction protein 3 (TJP3)	Vimentin (VIM)		
Claudin 2 (CLDN2)	Keratin 19 (KRT19)		

Tested epithelial markers, LAMC2, OCLN, TJP3, KRT19 and CTNNB1, showed a significant decrease in HEK-Mes cells. Mesenchymal markers ZEB2, SNAI1, ACTA2, VIM and FN1 mRNA levels were increased in HEK-Mes compared to HEK-Epi cells (Figure 18a). We performed then a western blot analysis of the epithelial marker β-catenin (CTNNB1), and mesenchymal markers vimentin (VIM) and smooth muscle actin (ACTA2). We observed a strong decrease of the CTNNB1 protein level, and an increased steady state level of VIM and ACTA2 proteins in HEK-Mes cells (Figure 18b). Differences of fold changes in HEK-Mes compared to HEK-Epi cells, between mRNA and protein levels suggests that regulation of some EMT markers between epithelial and mesenchymal states could happen at translational or post-translational levels. Even thought, differential expressions of EMT markers are

consistent with the fact that HEK-Epi and HEK-Mes cells exhibit epithelial and mesenchymal expression programs, respectively.

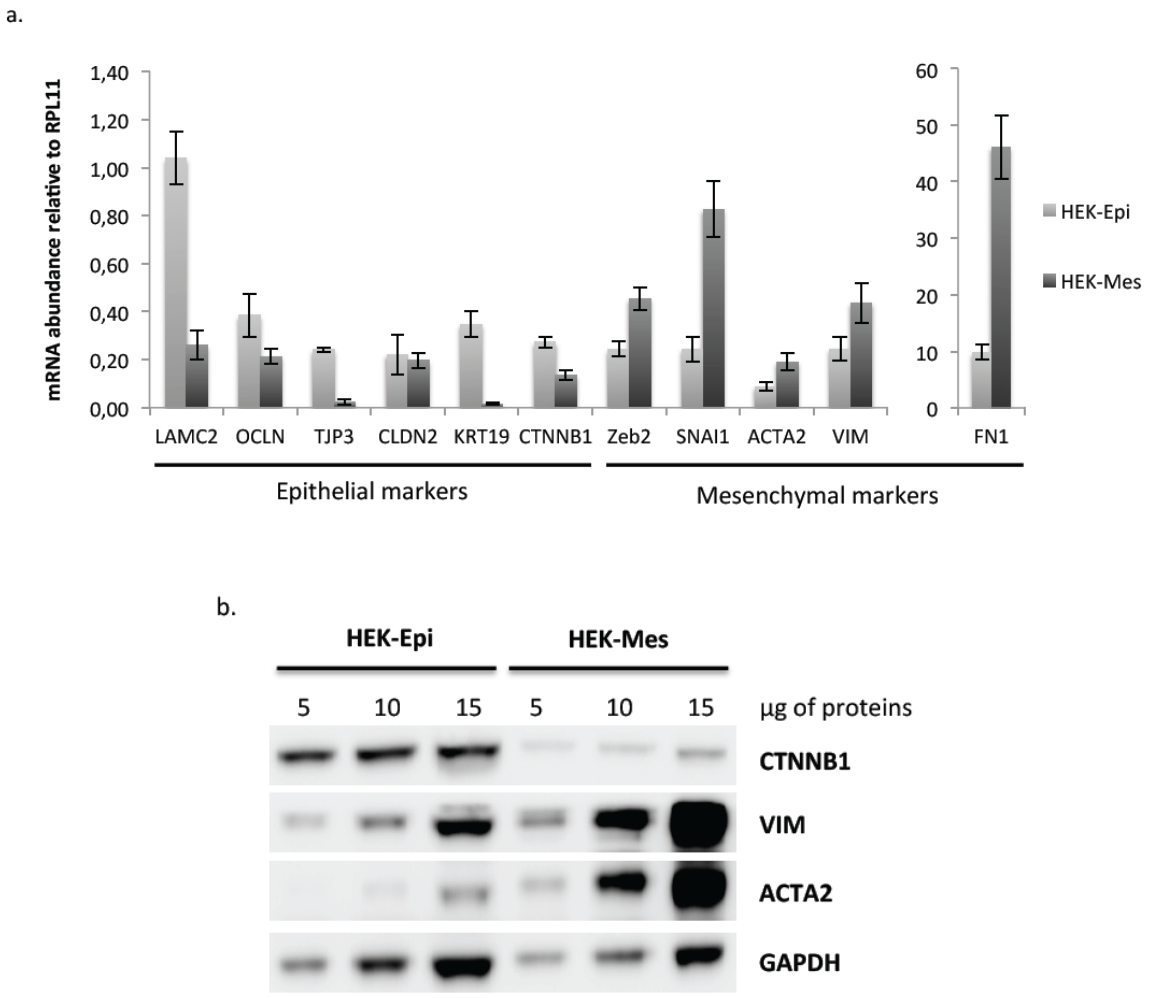

Figure 18 | HEK-Epi and HEK-Mes cells display EMT signatures. **a.** mRNAs abundance of EMT marker genes, quantified by random-primed RT-qPCR and relative to RPL11 level. Mean of three independent experiments. Error bars indicate standard deviation; Student t-test was used to determine the statistical significance: ** $p<0.01$, *** $p<0.0001$ **b.** Western blot detection of EMT markers. GAPDH is used as a loading control.

Taken together, these results suggested that HEK-Mes cells display an EMT-like phenotype, defined by the acquisition of migratory and invasive capacities, as well as a strong increase of mesenchymal genes expression and protein levels. This result confirmed that EMT occurred after beginning of telomere instability, and before crisis, when cells started to accumulate chromosome rearrangements. Interestingly, HEK-Mes cells exhibit some aspects

of human epithelial carcinogenesis: genome instability, immortalization, and acquisition of migratory and invasive properties. These first studies confirmed our interest in using this model to analyse differential expression and potential role of lncRNAs in EMT.

2. EMT is associated with changes in lncRNAs expression

a. LncRNAs signature of EMT in immortalized HEK-Epi and HEK-Mes cells

i. High-throughput RNA-sequencing allows the establishement of HEK-Epi and HEK-Mes full transcriptome

In order to establish a complete catalogue of differentially expressed lncRNAs in HEK cells before and after EMT, we used a high-throughput RNA sequencing approach, starting from total RNAs of HEK-Epi and HEK-Mes cells (Figure 19). Three RNA samples were extracted from each cell line using a Trizol reagent (Life Technologies). After ribosomal RNA depletion, we prepared strand-specific cDNA libraries using SOLiD Total RNA-seq kit (Life Technologies). Previous RNA-seq experiments performed in our laboratory showed that Illumina sequencing allowed easier and faster data treatment, compared to SOLiD sequencing. Therefore, we adapted our libraries adding Illumina adapters using PCR to each library fragments. Paired-end sequencing of cDNA libraries was performed on Illumina HiSeq 5500, by the IMAGIF genomic platform.

Data analysis was performed by Zohra Saci, using an original computational pipeline (Figure 19). Briefly, the quality of the reads was tested, and reads were trimmed to exclude SOLiD adapters from each sequence. The reads were then mapped using TopHat (CCB) and GENCODE v14 as a human genome of reference (Table 3). We observed that the average percentage of uniquely mapped reads was very low for HEK-Mes cells (42%) compared to HEK-Epi cells (63%), especially considering that the total number of reads sequenced was equivalent for each replicate of both cell lines. This phenomenon could be explained by chromosome rearrangements that occurred in HEK-Mes cells before immortalization, creating a very heterogeneous genome, distant from the one used as a reference. Then, incorrect read pairs were filtered. Only uniquely mapped reads forming correct pairs were used for next steps.

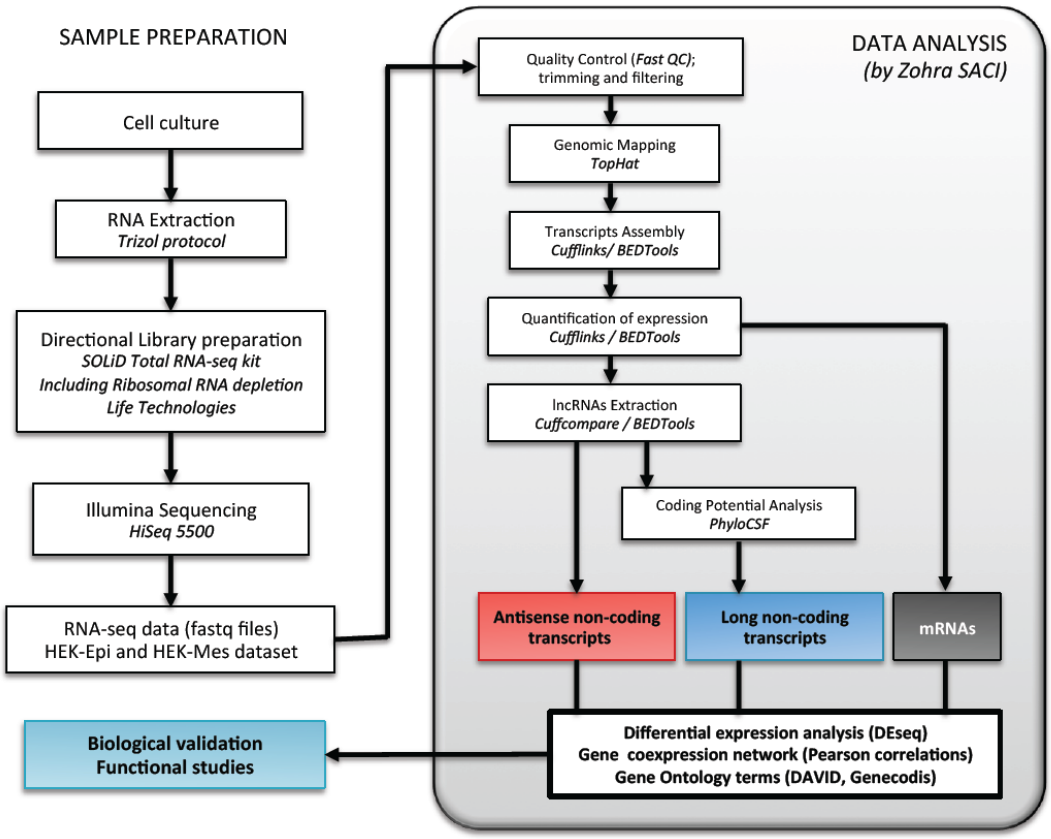

Figure 19 | Biological and bioinformatics pipeline used for the establishment of HEK-Epi and -Mes coding and non-coding transcriptome by RNA-seq. Total RNAs were extracted from cultured HEK-Epi and HEK-Mes cells, and RNA-seq libraries were prepared using SOLiD Total RNA-seq kit from Life Technologies. Samples were adapted for Illumina sequencing, and sequenced on HiSeq 5500 Illumina sequencer. Quality of the reads sequenced was then assessed, before mapping and assembly of transcripts. Expression of each transcript was quantified, antisense ncRNAs were extracted and additionally tested for their coding potential. Differential expression analysis were performed using DEseq, followed by Gene expression network establishment and gene ontology terms analysis.

Table 3 | RNA-sequencing statistics for HEK-Epi and HEK-Mes libraries.

Samples	Total number of reads	Uniquely mapped reads	% of mapped reads
HEK-Epi_1	225 716 306	145 907 150	64
HEK-Epi_2	210 977 885	151 442 582	71
HEK-Epi_3	214 096 046	120 905 169	56
HEK-Mes_1	222 885 773	88 388 676	39
HEK-Mes_2	225 144 444	97 496 641	43
HEK-Mes_3	219 105 561	98 084 365	44

Transcripts assembly and quantification were performed using Cufflinks (CBCB) and BEDTools (Broad Institute). Only assembled transcripts with length ≥ 200 nucleotides, present in at least one copy per cell (Reads per kilobase per million reads (RPKM) ≥ 1) (Mortazavi et al. 2008) and expressed in at least two biological replicates were considered for further analysis. These assembled transcripts were divided into three distinct classes that were then treated separately. The transcript was considered as mRNA when the reads were mapped on annotated exons, lncRNA when the assembled transcript was intergenic or overlapping annotated genes on the same strand, and antisense-lncRNA (as-lncRNA) when the assembled transcript overlapped annotated genes but on the antisense strand (Figure 20a). LncRNAs were specifically tested for their coding potential using PhyloCSF (Lin et al. 2011). Number of transcripts in each class is presented in Table 4. Remarkably, we identified numerous transcripts that do not overlap with any previously annotated transcription unit, considered as novel. This result was even more striking in the case of as-lncRNAs, for which the majority are novel transcripts.

Table 4 | mRNA, lncRNA and as-lncRNA repository in HEK-Epi and HEK-Mes cells.

Cell line	mRNA	lncRNA		as-lncRNA	
		annotated	novel*	annotated	novel*
HEK-Epi	11 954	1 476	44	383	1 099
HEK-Mes	12 603	1 779	58	478	2 557
Total	13 062	1 518	82	597	2 980

*only transcripts that do not overlap with any annotated transcription unit were considered

ii. Analysis of HEK-Epi and HEK-Mes cells full transcriptome reveals changes of lncRNAs and as-lncRNAs expression after EMT

Analysis of transcriptome density and distribution between HEK-Epi and HEK-Mes cells revealed notable differences between three distinct classes of transcripts. In total, 13 062 mRNAs were identified as expressed in HEK cells, but only 3% were expressed exclusively in HEK-Epi and 8% in HEK-Mes. Density plot of HEK-Mes versus HEK-Epi cells showed a rather homogeneous mRNAs expression between the two cell lines (Pearson correlation coefficient, R=0.83). Non-coding RNAs expression was much more deregulated between epithelial and mesenchymal states.

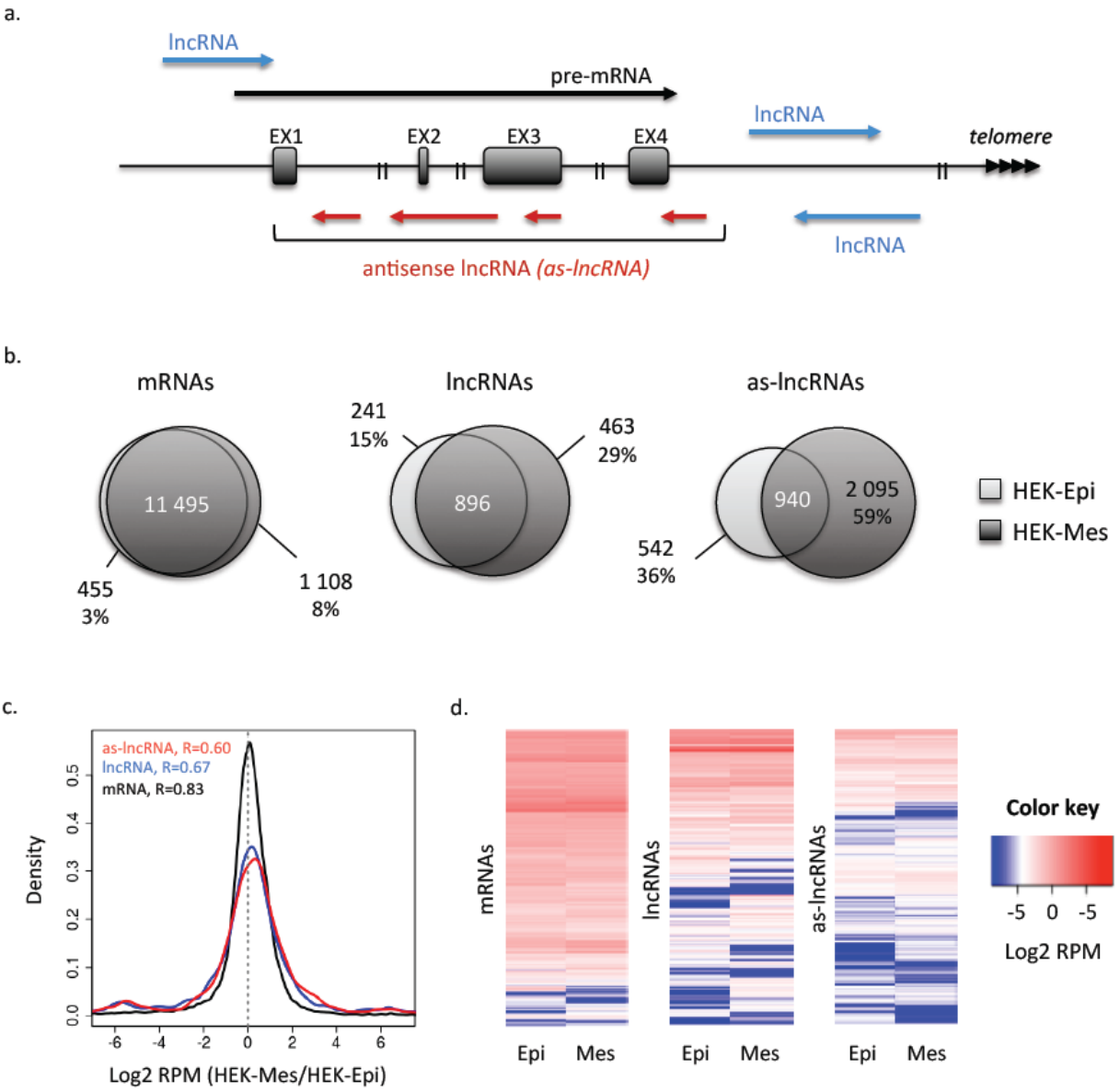

Figure 20 | Protein-coding and non-coding full transcriptome in HEK-Epi and HEK-Mes cells. **a.** Definition of transcript types. Transcripts overlapping coding genes, or located in intergenic regions, are described as long non-coding RNAs (lncRNAs). Transcripts detected in antisense orientation of annotated genes are described as antisense long non-coding RNAs (as-lncRNAs). **b.** Venn diagram of full transcriptome distribution in HEK-Epi and HEK-Mes cells. **c.** Transcript density distribution in HEK-Mes versus HEK-Epi cells. d. Heatmaps of HEK-Epi and HEK-Mes full transcriptome. RPM: Reads Per Million.

Indeed, percentages of lncRNAs and as-lncRNAs exclusively expressed in HEK-Epi or HEK-Mes were strongly increased compared to mRNAs. 1 600 lncRNAs were identified as expressed in HEK cells, with 29% specifically expressed in HEK-Mes cells, and 15% in

HEK-Epi cells. As-lncRNAs showed an even more cell-specific expression. Among 3 577 transcripts, 36 and 59% were expressed only in HEK-Epi and HEK-Mes cells, respectively. In correlation with these observations, density plots showed a slight shift in lncRNAs and as-lncRNAs expression towards HEK-Mes cells, with the Pearson correlation coefficient R=0.67 and 0.60, respectively (Figure 20b and c). Heat maps established for all three classes of transcripts showed that lncRNAs and as-lncRNAs are globally expressed at low levels, and revealed clusters of transcripts differentially expressed between HEK-Epi and HEK-Mes cells (Figure 20d).

iii. LncRNAs and as-lncRNAs are differentially expressed before and after EMT

- Differential expression analysis of mRNAs, lncRNAs and as-lncRNAs in HEK-Epi and HEK-Mes cells

To identify transcripts specific to epithelial or mesenchymal HEK cells, we performed a differential expression analysis (DEseq - Bioconductor) starting of the HEK-Epi and HEK-Mes RNA-seq dataset. We established two catalogues of transcripts. First, we defined an extended catalogue of transcripts using the combined criteria of p-value ≤ 0.05 and fold change ≥ 2. We then restricted the catalogue to a core list of the most significant differentially expressed transcripts, using the criteria of adjusted p-value ≤ 0.05 and fold change ≥ 2. Number of differentially expressed transcripts present in both extended and core catalogues can be found in Table 5. We established heat maps of three classes of deregulated transcripts, showing global lower expression of lncRNAs and as-lncRNAs compared to mRNAs. In addition, we identified clusters of transcripts specifically expressed in HEK-Epi or HEK-Mes cells (Figure 21a and b).

Table 5 | Numbers of differentially expressed mRNAs and lncRNAs

Transcript	Down-regulated, Mes Number of transcripts		Up-regulated, Mes Number of transcripts	
	p≤0.05	adj_p≤0.05	p≤0.05	adj_p≤0.05
mRNA	1 424	745	1 867	908
lncRNA	124	53	146	45
as-lncRNA	249	93	352	120

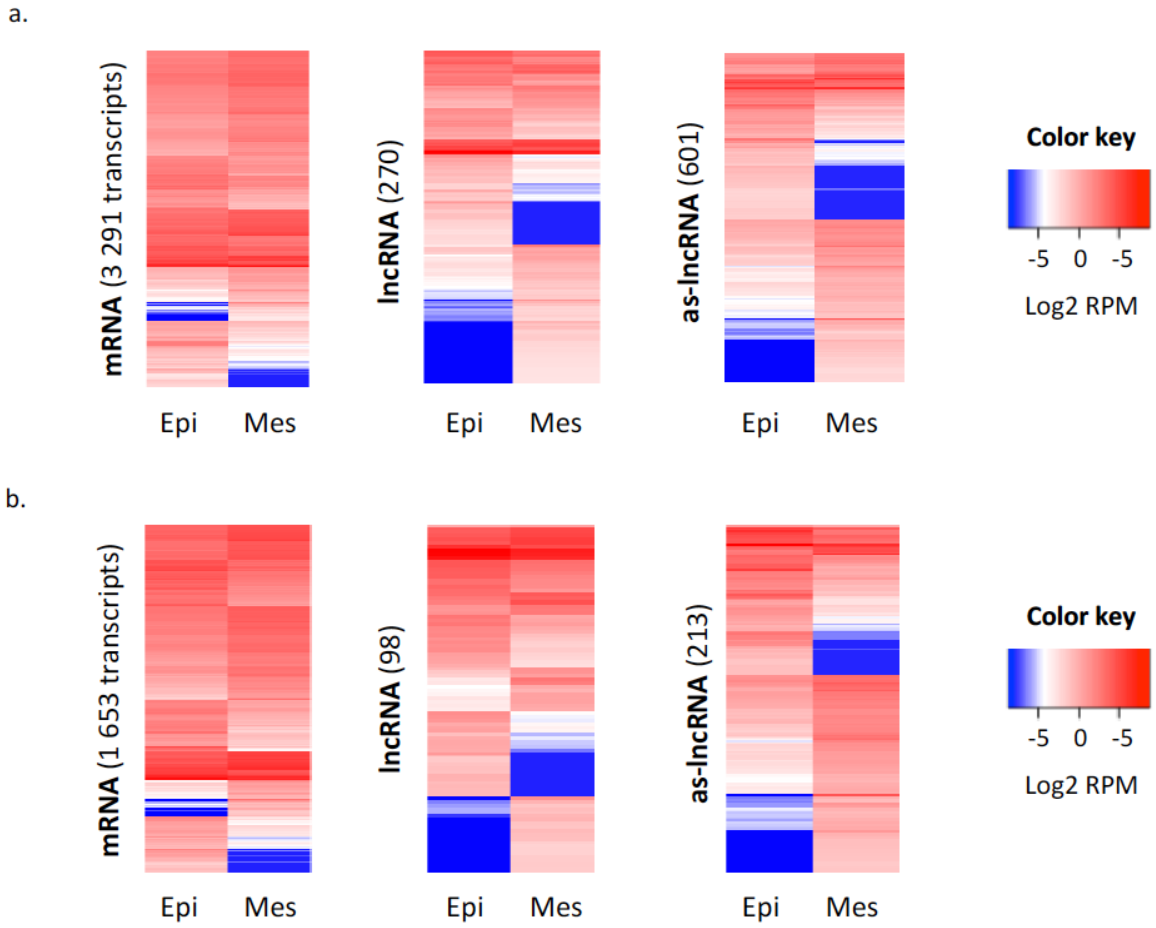

Figure 21 | Differential expression of coding and non-coding RNAs in HEK-Epi and HEK-Mes cells. **a.** Heat maps of extended list of deregulated mRNAs, lncRNAs and as-lncRNAs, between HEK-Epi and HEK-Mes, identified by DEseq analysis (FC>2, p-value<0.05). **b.** Heat maps of core list of deregulated mRNAs, lncRNAs and as-lncRNAs, between HEK-Epi and HEK-Mes, identified by DEseq analysis (FC>2, adjusted p-value<0.05).

- Biological processes and pathways associated with differentially expressed genes in HEK-Epi and HEK-Mes cells

To understand the biology underlying differentially expressed transcripts, we sought for associated biological processes. We used the Database for Annotation, Visualization and Integrated Discovery (DAVID) bioinformatics tool (Gene & Consortium 2000; Harris et al. 2004) to categorize genes from our EMT extended and core lists of mRNAs, according to their biological process annotation. We selected the 10 biological processes with the highest number of genes and p-value ≤ 0.05 (Figure 22a). With both extended and core lists of

differentially expressed mRNAs between HEK-Epi and HEK-Mes cells, enriched processes are linked to regulation of RNA and transcription, regulation of apoptosis, but also to cell adhesion, motion and proliferation. In addition with this analysis, we used the Kyoto Encyclopedia of Genes and Genomes (KEGG) to identify the significantly (p-value ≤ 0.05) enriched pathways of the same genes (Figure 22b). We found as the most relevant pathways those related with cancer, regulation of cellular communication (focal adhesion, extra-cellular matrix receptor interaction, cell adhesion molecules), but also TGF-β signalling. Thus, mRNAs identified as differentially expressed between HEK-Epi and HEK-Mes cells seemed to be involved in biological processes and pathways that could be easily linked to the observed EMT-like phenotype in HEK-Mes cells.

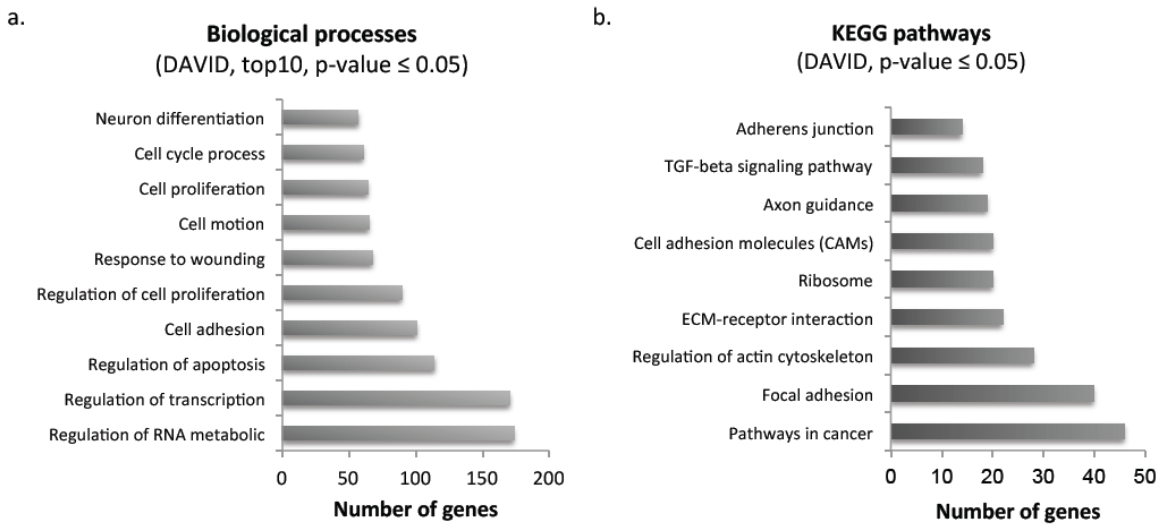

Figure 22 | Biological processes associated with differentially expressed mRNAs between HEK-Epi and HEK-Mes cells. **a.** Most enriched GO terms of biological processes identified by DAVID (p-value ≤ 0.05). **b.** KEGG pathways identified by DAVID analysis (p-value ≤ 0.05).

- Comparison with published EMT core gene dataset

We aimed then to examine whether our list of differentially expressed transcripts reflected classical transcription profiles of EMT. We compared our results with published EMT expression signatures identified in the previous comparative study joining 18 independent gene expression studies focusing on different cell types and treatment modalities. They established a list of 365 genes, up or down regulated upon EMT and shared between at least 10 datasets (Gröger et al. 2012). By comparison of this Gröger list with our own extended list

of mRNAs differentially expressed between HEK-Epi and HEK-Mes cells, we found several common genes: 62 among 744 down regulated genes, and 47 among 808 genes up regulated in HEK-Mes compared to HEK-Epi cells (Figure 23, Supplementary Table S2). Among common genes, we found ACTA2 and FN1, used in previous experiments as EMT markers, and FERMT1, a gene for which we identified a new as-lncRNA.

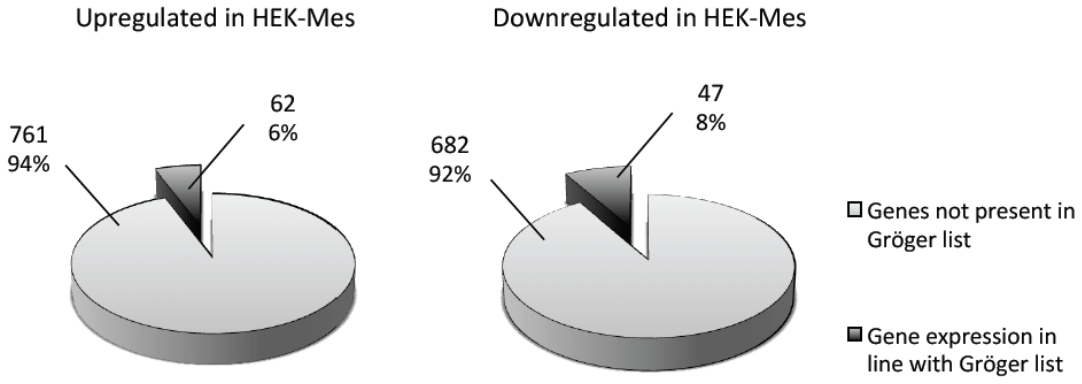

Figure 23 | Comparison with published list of EMT markers from Gröger *et al.* 2012. Lists of differentially expressed genes between HEK-Epi and HEK-Mes cells were compared with EMT list of 365 genes established by Gröger *et al*. from 18 gene expression studies dataset.

One hypothesis to explain the low overlap between Gröger list of EMT genes and our list of differentially expressed transcripts could be explained by the condition used for EMT initiation. Indeed, Gröger *et al*. analysed studies in which EMT has been artificially induced in epithelial cancer cells by specific stress conditions or treatments such as TGF-β1 or growth factors. They already observed a low overlap between cell lines and EMT induction modalities. In the HEK system, EMT is naturally occurring and potentially linked to the accumulation of genomic alterations. Therefore, changes in transcriptome associated with EMT might be different.

- Biological validation of RNA-sequencing analysis

To validate RNA-seq analysis, we selected several differentially expressed transcripts for all three classes of RNAs, showing high level of expression, significant fold-change between HEK-Epi and HEK-Mes cells and biological relevance. We measured their abundance by

random-primed (mRNAs) or oligo-specific (lncRNAs and as-lncRNAs) RT-qPCR in HEK-Epi and HEK-Mes cells.

Table 6 | Selected deregulated transcripts between HEK-Epi and HEK-Mes cells.

Transcript	HEK-Epi	HEK-Mes	Fold Change Mes/Epi	adj. p-value
mRNAs				
HAVCR1	5026	11	**0,00**	0,0000
MYO1D	2542	9	**0,01**	0,0000
IL7R	488	7190	**21,68**	0,0000
TFAP2A	6	1500	**343,71**	0,0000
FAT3	5	1281	**397,26**	0,0000
lncRNAs				
HOTAIR	20	251	**17,09**	0,0000
CTD-2314B22.3	25	879	**50,120**	0,0000
AL589743.1	22	1165	**76,110**	0,0000
as-lncRNAs				
FERMT1 as-lncRNA	1108	5	**0,006**	0,0000
PVRL3 AS1	1572	20	**0,018**	0,0000
MACROD2 as-lncRNA	7683	120	**0,023**	0,0000
IL6R as-lncRNA	226	5	**0,030**	0,0001
ZFHX4 AS1	259	17	**0,096**	0,0052
DMD as-lncRNA	971	241	**0,365**	0,0801
DCAF6 as-lncRNA x10	174	241	**2,044**	0,4108
SLC9A as-lncRNA	13	233	**27,420**	0,0000

We tested first the level of hepatitis A virus cellular receptor 1 (HAVCR1) and myosin 1D (MYO1D), mRNAs up regulated in HEK-Epi cells, as well as interleukin 7 receptor (IL7R), transcription factor AP-2 alpha (TFAP2A) and atypical cadherin 3 (FAT3), mRNAs up regulated in HEK-Mes cells. All transcripts showed the expected specificity of expression in HEK-Epi and HEK-Mes cells (Figure 24a). We observed that fold-changes obtained by RT-qPCR were not comparable with those obtained by RNA-seq data analysis. This could be explained by the fact that RNA-seq library preparation and RT-qPCR are two completely distinct protocols. RNA-seq libraries are prepared by depletion of ribosomal RNAs, fragmentation, reverse transcription and amplification of cDNA fragments by PCR, whereas RT-qPCR was performed directly with total RNA extracts.

Interestingly, lncRNAs and as-lncRNAs showing important and significant fold changes, both by RNA-sequencing and RT-qPCR, were PVRL3 AS1, MACROD2 as-lncRNA, up regulated in HEK-Epi cells, and HOTAIR, SLC9A as-lncRNA, AL589743.1, up regulated in HEK-Mes cells (Figure 24b). Among them, only PVRL3 AS1 and AL589743.1 have been previously annotated. PVRL3 AS1 is an antisense transcript to poliovirus receptor-related 3 gene (Bonaldo et al. 1996). If its role is not known, its sense gene encodes a protein functioning as adhesion molecule at adherens junctions and interacting with other proteins involved in regulation of cell motility, proliferation and survival (Rikitake et al. 2012). AL589743.1 has been already annotated as a new large non-coding RNA in mammals (Guttman et al. 2009), but its function is not known. MACROD2 as-lncRNA was one of the novel antisense non-coding RNAs identified by RNA-seq data analysis, antisense to "MACRO domain containing 2 gene". MACROD2 encodes a protein that interacts with mono(ADP-ribosyl)ation on target proteins (Feijs et al. 2013), and has been annotated as a region of high genomic instability involved in several diseases such as autism (Tsang et al. 2013) and cancer (Rajaram et al. 2013). MACROD2 mRNA level was significantly decreased in HEK-Mes compared to HEK-Epi cells. In the same way, SLC9A as-lncRNA is a novel identified antisense RNA to SLC9A gene, encoding Na(+)/H(+) exchangers with increasing apparent contribution to the pathophysiology of multiple human diseases (Fuster & Alexander 2014). Interestingly, SLC9A mRNA was significantly decreased in HEK-Mes compared to HEK-Epi cells, consistently with the idea that antisense non-coding transcripts could be involved in regulation of sense mRNA expression. Finally, HOTAIR has been widely studied for its regulatory role and its up regulation in a majority of human cancers (see Introduction). HOTAIR was selected for further functional studies as the best candidate lncRNA with a putative role in EMT. Among the tested lncRNAs and as-lncRNAs, only ZFHX4 AS1 showed no significant differential expression.

In conclusion, high-throughput RNA-sequencing allowed us to identify differentially expressed transcripts between immortalized HEK-Epi and HEK-Mes cells, establishing a signature of mRNAs and non-coding RNAs. These transcriptome changes seem to be highly specific to immortalized HEK cells, upon EMT acquired after beginning of telomere instability and chromosome rearrangements. To verify this statement, and assess the biological relevance of our signature, we tested several of our candidates in a dynamic, non-

immortalized model of HEK cells, and upon TGF-β1 induced EMT in immortalized HEK-Epi cells.

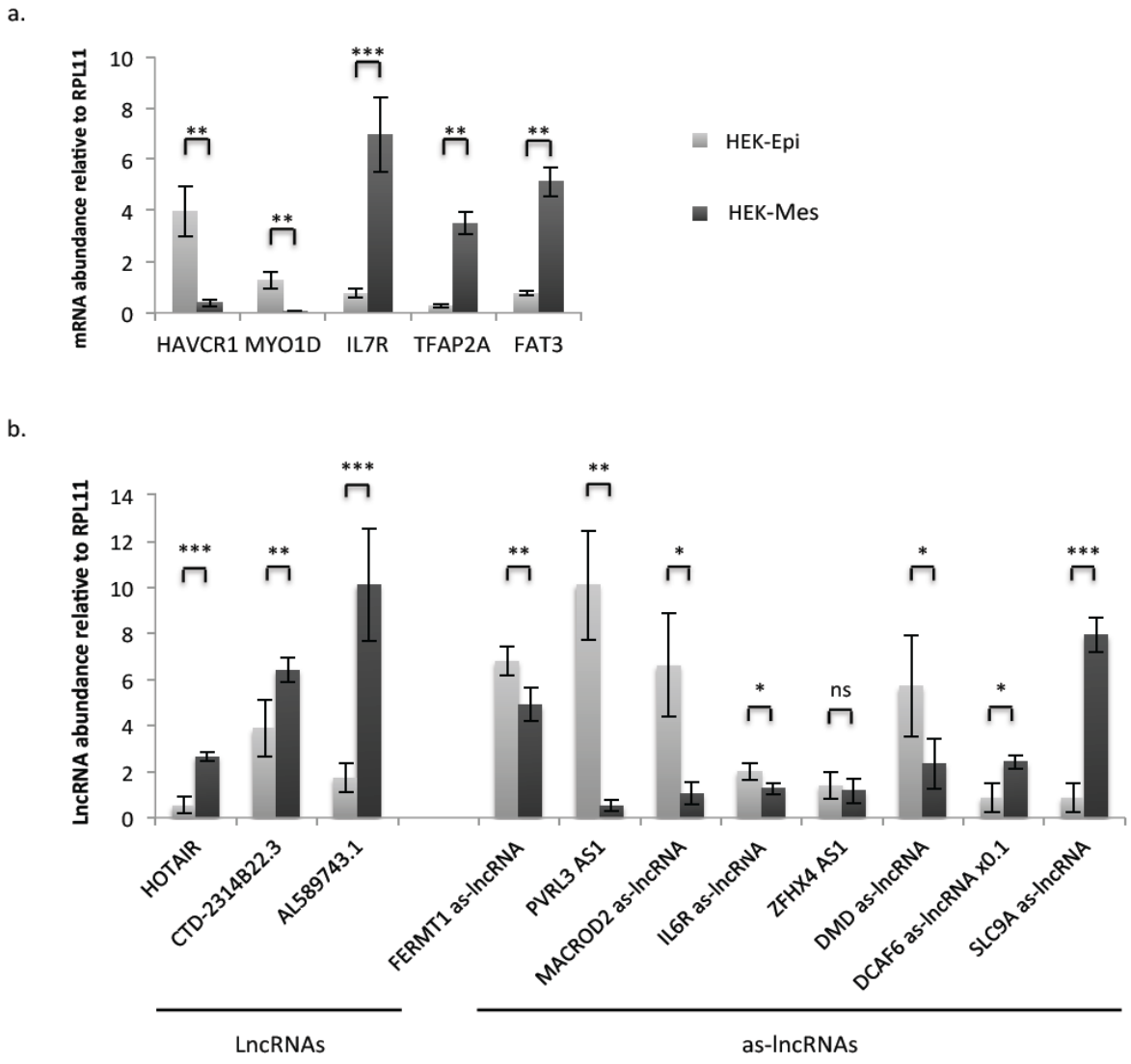

Figure 24 | Differential expression of coding and non-coding RNAs in HEK-Epi and HEK-Mes cells. **a.** Relative quantification by random-primed RT-qPCR of mRNAs identified as deregulated between HEK-Epi and HEK-Mes cells by DEseq analysis of RNA-seq data. **b.** Relative quantification by oligo-specific RT-qPCR of lncRNAs and as-lncRNAs identified as deregulated between HEK-Epi and HEK-Mes cells by DEseq analysis of RNAseq data. **a, b.** Error bars indicate standard deviation of three independent experiments; Student t-test was used to determine the statistical significance: * $p<0.05$, ** $p<0.01$, *** $p<0.0001$, ns not significant.

b. LncRNAs expression in dynamic non-immortalized model of HEK cells

In parallel with HEK-Epi and HEK-Mes cells immortalization, RNAs have been extracted by L.J. Castro-Vega from cells cultured without exogenous introduction of hTERT, the HEK dynamic system (Figure 25a). We analysed by random-primed (mRNAs) or oligo-specific (lncRNAs and as-lncRNAs) RT-qPCR RNA abundance of EMT markers and differentially expressed transcripts between HEK-Epi and HEK-Mes cells, in primary HEK cells and cells at PD 33 and 60. Among tested EMT markers, only KRT19 showed the expected change in expression, with a decrease in mRNA level as observed in HEK-Epi (immortalized at PD30) and HEK-Mes cells (PD60) (Figure 25b). In the same way, mRNAs identified as differentially expressed in HEK-Epi and HEK-Mes cells showed different changes in expression in the HEK dynamic system. We observed stable expression of HAVCR1, MYO1D, TFAP2A and FAT3 between PD33 and PD60. Only IL7R showed a strong and significant increase in PD60 compared to PD33, as observed in HEK-Mes compared to HEK-Epi cells (Figure 25c). To explain these differences with HEK immortalized cells, we can hypothesize that stable mesenchymal features were acquired in HEK-Mes cells because of immortalization by introduction of exogenous telomerase. We can also speculate that HEK dynamic cell system display a well-controlled balance between EMT and MET, explaining the observed incoherent expression of EMT markers.

Interestingly, even considering that dynamic HEK cells display a transient EMT phenotype, as-lncRNAs showed expression profiles following differential expression measured by RNA-sequencing and RT-qPCR in immortalized HEK-Epi and HEK-Mes cells. PVRL3 AS1 and MACROD2 as-lncRNA showed a strong decrease between PD33 and PD60. HOTAIR, SLC9A as-lncRNA and AL589743.1 exhibited an increased level between PD33 and PD60 (Figure 25d). This result might indicate that lncRNA and as-lncRNA signatures established in HEK immortalized cells could reflect the evolution of non-coding transcriptome in natural early EMT reprogramming.

c. LncRNAs expression in TGF-β1-induced EMT

We then tested whether the identified transcripts showed differential expression upon EMT induced in epithelial cells by treatment with TGF-β1. Interestingly, HOTAIR lncRNA has been associated with EMT induced by TGF-β1 in HCC1954 cells (Pádua Alves et al. 2013).

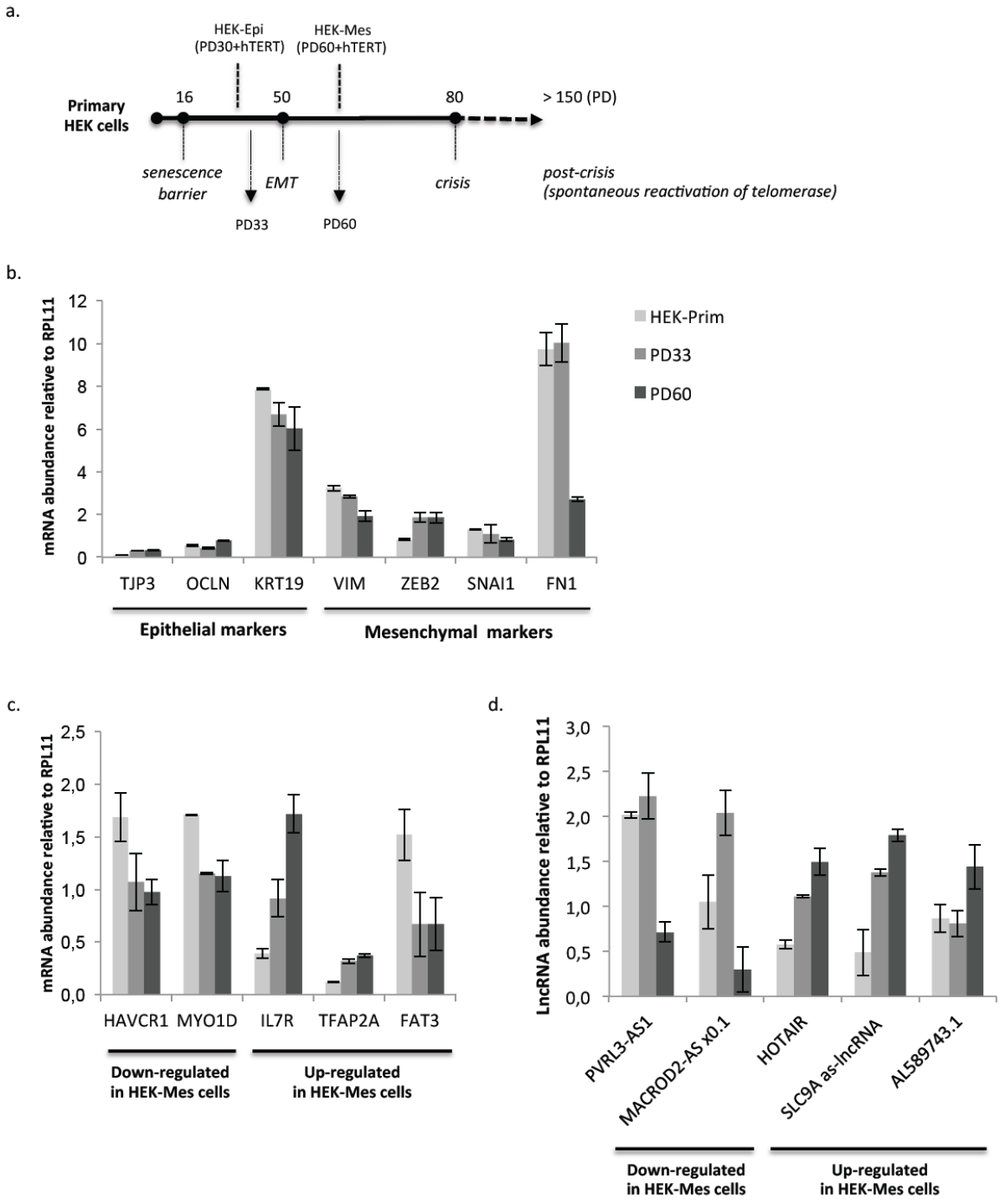

Figure 25 | RNAs expression in HEK dynamic system. **a.** Description of the HEK dynamic cell system. RNAs were extracted from primary HEK cells, and after 33 and 60 population doublings. **b.** Random-primed RT-qPCR quantification of EMT markers mRNAs, in HEK dynamic model. **c.** Random-primed RT-qPCR quantification of mRNAs identified by RNA-seq as differentially expressed between HEK-Epi and HEK-Mes immortalized cells. **d.** Oligo-specific RT-qPCR quantification of LncRNAs identified by RNA-seq as differentially expressed between HEK-Epi and HEK-Mes immortalized cells. Error bars represent standard deviation in three independent experiments.

To validate our approach, we treated HCC1954 immortalized, well-differentiated breast epithelial cells with 20 ng of human recombinant TGF-β1 (240-B-010, R&D Systems) during 72 hours. Random-primed RT-qPCR quantification of EMT markers mRNA showed a decrease of the TJP3 epithelial marker, and an increase of VIM, ZEB2, SNAI1 and FN1 mesenchymal markers expression level compared to untreated mock conditions, proving that EMT occurred in these cells (Figure 26a). Consistently with published results, HOTAIR showed an increased expression after TGF-β1 treatment (Figure 26b).

We treated HEK-Epi cells with TGF-β1 following the same protocol. We observed a decrease in level of epithelial markers KRT19 and CTNNB1 and an increase in level of mesenchymal markers VIM, SNAI1 and FN1, but these changes were quite low compared to changes obtained in HCC1954 (Figure 26c). It has been previously shown that in immortalized but not tumorigenic cells, such as HEK-Epi cells, short TGF-β1 treatment induces only transient EMT, reverted as soon as TGF-β1 is removed from the culture medium. More than 8 days of treatment are necessary to initiate the establishment of mesenchymal state even after the removal of TGF-β1 (Gregory et al. 2011). Thus, we hypothesized that the measured levels of EMT markers were reflecting only a transient transition, and tested longer treatment.

After 12 days of TGF-β1 treatment, EMT markers expression showed strong differences between mock and treated conditions. Epithelial markers KRT19 and TJP3 were decreased, mesenchymal markers VIM, SNAI1 and FN1 were strongly increased, suggesting that EMT has been robustly induced. Curiously, ZEB2, however known as the early up regulated gene upon TGF-β1 treatment, showed no significant change after 72 hours or 12 days of treatment (Figure 26d). Under microscopic evaluation, HEK-Epi cells after 12 days of TGF-β1 treatment showed an elongated morphology, close to the one observed for HEK-Mes cells. Phalloïdin-TRITC (1:1000, P5282, Sigma-Aldrich) staining, allowing the analysis of F-Actin fibers distribution, revealed that HEK-Epi treated cells displays an EMT-like phenotype (Figure 26e). In these cells, we measured by oligo-specific RT-qPCR the level of lncRNAs and as-lncRNAs identified as differentially expressed between HEK-Epi and HEK-Mes cells. Unexpectedly, the majority of the tested transcripts showed no change between treated and control HEK-Epi cells. Only AL589743.1 showed the expected tendency, with a strong increase between control and treated HEK-Epi cells.

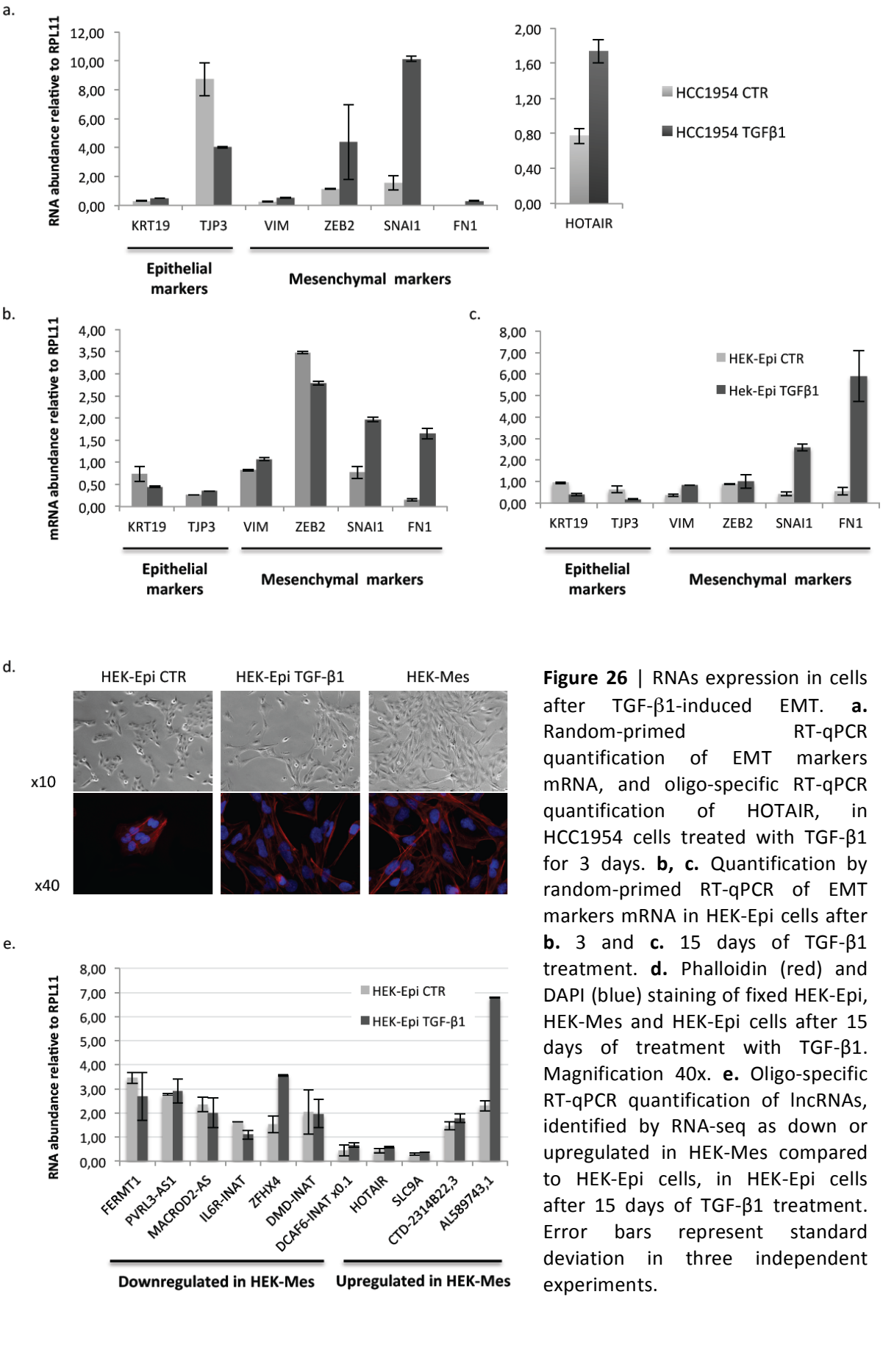

Figure 26 | RNAs expression in cells after TGF-β1-induced EMT. **a.** Random-primed RT-qPCR quantification of EMT markers mRNA, and oligo-specific RT-qPCR quantification of HOTAIR, in HCC1954 cells treated with TGF-β1 for 3 days. **b, c.** Quantification by random-primed RT-qPCR of EMT markers mRNA in HEK-Epi cells after **b.** 3 and **c.** 15 days of TGF-β1 treatment. **d.** Phalloidin (red) and DAPI (blue) staining of fixed HEK-Epi, HEK-Mes and HEK-Epi cells after 15 days of treatment with TGF-β1. Magnification 40x. **e.** Oligo-specific RT-qPCR quantification of lncRNAs, identified by RNA-seq as down or upregulated in HEK-Mes compared to HEK-Epi cells, in HEK-Epi cells after 15 days of TGF-β1 treatment. Error bars represent standard deviation in three independent experiments.

Interestingly, ZFHX4-AS1, which showed no significant change between HEK-Epi and HEK-Mes cells in RT-qPCR experiments contrary to RNA-seq data analysis, was significantly up regulated upon TGF-β1 treatment (Figure 26f).

These observations led us to think that transcripts identified as differentially expressed in HEK-Mes compared to HEK-Epi cells are not regulated by TGF-β1 signalling pathway. Alternative pathways could be involved in the induction of EMT. It has been shown in HEK EMT model that forced expression of miRNAs from the miR-200 family was sufficient for HEK-Mes cells to recover some epithelial characteristics (Castro-Vega et al. 2013). Therefore, in the HEK model, miRNAs down regulation seems to be critical in the induction of EMT. We can hypothesize that this down regulation is due to genomic instability since it has been shown that immortalized human bronchial epithelial cells exposed to tobacco carcinogens display a persistent dedifferentiation program marked by EMT (Tellez et al. 2011). If the EMT that occurred in the HEK model was due to genomic instability and miR down regulation, TGF-β1 treatment in HEK-Epi cells would recapitulate only partially the phenotype observed in HEK-Mes cells. This can be the reason why we did not observe the same differential expression of lncRNAs and as-lncRNAs in HEK-Epi treated cells compared to control cells and HEK-Mes compared to HEK-Epi cells. But we can also speculate that these transcripts define the mesenchymal identity of HEK cells, without being directly involved in EMT induction.

3. The role of HOTAIR in EMT

Previous studies have shown that HOTAIR overexpression increases invasiveness of epithelial cell lines. Its aberrant expression in numerous human cancers has been directly linked to metastasis process and poor survival rate (see Introduction). It already had been shown that HOTAIR when overexpressed induce gene silencing via recruitment of PRC2 and LSD1/CoREST/REST complexes, but the question of how HOTAIR could be involved in the formation of metastases, presumably achieved via the EMT, still remained to be answered.

Whole transcriptome analysis by RNA-seq of HEK-Epi and HEK-Mes cells identified HOTAIR as the most up regulated lncRNA in mesenchymal cells. This result was visualized using Ving, a bioinformatics tool developed in our lab (M. Descrimes, Y. Ben-Zouari) to

establish snapshots of transcripts read density from RNA-seq datasets (Figure 27a). We confirmed that HOTAIR was indeed overexpressed in HEK-Mes compared to HEK-Epi cells, showing a significant (p-value < 0.0001) 7.5 fold-change in a mean of 10 independent oligo-specific RT-qPCR experiments (Figure 27b). This observation led us to examine HOTAIR role in EMT using loss- and gain-of-function approaches to analyse associated phenotypes and establish a list of HOTAIR target genes by high-throughput RNA-sequencing.

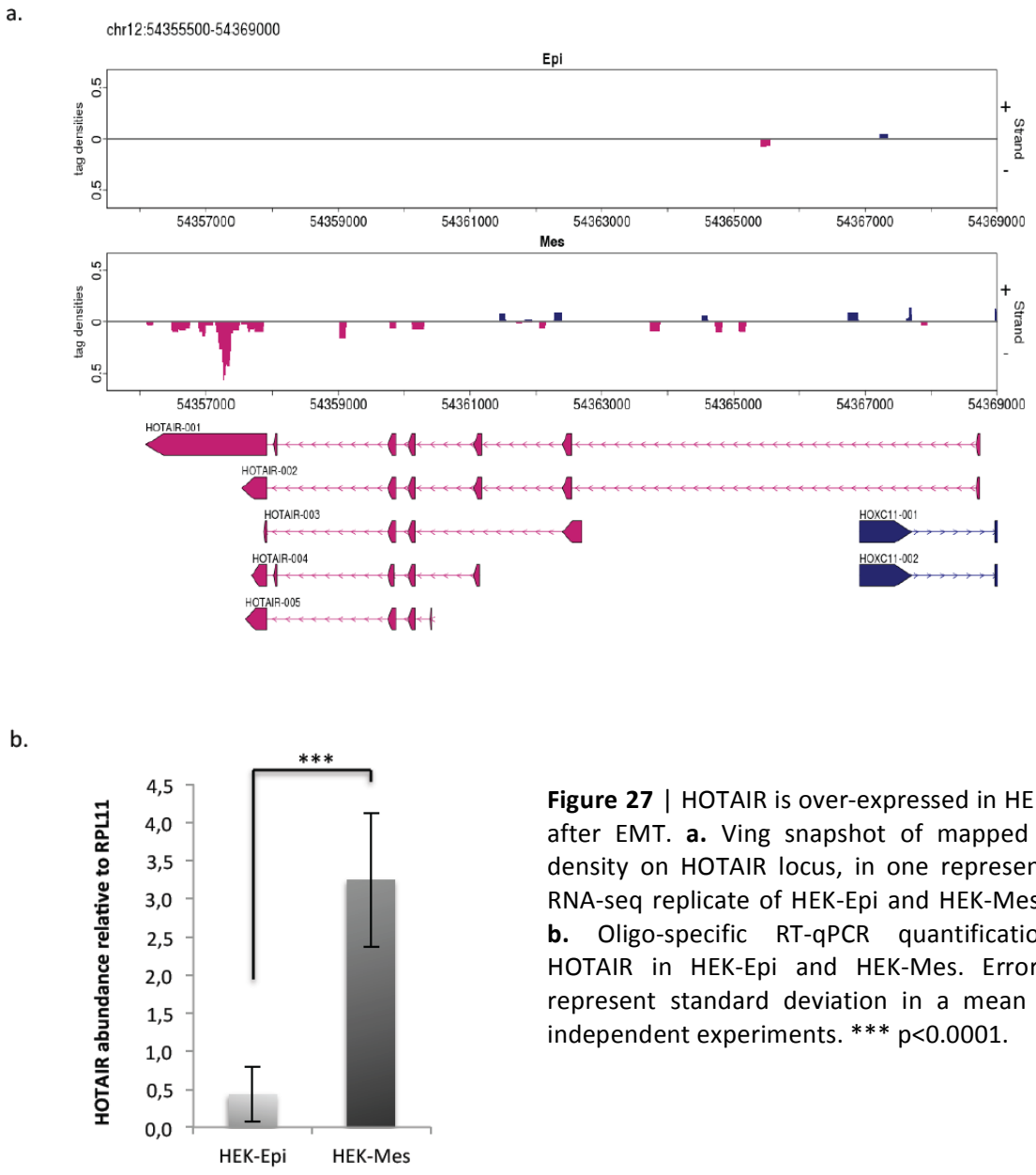

Figure 27 | HOTAIR is over-expressed in HEK cells after EMT. **a.** Ving snapshot of mapped reads density on HOTAIR locus, in one representative RNA-seq replicate of HEK-Epi and HEK-Mes cells. **b.** Oligo-specific RT-qPCR quantification of HOTAIR in HEK-Epi and HEK-Mes. Error bars represent standard deviation in a mean of 10 independent experiments. *** p<0.0001.

a. HOTAIR depletion reduces proliferation, migratory capacity and invasiveness of mesenchymal cells

i. HOTAIR depletion approaches

- HOTAIR depletion by siRNA

siRNAs are commonly used to deplete target RNAs. Double-stranded siRNAs are transfected into cells, once in the cytoplasm they form complexes with Dicer and are loaded into the RNA-induced silencing complex (RISC). siRNAs strands are separated, activating RISC that cleaves the RNA targeted by the siRNA sequence (Figure 28a). We used a pool of three commercial siRNAs (Sigma-Aldrich), all targeting the exon localized at the 3' extremity of HOTAIR and common to all annotated HOTAIR variants (siHOT) (Figure 28b). Our negative control was a siRNA targeting GFP (siCTR) (Sequences of siRNAs are available in Supplementary Table S3). Transfections were performed using Lipofectamine 2000 (Life Technologies) with 100 nM of siRNAs. Total RNAs were extracted after 48 hours of transfection. Oligo-specific RT-qPCR quantification of HOTAIR showed that 56% of HOTAIR was still detectable 48 hours after siRNA transfections, and we observed a high heterogeneity in the efficiency of HOTAIR depletion between 5 independent experiments (Figure 28c). This result could be explained by the fact that siRNAs have been originally designed to target mRNAs, for a majority localized in the cytoplasm of cells, whereas HOTAIR is mostly nuclear.

To examine whether the depletion of HOTAIR was efficient enough to observe effects on gene expression, we measured by RT-qPCR in the three independent experiments showing the most efficient depletion of HOTAIR, the mRNA level of genes identified as differentially expression upon HOTAIR over expression in breast cancer cell line MDA-MB-231 (Gupta et al. 2011). These genes showed no significant variation between siCTR and siHOT (SNAI1, LAMC2, JUB, SIRT2, GATA2, BDNF), or variations that were not consistent with published results. Indeed, LAMB3 and ABL2, showed to be induced by HOTAIR high expression, exhibited increased level in HOTAIR-depleted cells. On the contrary, we observed decreased levels of HOXD10, PCDH10 and PCDHB5, showed to be down regulated in presence of HOTAIR (Figure 28d).

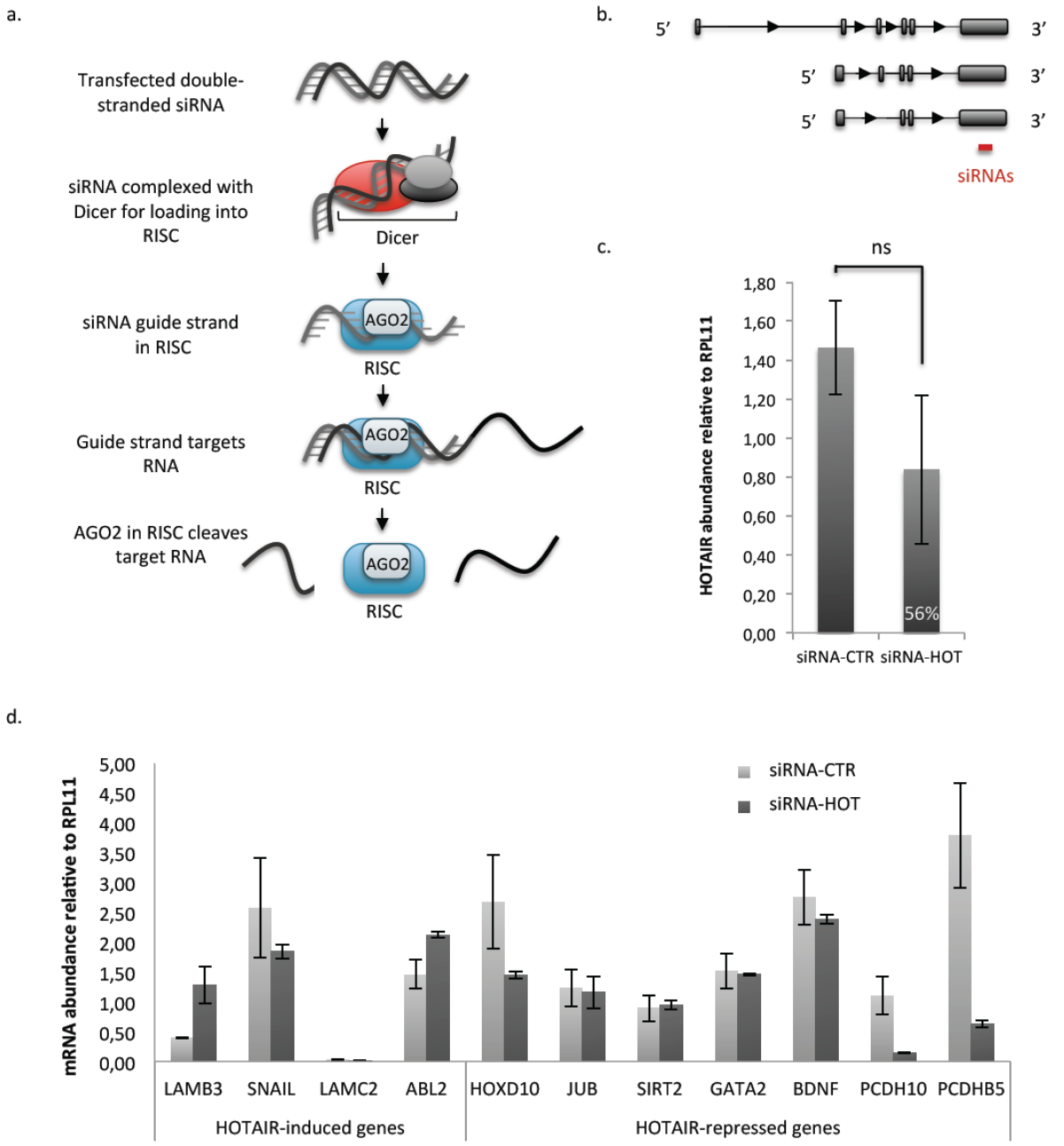

Figure 28 | HOTAIR depletion by siRNA. **a.** Schematic representation of HOTAIR depletion by siRNA. RNAs have been extracted from HEK-Mes cells after 48 hours of transfection, with 100 nM siRNA-CTR or siRNA-HOT. **b.** siRNA position on HOTAIR isoforms. **c.** Oligo-specific RT-qPCR quantification of HOTAIR abundance after siRNA transfection in HEK-Mes cells. **d.** Random-primed RT-qPCR quantification of mRNAs from genes previously identified as HOTAIR targets *(Gupta et al. 2010)*. Error bars represent standard deviation in five (**c.**) or three (**d.**) independent experiments. ns, not significant.

Three different hypotheses have to be considered to explain these results. First, we can assume that residual HOTAIR level after depletion could be enough to ensure HOTAIR-

mediated gene regulation. Second, 48 hours of depletion might be not sufficient to induce HOTAIR-dependent epigenetic reprogramming with further consequences on steady-state mRNA levels. Finally, it is important to consider that HOTAIR targets are not well defined, since they seem to be highly specific to the cell line. Indeed, two systematic studies were performed, in breast cancer cells exogenously expressing HOTAIR (Gupta et al. 2011), and in pancreatic cells depleted for HOTAIR by siRNA (Kim et al. 2013). Only 241 genes are common among 9,260 and 1,006 differentially expressed genes after HOTAIR over expression and knock down, respectively. We could imagine that HOTAIR-target genes in our specific model of HEK cells are different from reported studies.

- HOTAIR depletion by Antisense Oligonucleotides

Given the low and non-robust efficiency of HOTAIR depletion by siRNA, we used Antisense Oligonucleotides (ASO), in collaboration with ISIS Pharmaceuticals. ASOs are short single-stranded DNA molecules, bearing several modifications in the phosphodiester backbone. These patented modifications allow high stability and uptake into cells without any transfection reagent, resulting in a very low toxicity for the cells. Once these DNA molecules are delivered to the nucleus, they can pair to a complementary RNA target and induce RNaseH-mediated RNA-DNA pair degradation (Figure 29a). These molecules have already been proved to be highly efficient against nuclear lncRNAs as MALAT1 (Gutschner et al. 2013).

ISIS Pharmaceuticals proposed 5 ASOs against HOTAIR, validated among 156 in T47D cells. We performed several experiments to test their efficiency in HEK-Mes cells, and selected two of them (ASO-HOT1, ASO-HOT2) showing robust HOTAIR depletion after 5 days of treatment with 10 μM of ASO (Sequences of ASOs are available in Supplementary Table S3). These two ASO sequences are targeting the same exon of HOTAIR in the 3' region common to all annotated variants (Figure 29b). All results were compared to cells treated with a control ASO (ASO-CTR), a scramble DNA sequence not targeting any known RNA in the cells. We obtained a significant and reproducible depletion with both ASOs in three independent experiments, with a mean 33% of HOTAIR left after ASO-HOT1 treatment (p-value < 0.01) and 51% left after ASO-HOT2 treatment (p-value < 0.05) (Figure 29c). We used both ASOs to study changes in phenotypes of HOTAIR-depleted HEK-Mes cells, to exclude any off-target effects.

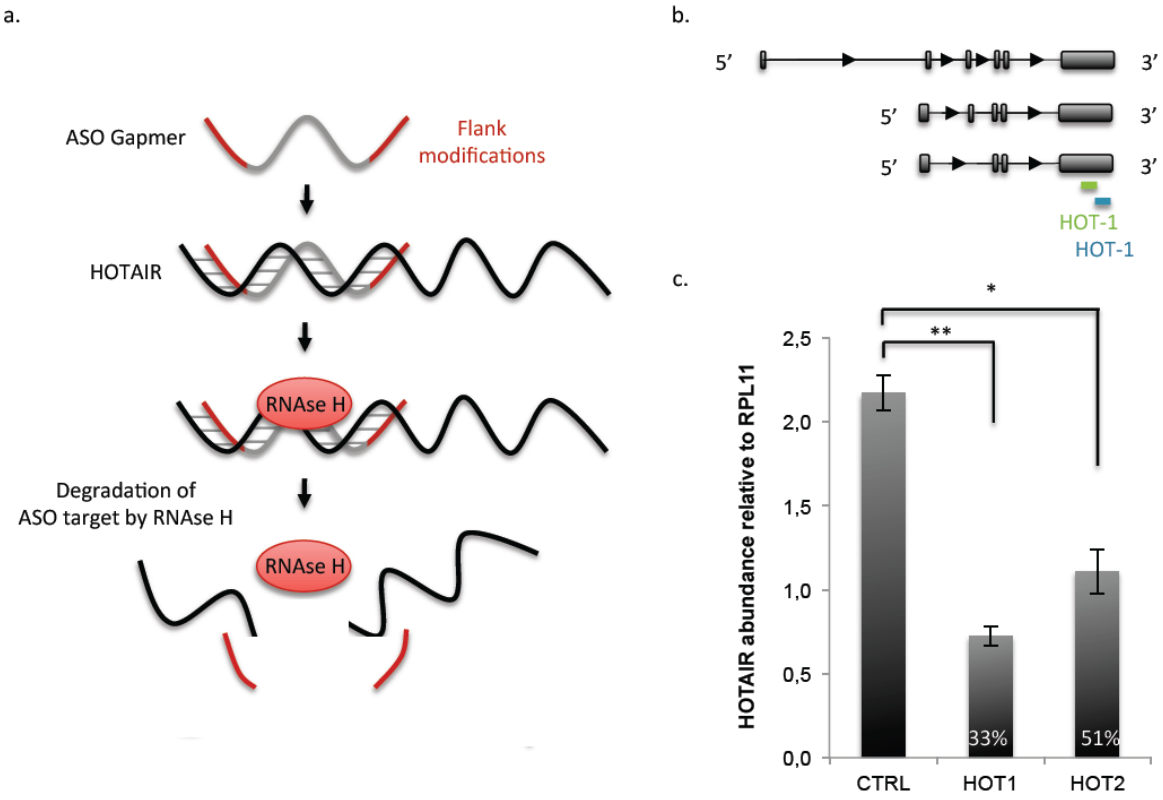

Figure 29 | HOTAIR depletion by ASOs. **a.** Schematic representation of Antisense Oligonucleotides-mediated knock-down of HOTAIR. ASOs induce RNase H degradation of their hybridization target. **b.** ASO HOT1 and HOT2 positions on known HOTAIR isoforms. **c.** Oligo-specific RT-qPCR quantification of HOTAIR abundance in HEK-Mes cells after 5 days of treatment with control ASO (ASO-CTR) or ASOs targeting specifically HOTAIR (ASO-HOT1 and ASO-HOT2). Percentage of HOTAIR RNA still detectable in the cells at the end of the treatment is indicated. Error bars indicate standard deviation in three independent experiments; * $p<0.05$, ** $p<0.01$.

ii. HOTAIR-depleted HEK-Mes cells show decreased proliferation, motility and invasiveness

We evaluated cell proliferation by counting cells in an exponentially growing population, each 24 hours after beginning of ASO treatment. Cells treated with ASO-CTR showed a small decrease in proliferation, compared to HEK-Epi and HEK-Mes cells, suggesting that ASO treatment *per se* exhibit a low toxicity affecting cell growth. Cells treated with ASO-HOT1 and ASO-HOT2 showed a decreased proliferation compared to ASO-CTR (Figure 30a). Proliferation decrease was more pronounced in cells treated with ASO-HOT2, which was quite surprising considering its lower efficiency. In parallel, we observed that cell viability was not affected by ASO treatment, using trypan blue coloration of cells *(data not shown)*.

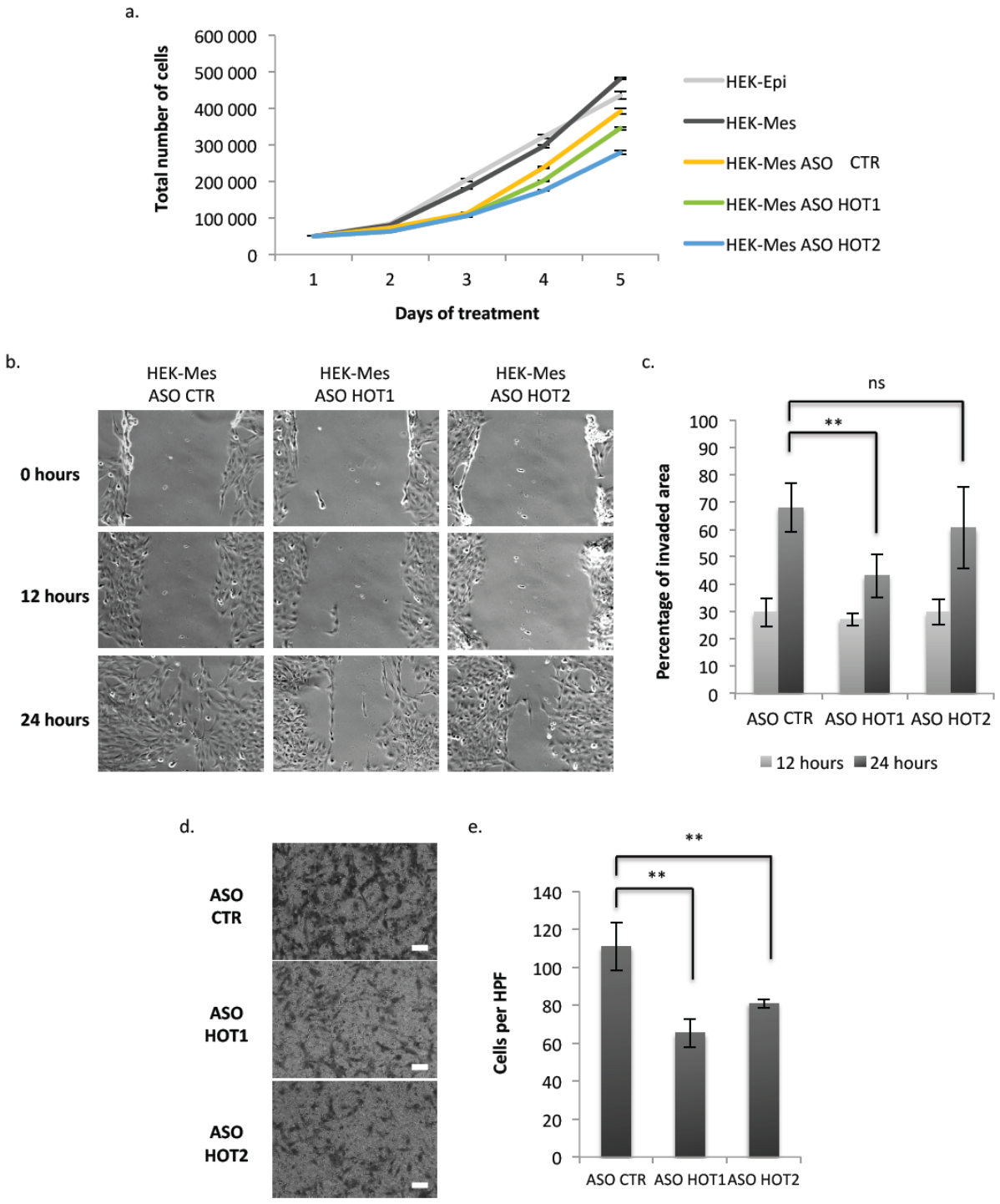

Figure 30 | Effect of HOTAIR depletion on proliferation, motility and invasiveness of HEK-Mes cells. **a.** Cell proliferation of HEK-Epi, HEK-Mes and HEK-mMes cells treated with ASO-CTR and ASOs targeting HOTAIR. **b.** Wound healing assay was used to assess cells motility. Phase contrast images show wound recovery at 0, 12 and 24h post-scratch, in 10x magnification. **c.** Histograms represent the estimated percentage of the invaded area, taking first picture at 0h as 20% invasion reference, and in a mean of 6 contrast phase images. **d.** Matrigel invasion assay was performed to assess cells invasion capacity. A mean of 10 high power fields was taken 72h after seeding 200.000 cells on membranes coated with matrigel, in three independent experiments. Scale bar = 100 µm. **e.** Counted number of cells per HPF. Error bars indicate standard deviation; ** $p<0.01$, ns, not significant.

Then, we performed a wound-healing assay to assess cells migration rate in the last 24 hours of 5 days ASO treatment. Quantification of invaded wound showed a significant 1.6 decrease (p-value < 0.01) after 24 hours for HEK-Mes cells treated with ASO-HOT1 compared to ASO-CTR, whereas cells treated with ASO-HOT2 showed a 1.1 decrease in wound recovery which appeared to be not significant in a mean of 6 HPF per sample (Figure 30b and c). Invasion assay showed a significantly decreased invasiveness of HEK-Mes cells upon HOTAIR depletion (p-value < 0.01), 1.7 and 1.4 folds for ASO-HOT1 and ASO-HOT2 when compared to ASO-CTR, respectively (Figure 30d and e).

These results showed that ASO-mediated depletion of HOTAIR induces a decrease in HEK-Mes cells proliferation, migration capacity and invasiveness. These effects seem to be sensitive to HOTAIR level, since more efficient depletion of HOTAIR showed stronger changes in cell migratory and invasive properties. It is worth to note that equivalent effects on HEK-Mes cells phenotype were observed in HOTAIR depletion experiments using siRNAs *(data not shown)*.

iii. HOTAIR expression is required for maintenance of β-catenin levels in HEK-Mes cells

To examine whether the EMT signature of HEK-Mes had changed in terms of EMT markers expression, we measured by random-primed RT-qPCR the level of epithelial and mesenchymal markers mRNA in ASO-treated HEK-Mes cells. We observed no significant variations in mRNA abundance of tested EMT markers between ASO-CTR, ASO-HOT1 and ASO-HOT2, with a high heterogeneity between three independent experiments (Figure 31a). Western blot analysis of EMT markers VIM and ACTA2 showed no variation between ASO-CTR, ASO-HOT1 and ASO-HOT2. The epithelial marker β-catenin showed a decrease in ASO-HOT1 treated cells compared to ASO-CTR (Figure 31b).

β-catenin has been shown to have a dual role, structural and signalling, through its structural composition that allows binding of numerous interaction partners, at the membrane, in cytosol and in the nucleus. β-catenin's partners share overlapping binding sites, ensuring their mutual exclusivity (Valenta et al. 2012; Lyashenko et al. 2011). In epithelial cells and in the absence of a Wnt stimulus, β-catenin is highly expressed, localized at the cytoplasmic side of the membrane, and interacts with E-cadherin as a component of cadherin-based cell-cell

connections (Hinck et al. 1994; Meng & Takeichi 2009). In the cytoplasm, free β-catenin, not bound to E-cadherin, is rapidly recognized by the destruction complex, and targeted for degradation (Kimelman & Xu 2006).

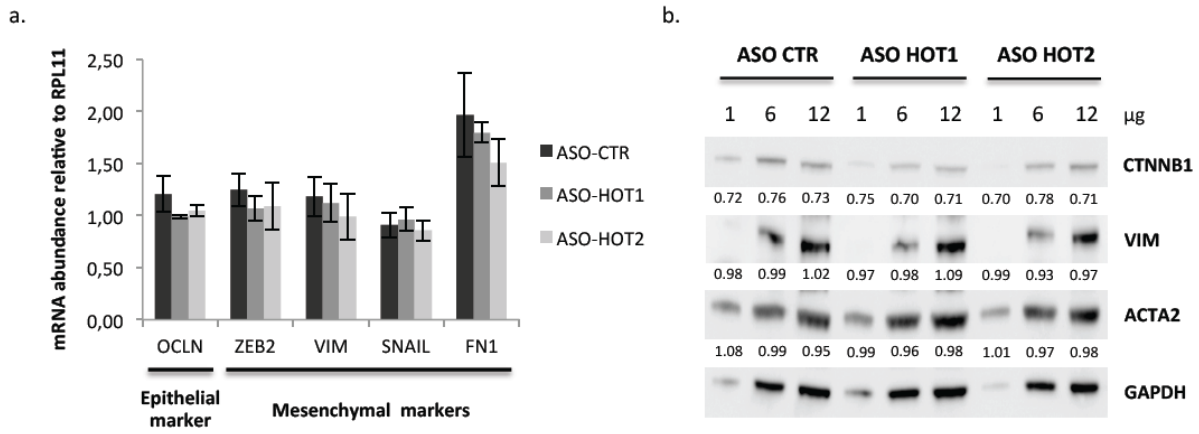

Figure 31 | EMT signature and transcriptome changes after HOTAIR knockdown. **a.** Random-primed RT-qPCR quantification of EMT markers mRNA abundance in HEK-Mes cells treated for 5 days with ASO-CTR or ASO targeting HOTAIR (HOT1 and HOT2). Error bars represent standard deviation in three independent experiments. **b.** Western blot detection of EMT protein markers. Numbers indicate protein level quantification relative to GAPDH.

Activation of Wnt signalling induces a cascade of target genes regulation, including the transcription factor Twist, which directly down regulates the expression of E-cadherin while also inducing expression of N-cadherin and fibronectin, inducing cellular invasion and EMT (Howe et al. 2003; Yang et al. 2007; Yang et al. 2004). Wnt signalling activation also leads to the disassembly of the destruction complex (Li et al. 2012; Clevers & Nusse 2012). This activation results in increased free cytoplasmic levels of β-catenin, which could directly interact with different nuclear pore complex components and enter into the nucleus. In the nucleus, β-catenin regulates transcription of Snail1 and Snail2, both involved in repression of E-cadherin transcription (Conacci-Sorrell et al. 2003; Barrallo-Gimeno & Nieto 2005), but also ZEB1, that transcriptionally represses epithelial markers, but activates mesenchymal genes (Sánchez-tilló et al. 2011). Altogether, these regulations result in the acquisition of the EMT-like phenotype and increased invasive properties of cells. Indeed, mutations in several factors, frequently resulting in hyper activation of Wnt/β-catenin signalling, have been detected in numerous human cancers and correlated with high metastasis rate and poor prognosis.

On the other hand, it has recently been shown that HOTAIR overexpression in oesophageal squamous cell carcinoma induces H3K27 methylation in promoter region of Wnt Inhibitory Factor-1 (WIF-1) (Ge et al. 2013), resulting in gene silencing and activation of Wnt/β-catenin signalling pathway, promoting tumor invasion and migration (Rubin et al. 2010).

Considering these observations, a decrease in β-catenin protein level could suggest that HOTAIR depletion in HEK-Mes cells induces a down regulation of the Wnt signalling pathway, reactivating free cytoplasmic β-catenin degradation by the destruction complex. This could explain the immediate phenotypic changes observed for HEK-Mes cells upon HOTAIR depletion. Down regulation of Wnt signalling pathway could be due to WIF-1 gene, which is expressed at low HOTAIR levels. Alternatively, other genes can be involved, given the high diversity of mutations resulting in WNT signalling pathway hyper activation in cancer (Anastas & Moon 2013). On the other hand, the absence of variations among other EMT markers at mRNA and protein levels could suggest that Mesenchymal-to-Epithelial Transition (MET) can not be induced by HOTAIR depletion. We can imagine that ASO treatment is not efficient or long enough to induce a complete HOTAIR-dependent epigenetic reprogramming. We can also speculate that HOTAIR is not a driver of EMT, but is necessary for maintenance of mesenchymal cells proliferation and migration capacities.

iv. HOTAIR-depleted HEK-Mes cells showed weak variations in protein-coding and non-coding transcriptome

To go further in the analysis of HOTAIR-depletion effects in HEK-Mes cells, we examined protein-coding transcriptome changes that could occur upon ASO treatment. First, we measured as previously mRNA levels of some genes identified as HOTAIR targets in previous study (Gupta et al. 2010). As shown upon siRNA transfection, we observed no significant changes (LAMB3, HOXD10), or changes that were contradictory with published results (PCDH10, PCDHB5, published as repressed by HOTAIR, but decreased in HOTAIR-depleted cells) (Figure 32). To identify new HOTAIR targets specific to our cell model, we performed by high-throughput RNA-sequencing of HEK-Mes cells upon HOTAIR depletion by ASOs.

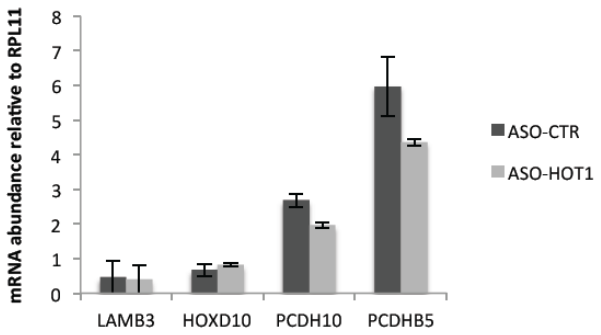

Figure 32 | Random-primed RT-qPCR quantification of mRNAs identified by Gupta *et al.* as putative targets of HOTAIR-mediated regulation, in HEK-Mes cells treated with ASO-CTR or ASO-HOT1. Error bars represent standard deviation in three independent experiments.

Total RNAs were extracted from three independent HEK-Mes cell cultures treated for 5 days with ASO-CTR, ASO-HOT1 and ASO-HOT2. We then prepared strand-specific cDNA libraries using TruSeq Stranded Total RNA-seq kit from Illumina. Samples were sequenced on a HiSeq 5500 Illumina sequencer, by the ICGex platform at Institut Curie. Total number of reads, mapped reads, cDNA duplicates and correctly paired reads are presented in Table 7. Data analysis was performed by Zohra Saci, following the bioinformatics pipeline previously described (Figure 19).

Table 7 | RNA-sequencing results for ASO-treated HEK-Mes cells.

Samples	Total number of reads	Mapped reads	% of mapped reads	Duplicates	% of duplicates	Properly paired	% of of properly paired
CTR1	169 527 366	159 042 977	93.82%	47 885 288	30%	136 426 478	86
CTR2	146 801 068	138 823 371	94.57%	41 387 401	30%	119 093 420	86
CTR3	147 845 989	138 725 995	93.83%	37 249 397	27%	121 587 524	88
ASO HOT1-1	150 647 096	141 789 820	94.12%	42 387 577	30%	121 643 292	86
ASO HOT1-2	140 113 447	130 204 126	92.93%	39 898 465	31%	110 980 396	85
ASO HOT1-3	151 098 636	127 748 540	84.55%	41 640 003	33%	106 315 068	83
ASO HOT2-1	143 714 726	133 865 126	93.15%	45 009 902	34%	112 545 196	84
ASO HOT2-2	160 735 194	151 929 901	94.52%	46 626 026	31%	129 265 258	85
ASO HOT2-3	155 907 239	146 359 589	93.88%	44 933 011	31%	125 840 352	86

As shown on transcript density plots, no major variations of coding and non-coding transcripts expression were observed between ASO-HOT1, ASO-HOT2 and ASO-CTR treatments. mRNAs and lncRNAs showed a strong correlation coefficient between ASO-HOT1 and ASO-CTR, as well as between ASO-HOT2 and ASO-CTR (R=0.99 in both cases). Only as-lncRNAs showed a slight global down regulation in cells treated with ASO against HOTAIR, compared to ASO-CTR, but the correlation coefficient between the two conditions was still very high (R=0.97) (Figure 33a and b). These results were confirmed by heat map visualisation of transcripts, showing equivalent expression of mRNAs and lncRNAs, and few variations of as-lncRNAs, between the three tested conditions (Figure 33c).

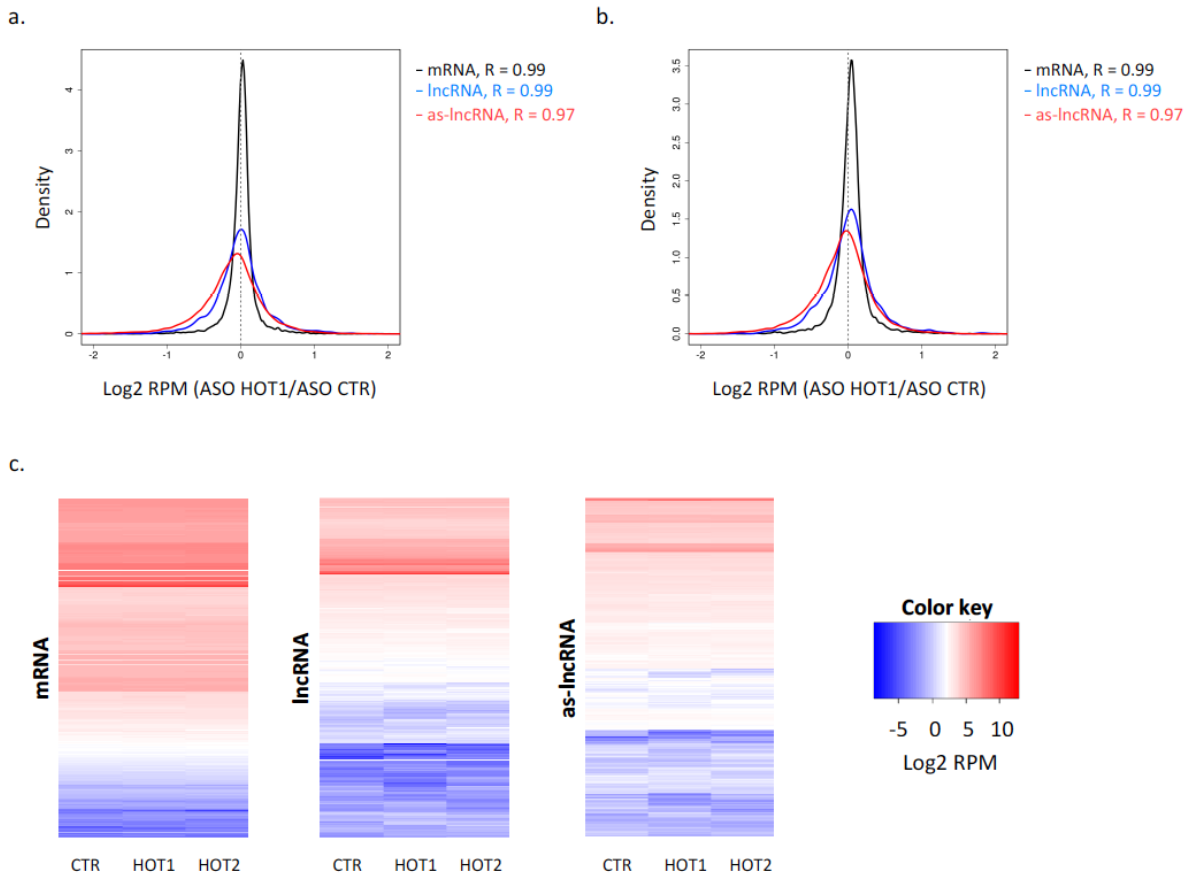

Figure 33 | Genome-wide analysis of HOTAIR depletion effects in HEK-Mes cells. **a.** Transcript density distribution in HEK-Mes ASO HOT1 versus ASO CTR. **b.** Transcript density distribution in HEK-Mes ASO HOT2 versus ASO CTR. RPM: Reads Per Million. **c.** Heatmaps of mRNAs, lncRNAs and as-lncRNAs in ASO-treated HEK-Mes cells full transcriptome.

We first analysed ASO-HOT1 and ASO-HOT2 as two independent samples. In ASO-HOT1 samples transcriptome, compared to ASO-CTR, we identified 131 up regulated

transcripts and 82 down regulated transcripts, including mRNAs, annotated or lncRNAs and as-lncRNAs. Among these transcripts, 4 up regulated and 4 down regulated RNAs showed adjusted p-value ≤ 0.05 in RNA-seq triplicates. Among them, only 3 down regulated transcripts, including HOTAIR passed our criteria of change ≥ 2 folds. In ASO-HOT2 transcriptome compared to ASO-CTR, we identified 140 up regulated transcripts and 211 down regulated transcripts. 16 down regulated and 4 up regulated transcripts showed adjusted p-values considered as significant. HOTAIR was one of them, with a fold change of 0.659 and adjusted p-value < 0.01. Only 2 transcripts in each dataset exhibited a fold change ≥ 2 folds. With the exception of HOTAIR, no overlap was observed between ASO-HOT1 and ASO-HOT2 datasets of differentially expressed genes established by comparison with ASO-CTR.

To examine whether common genes could be found as deregulated in HEK-Mes cells treated with the two different ASOs, we next analysed RNA-seq data using ASO-HOT1 and ASO-HOT2 as replicates of HOTAIR depletion, and compared them with ASO-CTR samples. We identified 6 transcripts significantly up regulated (adjusted p-value ≤ 0.05) among 111, and 10 significantly down regulated among 134, including mRNAs, lncRNAs and as-lncRNAs. HOTAIR was included in this list, with a fold change of 0.556 and adjusted p-value < 0.0001. 4 transcripts in each up and down regulated transcript datasets showed changes ≥ 2 folds.

Interestingly, among these transcripts, we found the down regulated mRNA of bolA family member 2B (BOLA2B) gene, that has been shown to be induced by stress (Santos et al. 1999) and involved in cell-cycle regulation as well as cell proliferation (Division et al. 2002). We also identified as down regulated novel lncRNAs antisense to pumilio RNA-binding family member 2 (PUM2) and euchromatic histone-lysine N-methyltransferase 1 (EHMT1) genes. PUM2 has been shown to be involved in the positive regulation of cellular proliferation in human adipose-derived stem cells (Shigunov et al. 2012). EHMT1 histone methyltransferase forms a heteromeric complex with EHMT2, which methylates H3K9 (Tachibana et al. 2005). This process is crucial for the transcription, signal transduction, proliferation and differentiation of cells (Collins & Cheng 2010; Chin et al. 2007). However, quantifications of these RNAs by random-primed RT-qPCR showed no variations between ASO-CTR, ASO-HOT1 and ASO-HOT2 treated HEK-Mes cells. We hypothesized that RT-qPCR

quantifications were not sensitive enough to detect the low fold changes calculated by RNA-seq data analysis between samples.

In conclusion, HOTAIR depletion by ASOs in HEK-Mes cells results in a strong decrease of cell proliferation, migratory capacities and invasiveness. This phenotype seems to be highly sensitive to HOTAIR levels, as a lower depletion using less efficient ASO results in intermediate effects. A decrease in β-catenin protein level upon HOTAIR depletion suggests that the Wnt/β-catenin signalling pathway is affected. The absence of variations among tested other EMT markers led us to think that EMT program was not reversed by HOTAIR depletion. But the very low number of differentially expressed transcripts between ASO-CTR, ASO-HOT1 and ASO-HOT2 treated cells shows that our ASO treatment doesn't induce major effects on the transcriptome of HEK-Mes cells. This probably means that HOTAIR depletion is not efficient enough or that 5 days are not long enough to reverse HOTAIR-mediated epigenetic modifications. It is worth to note that the gene WIF-1, involved in the regulation of Wnt/β-catenin signalling pathway, doesn't show any detectable variation in RNA level by RNA-seq.

b. HOTAIR gain-of-function study in HEK-Epi cells

i. HOTAIR full-length and truncated forms over expression in HEK-Epi cells

To understand if HOTAIR is a driver of EMT, we performed HOTAIR overexpression in HEK-Epi cells with naturally low HOTAIR levels. Starting from LZRS-HOTAIR plasmid created and provided by Howard Chang's laboratory, we amplified the full length HOTAIR cDNA as well as HOTAIR forms lacking 5' or 3' extremities. Indeed, it has been published that HOTAIR serves as a scaffold for at least two distinct histone modification complexes. A 5' domain of HOTAIR binds PRC2 via its EZH2 and SUZ12 subunits, whereas HOTAIR 3' domain binds the LSD1/CoREST/REST complex (Figure 34a). Once bound by HOTAIR, these two complexes seem to interact with each other, forming a higher complex. PRC2-binding activity was mapped to nucleotides 1-300 of HOTAIR, and LSD1 binding to nucleotides 1 500-2 146 (Tsai et al. 2010). Our aim was to identify transcriptome changes associated with HOTAIR over expression but also to dissociate the effects induced by HOTAIR via PRC2 and via LSD1 binding.

Using the Gateway technology (Life Technologies), we cloned HOTAIR full-length sequence (HOT), HOTAIR truncated for 300 nucleotides at its 5' extremity (HOTΔPRC2) and HOTAIR truncated for 646 nucleotides at its 3' extremity (HOTΔLSD1) (Figure 34b) into an expression vector (pLenti6.2/V5-DEST™ , Life Technologies) under the control of a strong constitutive CMV promoter. We also cloned GFP sequence into the same expression vector, and used it as a negative control for following experiments (Plasmids references can be found in Supplementary Table S4).

All the generated expression vectors were co-transfected with packaging vector (Gag/Pol/Rev/Tet) and envelope vectors (VSV-G) into HEK293T cells. Lentiviruses produced were then harvested and used to infect HEK-Epi cells (see Material and Methods section for detailed protocol) (Figure 34c). Cells were then cultured for approximately 10 population doublings in presence of blasticidine, to eliminate non-transduced cells. RNAs were then extracted from selected cells to test whether HOTAIR and its truncated forms were specifically and efficiently over expressed in the generated cell lines.

By oligo-specific RT-qPCR, we first measured the specificity of GFP, HOTAIR full-length and truncated forms expression in transduced HEK-Epi cells. Indeed, GFP mRNA was detected only in cells transduced with GFP plasmid, and HOTAIR expression was detected at both 5' and 3' extremities for cells infected with the plasmid carrying full-length HOTAIR sequence. From cells infected with plasmids designed for over expression of HOTAIR truncated forms HOTΔPRC2 and HOTΔLSD1, we detected HOTAIR expression only at 3' or 5' extremity, respectively *(Data not shown)*. We quantified then HOTAIR level using qPCR oligonucleotides common to all HOTAIR forms. Full-length HOTAIR, truncated forms HOTΔPRC2 and HOTΔLSD1 were highly expressed with 48, 27 and 123 fold enrichment compared to HOTAIR level in HEK-Epi cells, respectively (Figure 34d). Two hypotheses can explain this difference in HOTAIR expression. First, the expression construct can be integrated to different genome regions that are more or less "open" and prompts for active transcription. Second, we can speculate that 5' and 3' regions of HOTAIR may contain important signals for the transcript turnover, and once deleted, affect its stability. Notably, all HOTAIR forms were at least 20 times more abundant in transduced HEK-Epi cells than in HEK-Mes cells.

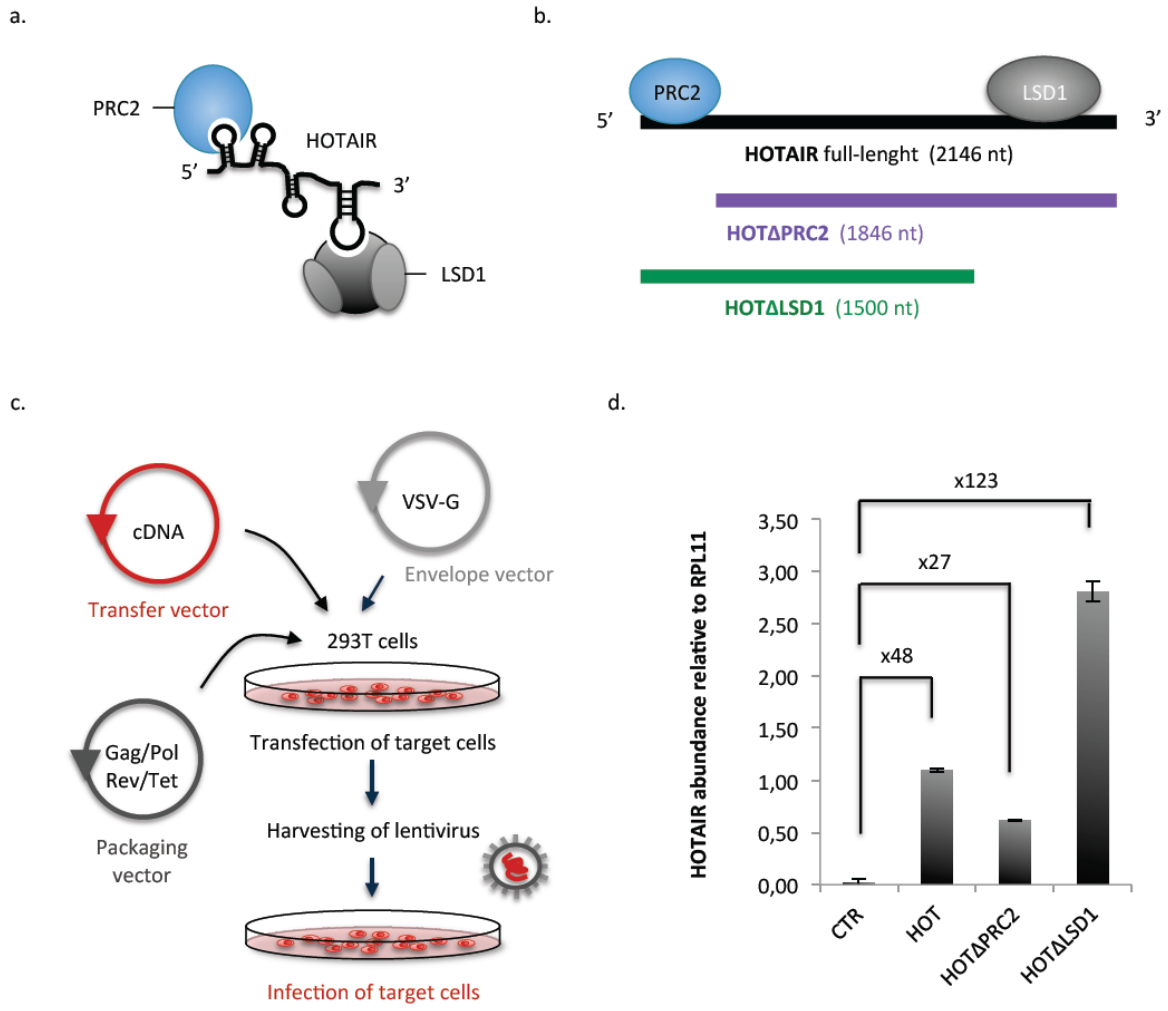

Figure 34 | HOTAIR over-expression in HEK-Epi cells. **a.** Schematic representation of HOTAIR secondary structures and interactions at its 5' and 3' extremities with PRC2 and LSD1/CoREST/REST complexes, respectively. **b.** Schematic representation of HOTAIR constructs used for HEK-Epi cells infection. HOTAIR full-length construct (HOT) correspond to HOTAIR annotated variant 1. HOTΔPRC2 and HOTΔLSD1 are HOTAIR constructs lacking PRC2 and LSD1 interaction domains, respectively. **c.** Infection procedure. HEK293T cells were transfected with envelope and packaging vectors, as well as a transfer vector carrying one HOTAIR cDNA construct, to produce lentivirus used to infect HEK-Epi cells. **d.** Oligo-specific RT-qPCR quantification of HOTAIR in infected HEK-Epi cells, using oligonucleotides detecting all HOTAIR forms. Fold-changes between CTR and HOTAIR forms are indicated.

ii. HEK-Epi cells overexpressing HOTAIR show increased proliferation, migratory capacity and invasiveness

Under microscopic evaluation, no major changes were visible in cells epithelial morphology upon HOTAIR full length and truncated forms over expression *(Data not*

shown). We then examined the effect of over expressions on cell proliferation, migratory capacity and invasiveness.

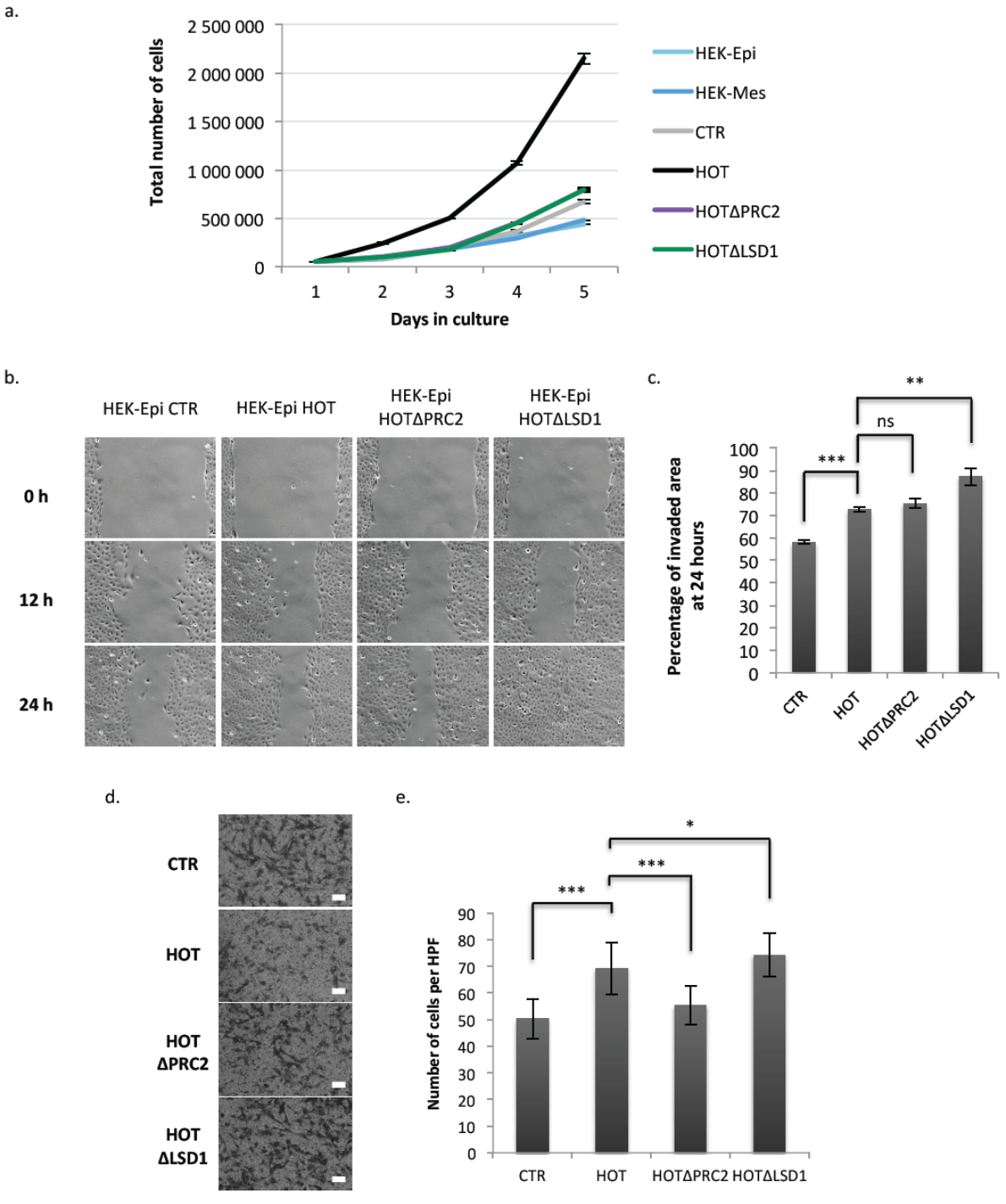

Figure 35 | Effects of HOTAIR over-expression on proliferation, motility and invasiveness of HEK-Epi cells. **a.** Cell proliferation rates of HEK-Epi, HEK-Mes, and HEK-Epi cells 10 population doublings after infection with GFP (CTR) and HOTAIR constructs. **b.** Wound healing assay, used to assess cells motility. Phase contrast images show wound recovery at 0, 12 and 24h post-scratch, in 10x magnification. **c.** Histograms represent the estimated percentage of the invaded area, taking first

picture at 0h as 20% invasion reference, and in a mean of 6 contrast phase images. **d.** Matrigel invasion assay was performed to assess cell invasion capacity. A mean of 10 phase contrast images was taken 72h after seeding 200.000 cells on membranes coated with matrigel, in three independent experiments. Scale bar: 100 µm. **e.** Counted numbers of invading cells per HPF. Error bars indicate standard deviation; * p<0.05, ** p<0,001, *** p<0,0001, ns, not significant.

Proliferation rate showed a strong increase in HEK-Epi cells over expressing full-length HOTAIR, whereas over expression of truncated transcripts led to a minor increase compared to the control. It is worth to note that the proliferation rate of cells over expressing GFP showed also a small increase compared to non-transduced HEK-Epi cells (Figure 35a). Trypan blue coloration of cells, performed in parallel, showed that cells viability was not affected by GFP or HOTAIR over expression *(data not shown)*.

We performed a wound-healing assay to assess cells migratory capacities. 24 hours after wound generation in the confluent monolayer, we observed that HEK-Epi cells over expressing GFP showed a marked increase in wound recovery compared to non-transduced HEK-Epi cells. Cells over expressing full-length HOTAIR showed a significant 1.2 increase in wound recovery (p-value < 0.0001) compared to CTR cells over expressing GFP. We then compared full-length and truncated HOTAIR forms. HOTΔPRC2 showed no significant difference in wound recovery, but we observed a significant 1.2 increase (p-value < 0.001) for HOTΔLSD1 cells, with complete wound recovery 24 hours post-scratch (Figure 35b and c).

Invasion assay showed that cells over expressing full-length HOTAIR exhibited a significant 1.4 increase in invasive rates, compared to control GFP (p-value < 0.0001). Cells over expressing HOTAIR truncated forms showed different changes in invasiveness, with no change for HOTΔPRC2, and a higher 1.5 increase for HOTΔLSD1, compared to the control (Figure 35d and e).

In conclusion, phenotypic studies showed that HOTAIR over expression in HEK-Epi cells affects cell proliferation rate, migratory capacities and invasiveness. We noticed that cells transduction *per se* induces slight increase in proliferation and migratory capacities of cells, seen in HEK-Epi cells infected with GFP expression plasmid and possibly due to the huge expression of an exogenous protein (Goto et al. 2003; Agbulut et al. 2007). Further increase of proliferation rate in cells expressing full-length HOTAIR suggests that HOTAIR

regulates proliferation of HEK-Epi cells, but this regulation seems to require both PRC2 and LSD1-interacting domains. We observed increased migration capacity and invasiveness of HEK-Epi cells upon both HOTAIR full-length and truncated forms over expression. Interestingly, the effects were even more marked in cells over expressing HOTAIR lacking LSD1-interacting domain. As hypothesized previously with HOTAIR depletion experiments, we can speculate that the effects of HOTAIR are dose-dependent, and that a higher abundance of HOTAIR in HOTΔLSD1 cells explains the increased migratory capacity and invasiveness. But this hypothesis also suggests that LSD1 binding has a minor or no role in HOTAIR function. A second hypothesis is in contrast that LSD1-interacting domain of HOTAIR plays a negative role in modulation of HOTAIR function.

iii. Effects of HOTAIR over expression on the EMT signature of HEK-Epi cells

We then examined the EMT signature of HEK-Epi cells over expressing full-length and truncated HOTAIR forms. We quantified by random-primed RT-qPCR the abundance of EMT markers mRNA in CTR, HOT, HOTΔPRC2 and HOTΔLSD1 cells.

Among tested epithelial markers, KRT19 and CTNNB1 showed equivalent expression in CTR, HOT and HOTΔLSD1 and a significant decrease in HOTΔPRC2 cells. TJP3 showed an equivalent decrease in all HOTAIR over expression conditions compared to the control. Among mesenchymal markers, ZEB2 showed comparable expression in CTR and HOT cells and a decreased level in HOTΔPRC2 and HOTΔLSD1. SNAI1 showed an increased level over expressed in HOT compared to CTR condition, but a decrease in presence of HOTAIR truncated forms, compared to the control. VIM expression seemed to be increased in HOT cells only. FN1 showed a comparable 1.6 decrease in HOT and HOTΔPRC2 cells compared to CTR cells, and more pronounced 4.0 decrease in HOTΔLSD1 (Figure 36a).

Western blot analysis of EMT markers revealed no significant variation of β-catenin and smooth muscle actin levels. Consistently with RT-qPCR results, VIM showed an increased level in cells over expressing full-length HOTAIR, but this effect was lost in cells over expressing HOTAIR lacking PRC2 and LSD1-binding domains (Figure 36b).

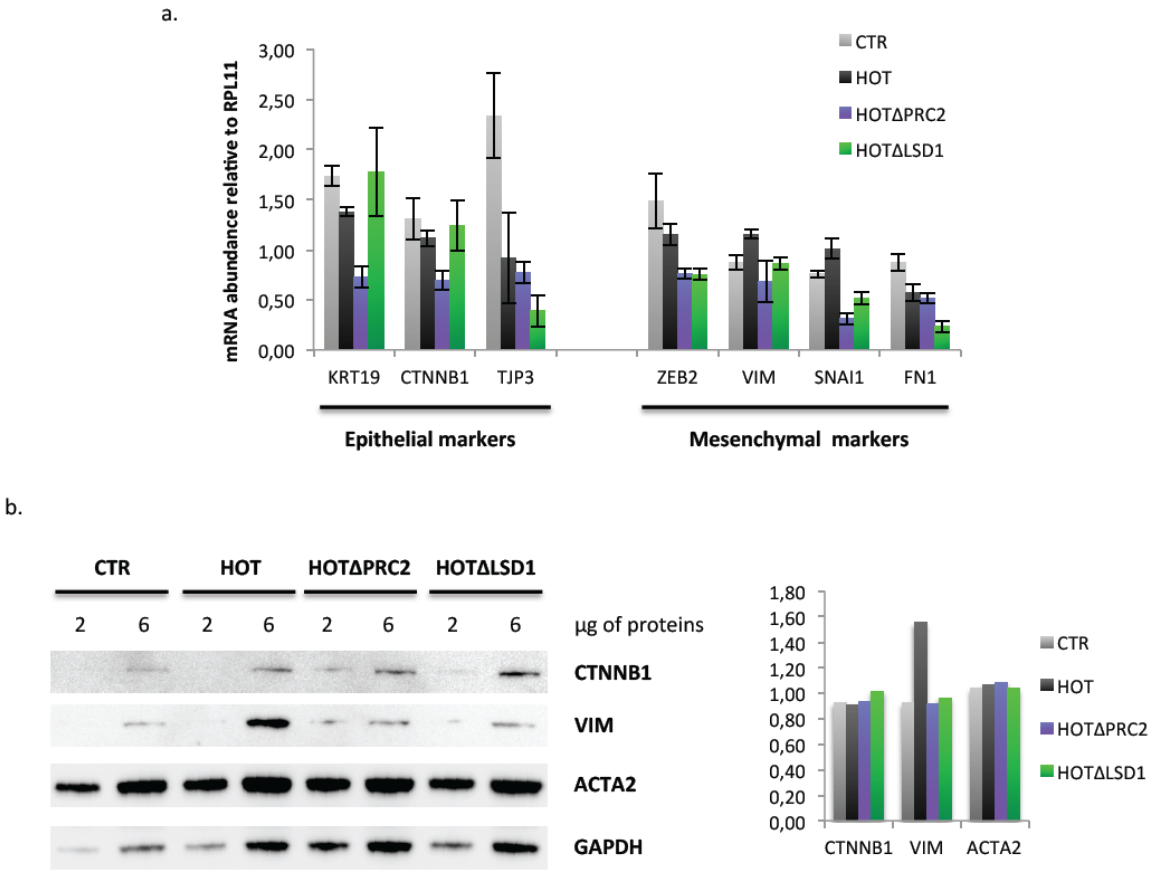

Figure 36 | EMT signature after HOTAIR over-expression. **a.** Quantification by random-primed RT-qPCR of EMT markers mRNA abundance in HEK-Epi cells 10 population doublings after infection with HOTAIR constructs. **b.** Western blot detection of EMT protein markers. Protein levels were quantified using ImageQuant software (GE Healthcare), and normalised by GAPDH as a loading control.

Altogether, these results showed that full-length HOTAIR overexpression in HEK-Epi cells promotes a decrease of the epithelial marker TJP3 at RNA level, and an increase of mesenchymal marker VIM. However, epithelial and mesenchymal markers signature is different from HEK-Mes cells, suggesting that the EMT program is not completely induced in HEK-Epi cells over expressing HOTAIR.

Interestingly, several tested EMT markers showed different changes between full-length and truncated HOTAIR forms. As an example, mesenchymal marker VIM was up regulated only in cells over expressing full-length HOTAIR, suggesting that both PRC2 and LSD1-interacting domains are necessary to establish a complete HOTAIR-mediated regulation of this gene. The epithelial marker TJP3 seemed to be down regulated in presence of full-length HOTAIR and both truncated forms. KRT19 and CTNNB1 mRNA levels were only decreased

in HOTΔPRC2, suggesting that their regulation takes place only when PRC2 is not bound to HOTAIR. With the exception of VIM and ZEB2, over expression of HOTAIR lacking LSD1-interacting domain showed more pronounced changes in mRNA levels. This result suggests the importance of LSD-binding domain in HOTAIR regulatory function.

iv. Effects of HOTAIR over expression on HEK-Epi cells transcriptome

To examine the transcriptome changes induced by HOTAIR full-length and truncated forms over expression, we analysed first by random-primed RT-qPCR the mRNA abundance of published HOTAIR target genes, identified by HOTAIR over expression in breast cancer cells MDA-MB-231 (Gupta et al. 2010).

LAMB3, published in these cells as induced by high HOTAIR level, showed no significant variation in HEK cell lines over expressing HOTAIR. Among the tested genes published as repressed in presence of high HOTAIR level, SIRT2, GATA2, PCDH10 and PCDHB5 showed indeed decreased levels in cells expressing full-length HOTAIR compared to CTR cells.

By comparison between full-length HOTAIR and truncated forms, we observed that GATA2 showed an increased level in HOTΔLSD1 compared to HOT and HOTΔPRC2 cells, suggesting that its regulation requires LSD1-interacting domain of HOTAIR. PCDH10 and PCDHB5 exhibited lower levels in HOTΔLSD1 compared to HOT and HOTΔPRC2 cells. This result suggests that down regulation of these genes is even more efficient with HOTAIR lacking LSD1-interacting domain, and supports the idea that LSD1-binding domain modulates HOTAIR function. HOXD10, JUB and BDNF genes showed no variation between HOT and CTR cells, but strong decreases in cells over expressing truncated forms (Figure 37).

Given these results, we can hypothesize that some of these genes, like GATA2, are direct targets of HOTAIR-mediated regulation. On the contrary, genes showing no variations between CTR and HOT cells, but a strong decrease in HOTΔPRC2 and HOTΔLSD1 cells, might reveal the indirect effects of a high exogenous expression of HOTAIR without the complete modulation of its function by both PRC2 and LSD1 binding.

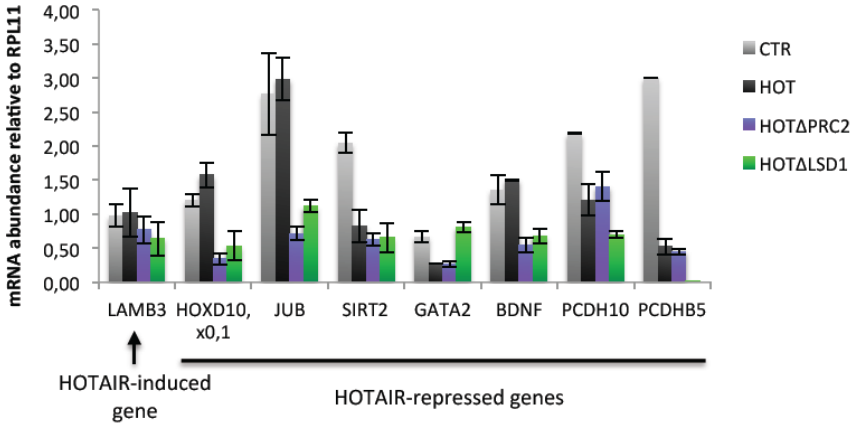

Figure 37 | Transcriptome changes after HOTAIR over-expression. Quantification by random-primed RT-qPCR of mRNAs identified by Gupta et al. as putative targets of HOTAIR-mediated regulation. Error bars represent standard deviation in three independent experiments.

- Transcriptome establishment by RNA-seq

We then analysed the global transcriptome of cells over expressing GFP, HOTAIR full-length and HOTAIR truncated forms using high-throughput RNA-sequencing approach. Starting from total RNA extracts of the four cell lines, we prepared duplicates of RNA-seq strand-specific cDNA libraries using TruSeq stranded Total RNA-seq Kit (Illumina). Samples were sequenced on an Illumina HiSeq 5500 sequencer at the ICGex platform of Institut Curie, and Zohra Saci performed bioinformatics analysis of RNA-seq data, following the pipeline presented previously (Figure 19). Number of total reads, mapped reads, duplicates and properly paired reads are presented in Table 8.

Table 8 | RNA-sequencing results for HEK-Epi cells over expressing HOTAIR forms.

Samples	Total number of reads	Mapped reads	% of mapped reads	Duplicates	% of duplicates	Properly paired	% of of properly paired
HOT1	167 931 347	156 168 128	93.00%	46 048 414	29%	131 678 062	84
HOT2	171 647 287	160 875 048	93.72%	46 918 340	29%	135 423 816	84
HOTΔPRC2-1	171 873 423	161 160 509	93.77%	43 504 001	27%	134 551 780	83
HOTΔPRC2-2	168 599 219	158 235 314	93.85%	42 633 943	27%	131 755 616	83
HOTΔLSD1-1	177 613 452	165 811 081	93.36%	46 947 232	28%	137 775 758	83
HOTΔLSD1-2	176 665 303	163 536 913	92.57%	46 723 369	29%	136 383 718	83
CTR-1	173 960 563	161 157 333	92.64%	42 824 236	27%	135 625 646	84
CTR-2	166 551 057	155 055 666	93.10%	41 702 225	27%	129 521 000	84

We first established density plot of mRNAs expressed in each cell line over expressing a HOTAIR construct compared to CTR cell line. We observed that HOTΔPRC2 exhibited the highest, and HOTΔLSD1 the lowest mRNA densities (Figure 38a). We then established heat maps representing mRNAs level of expression among the different conditions. HOT and HOTΔPRC2 samples were very closed to each other, with mRNA expressions showing equivalent variations. Interestingly, clustering of the sequenced samples separated HOTΔLSD1 from the other conditions, with a complete switch of mRNAs expression (Figure 38b). We examined also a density of reads mapped to the HOTAIR locus. RNA-seq data analysis revealed equivalent variations of HOTAIR expression levels between CTR and HOT, HOTΔPRC2, HOTΔLSD1 transduced HEK-Epi cells as measured by RT-qPCR, with 94, 52 and 217 fold enrichments, respectively (Figure 38c).

- Differential expressions of mRNAs, lncRNAs and as-lncRNAs

We performed a differential expression analysis expression between each HOTAIR construct and CTR using DEseq (Bioconductor) to identify the transcripts that were affected by HOTAIR over expression. We selected all mRNAs, lncRNAs and asRNAs that passed the combined criteria of adjusted p-value ≤ 0.05 and fold change ≥ 2 compared to the CTR.

We observed that 1 564 transcripts exhibited differential expression between HOTAIR full-length and CTR over expression. We noted that among these transcripts, 139 showed the same variation in expression in HEK-Mes cells, in which HOTAIR was naturally over expressed compared to HEK-Epi cells. DEseq performed between HOTΔPRC2 and CTR selected 660 transcripts. HOTΔLSD1 versus CTR DEseq analysis showed the major effect of the over expression of HOTAIR lacking LSD1-interacting domain on the transcriptome, with 4 951 transcripts deregulated. Consistently with the role of HOTAIR in gene silencing, the majority of differentially expressed transcripts were down regulated upon HOTAIR over expression: 908 down regulated transcripts among 1 564 for HOT cells, 434 among 660 for HOTΔPRC2, and 2 909 among 4 951 for HOTΔLSD1. We then examine whether the differentially expressed genes were common between the cells over expressing full length and truncated HOTAIR forms. We found 317 common genes between HOT and HOTΔLSD1, and 28 common genes between HOT and HOTΔPRC2. 158 genes were commonly deregulated in HOT, HOTΔLSD1 and HOTΔPRC2 (Figure 39).

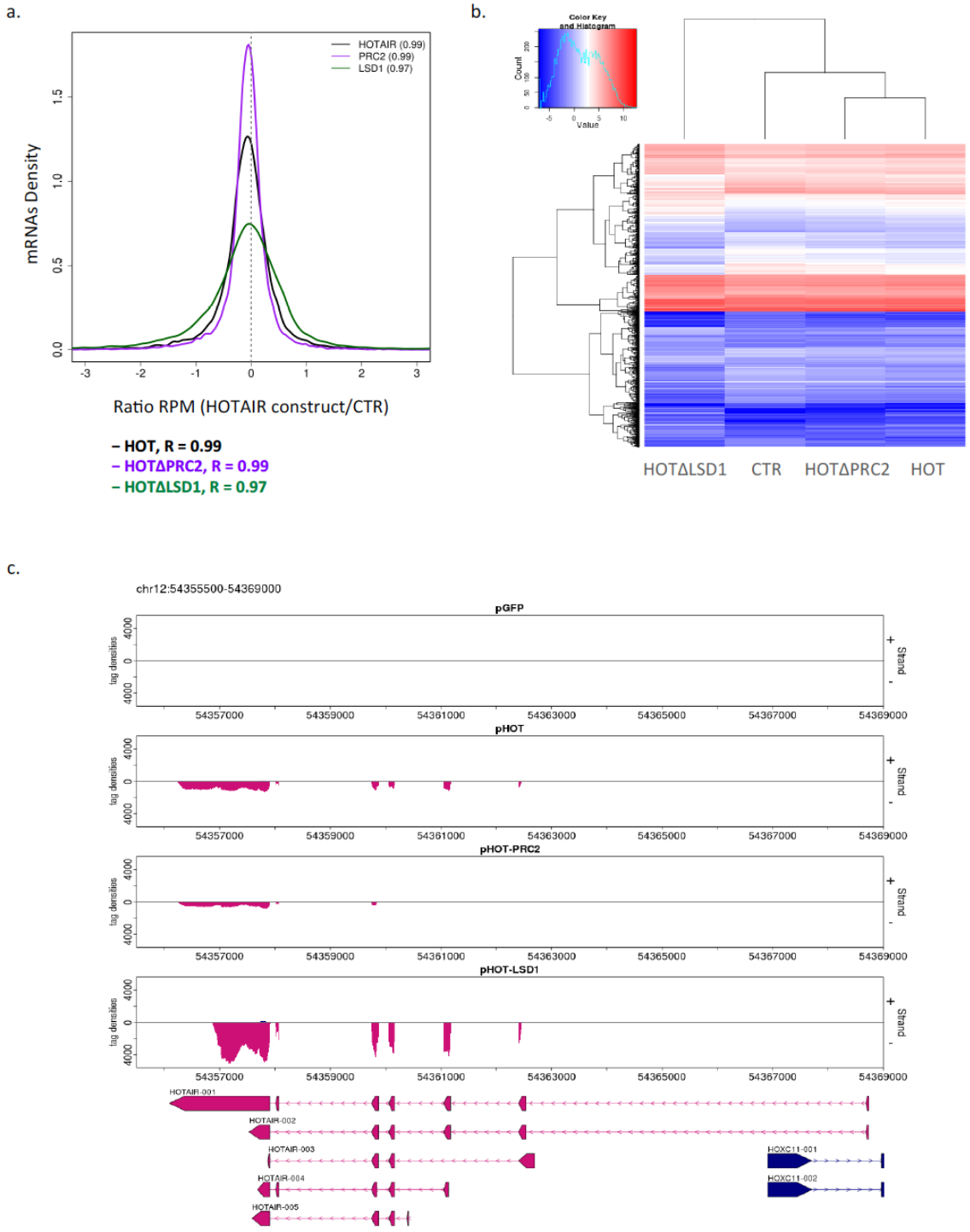

Figure 38 | Genome-wide analysis of HOTAIR over-expression effects in HEK-Epi cells. **a.** mRNAs density distribution in HEK-Epi cells infected with HOTAIR constructs versus HEK-Epi cells infected with CTR plasmid. RPM: Reads Per Million. **b.** Heatmaps of deregulated mRNAs. **c.** Ving snapshot of mapped reads density on HOTAIR locus, in one representative RNA-seq replicate of cells infected with control plasmid (CTR), full length HOTAIR (HOT), HOTΔPRC2 and HOTΔPRC2LSD1.

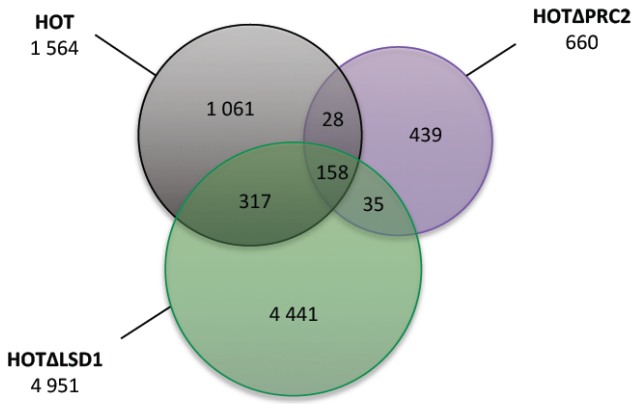

Figure 39 | Numbers of differentially expressed transcripts identified by DEseq analysis. Venn diagram representation of significant specific and common differentially expressed transcripts in HEK-Epi cells over expressing full length HOTAIR, HOTΔPRC2 and HOTΔLSD1, compared to HEK-Epi cells over expressing GFP as a control.

Genes identified as differentially expressed upon HOTAIR over expression could be direct or indirect targets of HOTAIR-mediated regulation. Comparison between datasets generated by DEseq analysis allowed the establishment of 5 distinct classes of potential target genes. **(1)** 1 061 genes deregulated only upon HOTAIR full-length overexpression could be direct or indirect targets of a HOTAIR-mediated regulation involving both LSD1 and PRC2-interacting domains, and possibly interaction between the two protein complexes. **(2)** 317 genes commonly deregulated in HOT and HOTΔPRC2 samples could be potential targets of regulation mediated by HOTAIR via LSD1 binding. **(3)** In the same way, 317 commonly deregulated genes in HOT and HOTΔLSD1 samples could be potential targets of regulation mediated by HOTAIR via PRC2 binding. **(4)** 158 genes were commonly deregulated in HOT, HOTΔPRC2 and HOTΔLSD1 samples. This result suggests that potential HOTAIR target genes could be regulated via a mechanism not involving PRC2 and LSD1 interactions. **(5)** We found 4 441 and 439 genes deregulated only in cells overexpressing HOTΔLSD1 and HOTΔPRC2, respectively. Effects observed on expression of these genes could be due to an unspecific targeting by high level of exogenous HOTAIR, which cannot be well controlled by PRC2 and LSD1 interactions.

The very high proportion of transcripts showing differential expression only in HOTΔLSD1 cells suggests that LSD1 binding is crucial for gene regulation mediated by HOTAIR, with two hypotheses. First, we could speculate that LSD1 binding modulates HOTAIR function, impeding the binding or interacting with PRC2 to repress HOTAIR activity. We can also think that LSD1 binding to HOTAIR regulates major repressors or activators of transcription, inducing global up or down regulation of transcription.

- GO terms and KEGG Pathways in HEK-Epi cells overexpressing HOTAIR

We then examined biological process annotation of mRNAs differentially expressed in HEK-Epi cells over expressing HOTAIR forms compared to CTR. We selected the significant (p-values ≤ 0.05) biological processes. First, our analysis revealed that HEK-Epi cells over expressing full-length HOTAIR displayed significant changes in expression of genes involved in cell-cell signalling, regulation of cell proliferation and migration, positive regulation of cell cycle, but also inflammatory response (Figure 40a), consistently with observed changes in cells phenotype. Interestingly, in the extended GO terms list, we also found the Wnt signalling pathway. Among identified KEGG pathways, we found as the most relevant those related with the MAPK signalling pathway, regulation of actin cytoskeleton and cell adhesion molecules. We also found an enrichment of cancer genes, implicated in basal cell carcinoma, endometrial cancer, glioma and melanoma (Figure 40b) (Supplementary Table S5).

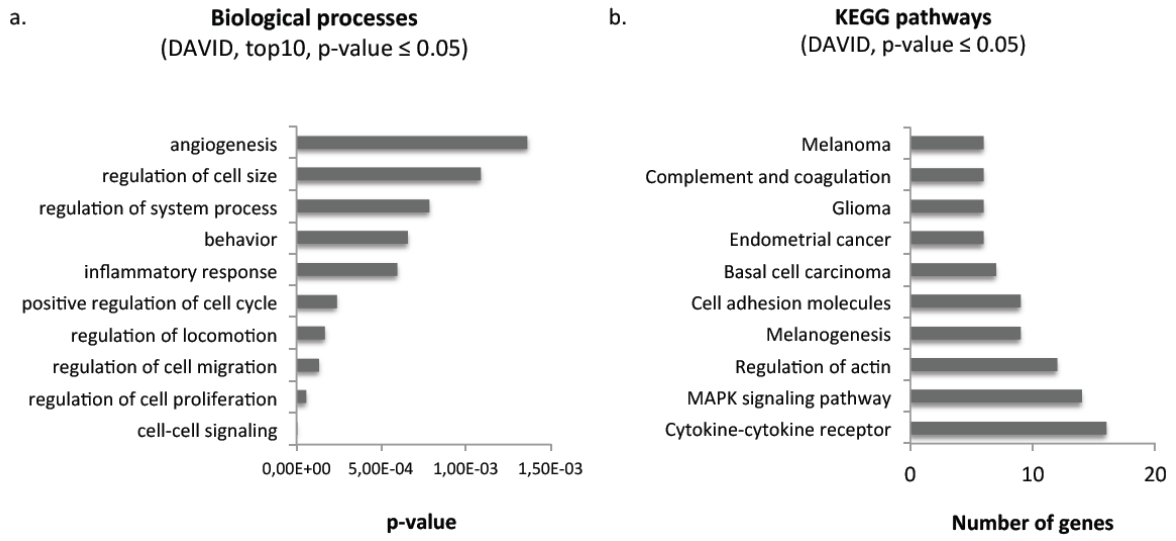

Figure 40 | Biological processes associated with differentially expressed mRNAs between HEK-Epi over expressing full-length HOTAIR and GFP. **a.** Most enriched GO terms of biological processes identified by DAVID (p-value ≤ 0.05). **b.** KEGG pathways identified by DAVID analysis (p-value ≤ 0.05).

To better understand the observed effects of HOTAIR truncated forms over expression in HEK-Epi cells, we then analysed biological process annotation of differentially expressed mRNAs specific to HOTΔLSD1 and HOTΔPRC2 cells. A very high number of GO terms found as relevant for HOTΔLSD1 specific genes revealed the importance and heterogeneity of the effects induced by HOTAIR lacking LSD1-interacting domain over expression. Among the most significant GO terms, we found collagen fibril organization, regulation of cell proliferation, but also cell adhesion and response to wounding. We found as the most relevant KEGG pathways those related with focal adhesion, purine metabolism, p53 and TGF-Beta signalling pathway. Differentially expressed mRNAs in HEK-Epi cells over expressing HOTΔPRC2 were found to be significantly related with only 4 GO terms: negative regulation of apoptotic process, lymphocyte chemotaxis, cell activation and positive regulation of interleukin-1 secretion. No KEGG pathway was found as significantly associated (Supplementary Table S6). These results are consistent with the observed changes in cells phenotype. However, the high number and variety of GO terms relevant for HOTΔLSD1 deregulated genes suggest that HOTAIR lacking LSD1-interacting domain has a very heterogeneous regulatory effect on protein-coding transcriptome.

- Comparison with published list of HOTAIR-target genes and EMT markers

Two types of analysis were performed to examine whether the genes we identified were already shown as regulated by HOTAIR or linked to the EMT process.

First, we compared our dataset of differentially expressed transcripts between HOT and CTR cells, to the list of HOTAIR target genes established in MDA-MB-231 cells over expressing HOTAIR (Gupta et al. 2010). A total of 30 genes, 13 down regulated and 17 up regulated, were common between the two lists (Supplementary Table S7). This low number reflected the high cell-specificity of genes targeted by HOTAIR, but also validated the possibility to isolate a core list of HOTAIR target genes, common to different cell lines.

Then, we crossed our list of differentially expressed transcripts between HOT and CTR cells with the list of 365 EMT-associated genes (Gröger et al. 2012). 29 down regulated genes and 58 up regulated genes were common between the two lists (Supplementary Table S7). This result suggests that the expression of several EMT markers could be regulated by

HOTAIR. DAVID analysis revealed that the significantly (p-value ≤ 0.05) enriched GO terms associated to these genes are the regulation of cell growth, proliferation, adhesion and motion. This result supports the idea that HOTAIR is not directly involved in the induction of the EMT program, but regulates cell proliferation, acquisition of migratory capacity and invasive potential.

- Biological validation of RNA-seq data generated from HEK-Epi cells over expressing HOTAIR forms

To examine whether our RNA-seq analysis had a biological significance, we selected several differentially expressed mRNAs showing high expression level and significant fold-change between HEK-Epi cells over expressing GFP and HOTAIR constructs. We quantified by random-primed RT-qPCR the abundance of mRNAs in HEK-Epi cells over expressing GFP or HOTAIR constructs.

We first selected several mRNAs showing down or up regulation upon HOTAIR over expression in our RNA-seq dataset. The nuclear protein transcriptional regulator 1 (NUPR1), that has been shown to exhibit a protective role against metastasis in pancreatic cancer (Cano et al. 2012), and the tribbles pseudokinase 3 (TRIB3), a critical protein in the regulation of cell cycle arrest (Yu et al. 2013), were indeed down regulated in HOT cells compared to CTR cells. The protein disulphide isomerase PDIA4, involved in resistance to cisplatin-induced cell death in lung adenocarcinoma (Tufo et al. 2014), and Heat-shock protein 90 (HSP90B1), a molecular chaperone crucial for function of many proteins and now a potential target for cancer therapy (Garcia-Carbonero et al. 2013), were confirmed as up regulated in HOT cells compared to CTR cells (Figure 41a).

We then selected several genes down regulated upon over expression of full-length HOTAIR, and up regulation in either HOTΔLSD1 or HOTΔPRC2, suggesting that their down regulation was mediated by LSD1 or PRC2 binding to HOTAIR transcript, respectively. G protein-coupled receptor 1 (GPR1) showed indeed a down regulation in HOT and HOTΔPRC2 cells compared to CTR, which was not observed anymore in HOTΔLSD1 cells. This result confirmed that GPR1 could be specifically regulated by LSD1 interacting with HOTAIR. On the contrary, inhibin beta E (INHBE) and cadherin-related family member 1 (CDHR1) showed down regulation in HOT and HOTΔLSD1 cells compared to CTR, but not

in HOTΔPRC2 cells. This suggests that specific PRC2-binding to HOTAIR transcript regulates the expression of these genes (Figure 41b).

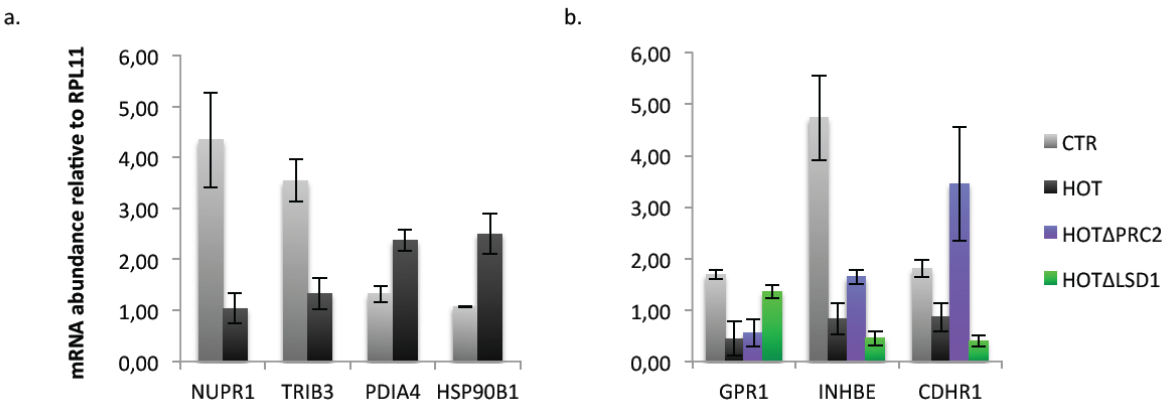

Figure 41 | Quantification by random-primed RT-qPCR of mRNAs identified as deregulated in HEK-Epi cells overexpressing **a.** HOTAIR full-length and **b.** HOTAIR full-length and truncated constructs, compared to HEK-Epi cells over expressing GFP as a control. Error bars represent standard deviation between three technical replicates.

Altogether, these results showed that RNA-sequencing allowed the identification of transcripts differentially regulated in HEK-Epi cells over expressing HOTAIR, as well as classes of transcripts specifically regulated by PRC2 and LSD1 via interaction with HOTAIR. Focusing on mRNAs, we showed that these transcripts are mostly involved in cell proliferation, migration and adhesion. Major and heterogeneous transcriptome changes induced by over expression of HOTAIR lacking LSD1-interacting domain confirmed its importance in HOTAIR-mediated regulation.

B. Identification and characterization of novel antisense lncRNAs in human cells

Among the emerging classes of lncRNAs, long intervening non-coding RNAs have been extensively described as being involved in *trans*-regulation, mostly at the epigenetic level, of genes important in cell differentiation, development and cancer (Rinn & Chang 2012). But other less-studied classes of lncRNAs exhibit interesting features and potential as therapeutic targets, such as Natural Antisense Transcripts (NATs) (Li & Chen 2013) (see Introduction). These transcripts have been shown to form RNA/RNA hybrids, triggering

RNAi machinery, but also to perturb sense gene expression. However, despite their regulatory importance, antisense transcripts have been poorly studied and no systematic study has yet addressed their comprehensive functional description in eukaryotes and, particularly, in humans.

In this part, I will present preliminary results of two different studies we started in parallel with the analysis of lncRNAs role in EMT. First, by revisiting publicly available RNA-seq datasets with our specific bioinformatics pipeline, we unveiled the existence of several novel, non-annotated as-lncRNAs, including a specific class of antisense transcripts restricted to introns of protein-coding genes. Second, previous work performed in our lab revealed the existence of cryptic non-coding RNAs sensitive to degradation by XRN1 5'-3' exoribonuclease in yeast *Saccharomyces cerevisiae* (XRN1-sensitive Unstable Transcripts, XUTs). By analogy with this study, and given the high degree of XRN1 conservation among eukaryotes, we examined the existence of such transcripts in human cells.

1. INATs, a novel class of Intronic Antisense Transcripts in human cells

a. Identification of novel as-lncRNAs from ENCODE available RNA-seq datasets

Using the previously described bioinformatics pipeline, we revisited 14 strand-specific RNA-seq datasets published by the ENCODE project (Djebali et al. 2012) (Table 9). All unique transcripts present in at least two replicates for each cell line with RPKM ≥ 1 were considered for further analysis. We then filtered all transcripts < 100 nucleotides in length, and applied PhyloCSF1 filter (score > 100) to all antisense transcripts to extract a set of antisense non-coding transcripts. Using this method, we successfully detected an average of 1 064 already annotated antisense transcripts per cell line, consistently with ENCODE published results. In addition, we identified between 1 721 and 6 128 novel as-lncRNAs per cell line. It is worth to note that higher numbers of novel transcripts were identified in cancer cells lines compared to normal cells, with average numbers of 3 484 and 2 187 transcripts per cell line, respectively. We examined then the localisation of these novel as-lncRNAs on the genome. Interestingly, 5 to 22% of identified antisense transcripts, which length varied between 200 and 4 000 nucleotides, were specifically localized in introns of protein-coding genes, not overlapping with any annotated exon. In addition, we examined the position in the

genome of novel as-lncRNAs identified by RNA-seq in HEK-Epi and HEK-Mes cells, and found 914 and 1 794 antisense intronic transcripts, respectively (Table 10). We focused our study on these Intronic Antisense Transcripts (INATs).

Table 9 | ENCODE dataset used in this study.

Cell type name	Description	Lineage	Tissue	Karyotype	Number of replicates	Type of RNA	Total reads	Uniquely mapped
A549	alveolar basal epithelial cells	endoderm	lung epithelium	Lung carcinoma	2	polyA+	190 108 518	143 596 236
AG04450	fetal lung fibroblast	endoderm	lung	Normal	2	polyA+	235 824 956	174 327 823
BJ	foreskin skin fibroblasts	-	skin	Normal	2	polyA+	214 852 470	162 102 709
GM12878	B lymphoblastoid	mesoderm	blood	Normal	2	polyA+	235 752 640	167 856 226
HepG2	hepatoblastoma cells	endoderm	liver	Liver carcinoma	2	polyA+	248 345 298	188 658 976
HSMM	skeletal muscle myoblasts	mesoderm	muscle	Normal	2	polyA+	230 526 592	180 638 612
HUVEC	umbilical vein endothelial cells	mesoderm	blood vessel	Normal	2	polyA+	174 436 896	136 266 915
K562	erythroleukaemia cells	mesoderm	blood	Myelogenous leukemia	2	polyA+	227 177 516	159 808 327
MCF7	mamary gland adenocarcinoma	ectoderm	breast	Adenocarcinoma	2	polyA+	256 356 220	188 551 970
NHEK	epidermal keratinocytes	ectoderm	skin	Normal	2	polyA+	276 806 018	207 373 122
NHLF	lung fibroblast	endoderm	lung	Normal	2	polyA+	264 182 940	211 153 950
SK-N-SH	neuroblastoma differentiated with retinoic acid	ectoderm	brain	Neuroblastoma	2	polyA+	280 724 734	214 590 184
HeLa-S3	S3 cervical carcinoma cells	ectoderm	cervix	Cervical cancer	2	polyA+	242 236 968	176 440 983
ESC	H1 Embryonic Stem Cells	inner cell mass	ESC	Normal	2	polyA+	250 790 392	181 112 441

Table 10 | Identified as-lncRNAs in ENCODE cell lines.

Cell line	as-lncRNAs, Total	Annotated	Novel	INATs
HeLa-S3	4122	1102	3020	715
SK-N-SH	7442	1314	6128	407
MCF7	4639	1269	3370	636
K562	3704	1020	2684	633
HepG2	4378	1132	3246	555
A549	3516	1061	2455	567
ESC	4353	1198	3155	555
NHLF	2664	943	1721	443
NHEK	2566	967	1599	565
HUVEC	3060	850	2210	441
HSMM	2835	1015	1820	438
GM12878	4284	1111	3173	684
BJ	2914	972	1942	460
AG04450	2839	964	1875	494

Additional bioinformatics analysis revealed a proportion of INATs overlapping with annotated pseudogenes, as well as INATs constituted of repeated sequences. For further experiments, we selected several candidates among INATs without any annotation or repeats.

Before any characterization of these transcripts, we tested the biological relevance of RNA-seq analysis. We first examined the existence of INATs by Northern blot experiments. Total RNAs from HeLa and MCF7 cells were loaded on a 1% agarose gel and transferred onto a nitrocellulose membrane. We probed the membrane with P-32 labelled oligonucleotides complementary to the sequence of two different INATs, antisense to one intron of DMD and DCAF6 genes. For both DMD and DCAF6 INATs, we observed signals at a size corresponding to assembly transcripts by RNA-seq, 414 and 1 014 nucleotides, respectively (Figure 42a). Oligo-specific RT-qPCR experiments allowed quantification of INATs from DCAF6, DMD, PCNP and BTRC loci, in HeLa and MCF7 cells (Figure 42b). These results confirmed that INATs identified by RNA-seq analysis are not artefacts of sequencing or bioinformatics analysis, but real transcripts detectable in cells.

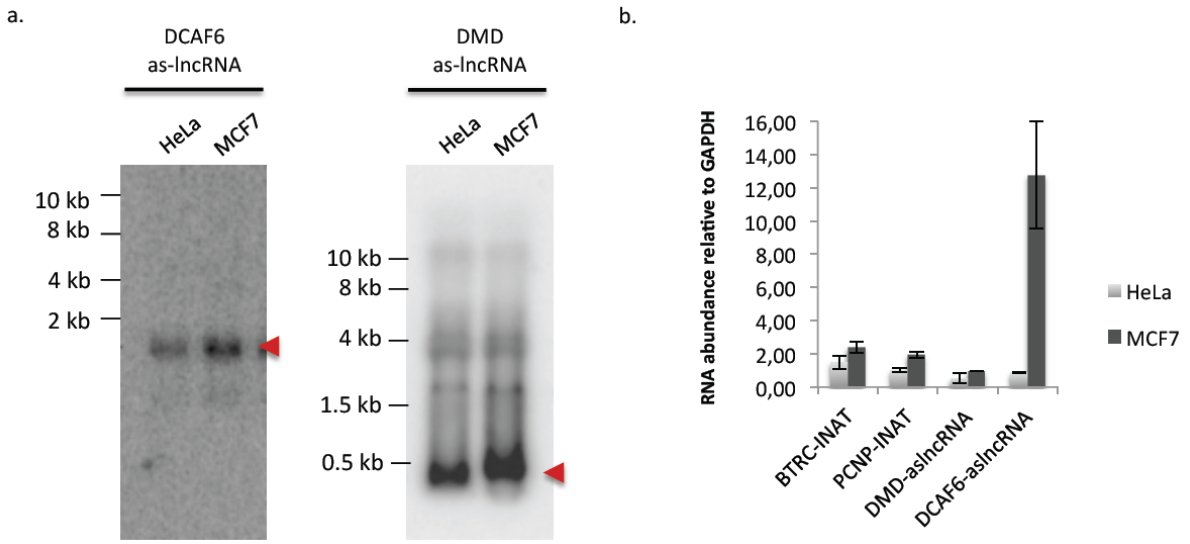

Figure 42 | INATs detection in HeLa and MCF7 cells. a. Northern blot detection of DCAF6 as-lncRNA and DMD as-lncRNA in HeLa and MCF7 cells. b. Oligo-specific RT-qPCR quantification of BTRC-INAT, PCNP-INAT, DCAF6 as-lncRNA and DMD as-lncRNAs in HeLa and MCF7 total RNA extracts. Error bars represent standard deviation in three independant experiments.

b. INATs biogenesis

INATs were detected in ENCODE RNA-seq datasets generated from PolyA+ RNA extracts, suggesting that these transcripts are polyadenylated. To confirm this hypothesis, we performed PolyA+ RNAs purification from HeLa cells total RNAs and measured the abundance of several mRNAs, lncRNAs and INATs in both polyA+ and polyA- isolated fractions. GAPDH, RPL11, MALAT1 RNAs, known polyadenylated transcripts, and 7SL, non-polyadenylated RNA, were used to control the efficiency of PolyA+ RNA purification. INATs from PCNP, DCAF6 and DMD loci were highly enriched in PolyA+ RNAs fraction (Figure 43).

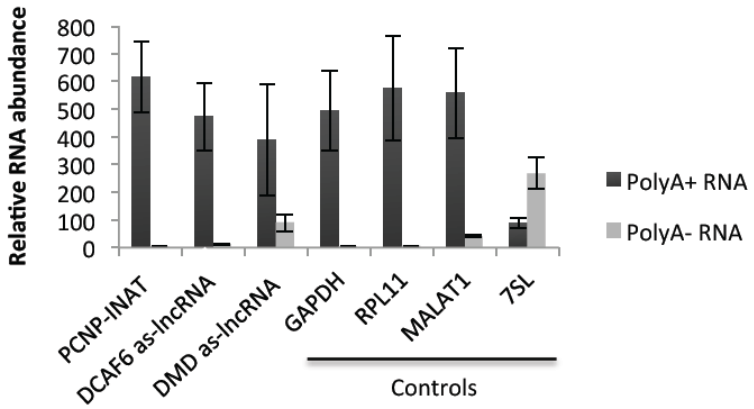

Figure 43 | INATs are polyadenylated transcripts. Oligo-specific RT-qPCR quantification of mRNAs and INATs in polyA+ and polyA- fractions of HeLa RNA extracts.

Polyadenylation is a classical feature of RNA Polymerase II (RNAPII) transcripts. To further explore whether INATs share common features with RNAPII-transcribed genes, we analysed chromatin signatures at putative INAT transcription start sites (TSS). We selected histone H3K4me3 and H3K27ac as representative marks of actively RNAPII-transcribed genes (Kouzarides 2007). Using ENCODE ChIP-seq data, we performed genome-wide meta-analysis of H3K4me3 and K27ac distribution along and around INAT's metagene in 14 cell lines. As expected, INATs loci show non-random peaks of H3K4me3 and H3K27ac at their putative TSS, similar to RNAPII promoter patterns (Figure 44a and b). We confirmed histone H3K4me3 and RNAPII patterns by gene-specific ChIP experiments for DCAF6-INAT, localized in intron 17 of DCAF6 gene, in HeLa cells. We measured by RT-qPCR the level of

immunoprecipitated DNA at the promoter region of DCAF6, and all along the intron where the INAT was mapped. U1, CDK4 and GAPDH were used as controls of active transcription by RNAPII, and U6 as a negative control. We observed a high level of RNAPII detected at the localization of the INAT inside the intron (Figure 44c), as well as enrichment of H3K4me3 at the putative promoter of DCAF6-INAT (Figure 44d), both correlated with active transcription.

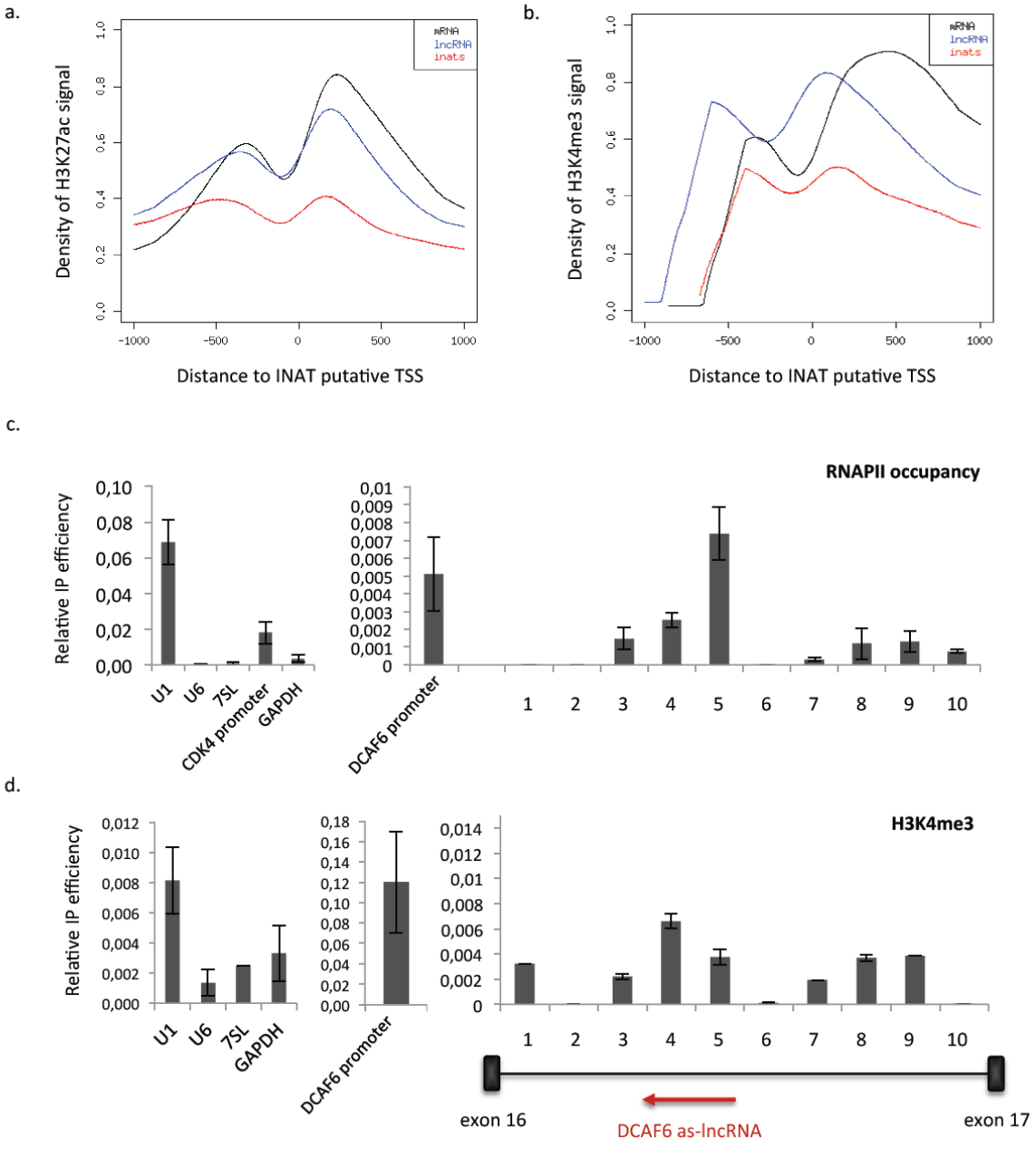

Figure 44 | Chromatin signature and RNAPII occupancy at INATs localization. **a.** Density of H3K27ac signal at putative INATs TSS, determined from ENCODE ChIP-seq dataset. **b.** Density of H3K4me3 signal at putative INATs TSS. **c.** RNAPII occupancy and **d.** H3K4me3 measured by ChIP experiments, all along DCAF6 intron exhibiting antisense transcription.

Altogether, these preliminary results suggest that INATs are discrete RNAPII transcription units strictly restricted to introns of protein-coding gene. We then addressed several issues to get insights into INATs biological function.

c. Insights into INATs biological function

• INATs intronic localization

We found some interesting features by analysing the genes exhibiting intronic antisense transcription. As an example, DMD-INAT is localized in an intron of DMD gene, extensively studied for its complexity and implication of its mutations in Duchenne (DMD) and Becker (BMD) Muscular Dystrophies (Muntoni et al. 2003). DMD, the largest gene of the human genome, harbours a variety of sense and antisense lncRNAs. Gain-of-function experiments revealed that DMD lncRNAs contributes to the orchestration and homeostasis of the muscle dystrophin expression pattern by either selective targeting or down modulating the dystrophin promoter transcriptional activity (Bovolenta et al. 2012). We can speculate that DMD-INAT identified by RNA-seq analysis is a novel lncRNA involved in this regulation. On the other hand, DCAF6-INAT, antisense to DDB1 and CUL4 associated factor 6 gene, is localized in the intron downstream an exon annotated as alternative in DCAF6 mRNA variants, suggesting that its transcription might be involved in the regulation of alternative splicing.

• INATs cell-specific expression

We examined the expression of INATs across the 14 analysed cell lines from the ENCODE project. As for lncRNAs, we observed that a high proportion of INATs were expressed in only one cell line (39 and 45%, respectively). On the contrary, a majority of mRNAs (48%) seemed to be expressed in all tested cell lines (Figure 45). This result suggests that in the same way as lncRNAs, and contrary to mRNAs, INATs exhibit a cell-specific expression pattern.

In addition to this observation, it is worth to note that INATs seemed to be differentially expressed before and after the EMT. Indeed, oligo-specific RT-qPCR quantification of several INATs showed that BTRC, DMD and PCNP-INATs are down regulated, and DCAF6-INAT up regulated, in HEK-Mes compared to HEK-Epi cells (Figure 46). Moreover,

the high number of INATs expressed in HEK-Mes compared to HEK-Epi cells (1 794 and 914, respectively), suggests that they might be not only cell-type specific but also very sensitive to changes in cell identity.

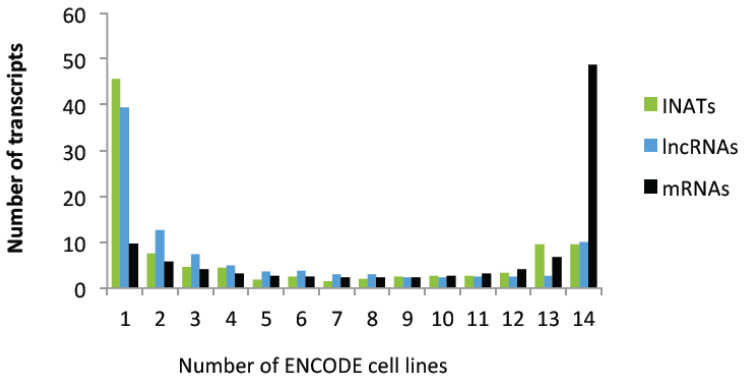

Figure 45 | INATs are cell-specific transcripts. Repartition of mRNAs, lncRNAs and INATs expressed in one or several cell lines from ENCODE RNA-seq datasets.

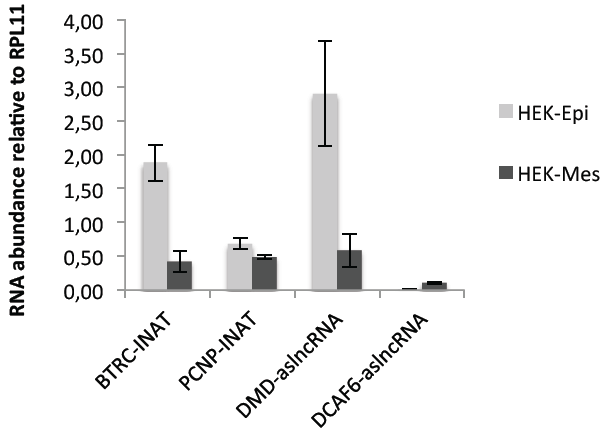

Figure 46 | INATs are differentially expressed in HEK-Epi and HEK-Mes cells. Oligo-specific RT-qPCR quantification of BTRC-INAT, PCNP-INAT, DMD as-lncRNA and DCAF6 as-lncRNA in HEK-Epi and HEK-Mes total RNA extracts. Error bars represent standard deviation in three independent experiments.

- INATs cellular localization

We then analysed cellular localization of INATs using RNA-seq data generated by the ENCODE project from fractionated HeLa cells. We identified two subclasses of INATs: a cytoplasmic subclass, showing very heterogeneous expression levels, and a nuclear subclass (Figure 47a). This result was confirmed for several INATs, by oligo-specific RT-qPCR

quantification in RNAs extracted from fractionated HeLa cells. We used GAPDH mRNA, localized in cells cytoplasm, and MALAT1 lncRNA, known to be retained in cell nucleus (Hutchinson et al. 2007; Clemson et al. 2010), to control the efficiency of cells fractionation. Results showed that DCAF6 and PCNP-INATs are enriched in the nuclear fraction, whereas DMD and RAB3GAP2-INATs are localized in the cytoplasm of HeLa cells (Figure R47b). Together, these results argue to highly controlled trafficking of INATs, as well as to a regulatory potency of these ncRNAs in various cellular processes.

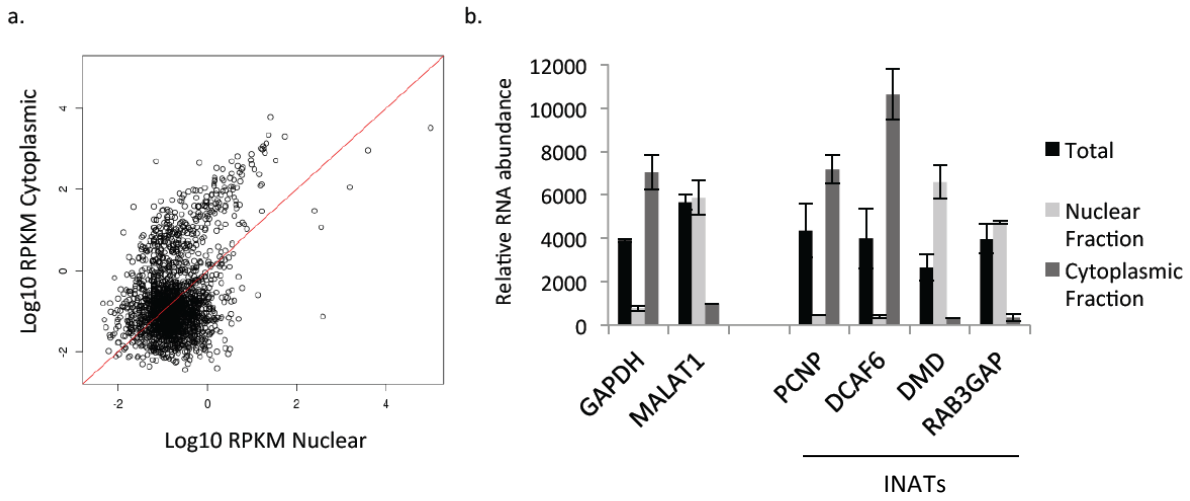

Figure 47 | INATs are localized in the cytoplasm and in the nucleus of HeLa cells. **a.** INATs cellular localization analysis in ENCODE RNA-seq datasets from fractionated HeLa cells. **b.** Oligo-specific RT-qPCR quantification of INATs in RNA extracts from HeLa nucleus and cytoplasm.

- INATs loss-of-function study

To examine whether INATs play a role in the regulation of gene expression or alternative splicing, we performed DCAF6-INAT depletion using ASOs, in HeLa cells. We designed four antisense oligonucleotides all along DCAF6 INAT (Figure 48a), and tested several concentrations and time of treatment to optimize DCAF6-INAT depletion. Finally, we treated the cells using 5 µM ASO, and extracted total RNAs after 72 hours. We used an ASO against GFP as a negative control. Oligo-specific RT-qPCR measure showed that two among four tested ASOs induced a very efficient (90%) depletion of DCAF6-INAT (Figure R48b).

After DCAF6-INAT depletion, and following the idea that DCAF6-INAT could be involved in the regulation of exon 17 alternative splicing, we measured by oligo-specific RT-

qPCR the abundance of DCAC6 mRNA using primer pairs amplifying exon junctions. Our aim was to amplify junctions between exons 15 and 16, as a control, and junctions between exons 16 and 17, 17 and 18, 16 and 18, to test their enrichment or loss upon DCAF6-INAT depletion. Interestingly, no junction between exon 17 and other exon could be detected, possibly indicating that this exon was not included in DCAF6 mRNA at all in HeLa cells, or suggesting that this exon is only an artefact of the annotation used as a reference. But our results showed that the level of mRNA detected by either 15-16 or 16-18 primers was decreased in cells after DCAF6-INAT depletion (Figure 48c). This preliminary observation suggests that DCAF6-INAT could be involved in positive regulation of DCAF6 mRNA transcription or stability.

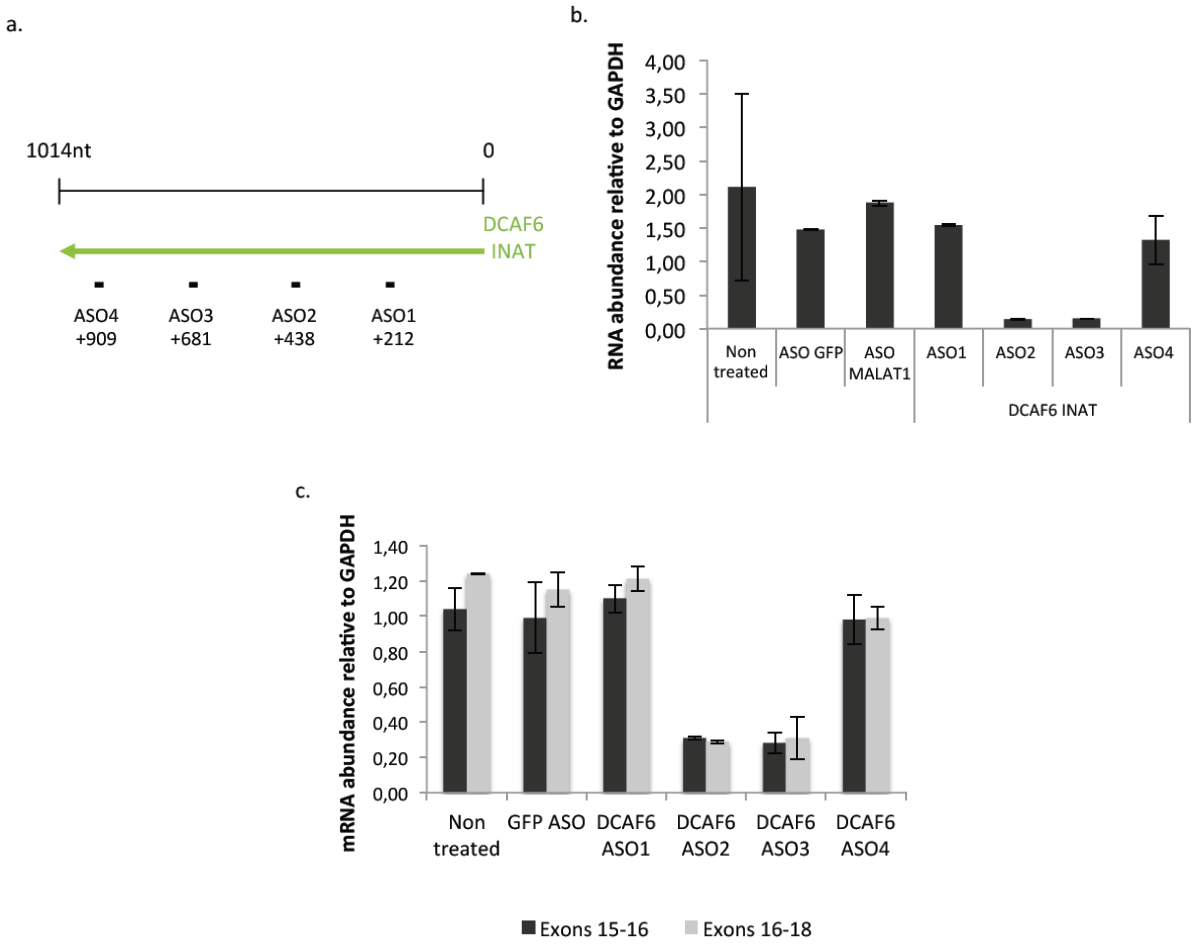

Figure 48 | DCAF6 INAT depletion in HeLa cells induces a decrease in DCAF6 mRNA level. **a.** Localization of tested ASO on DCAF6 INAT locus. **b.** Oligo-specific RT-qPCR quantification of DCAF6 INAT level in HeLa cells after 72 hours of treatment with ASOs. **c.** Oligo-specific RT-qPCR quantification of DCAF6 mRNA exon junctions in HeLa cells 72 hours after ASO treatment.

Altogether, these preliminary results showed that revisiting published RNA-seq datasets using a specific bioinformatics pipeline allowed to reveal the existence of several thousands of antisense transcripts in normal and cancer cell lines, drawing a more complete human antisense lncRNAs catalogue. Among novel as-lncRNAs, we identified a class of transcripts restricted to introns of protein-coding genes, transcribed by RNAPII, and that seems to be highly cell-specific and sensitive to changes in cell identity. Further validation and a complete study of INATs are necessary to confirm our results, but first insights into INATs function suggests that they could play a role in gene expression regulation.

2. hXUTS, a novel class of cryptic non-coding transcripts in human cells?

Recently, our laboratory identified by the use of strand-specific RNA-sequencing a novel class of 1 658 lncRNAs in *Saccharomyces cerevisiae*. These transcripts, called XUTs (Xrn1-sensitive Uncharacterized Transcripts) were detected in yeast strains depleted for Xrn1 5'-3' cytoplasmic exoribonuclease. They are by a majority (66%) antisense to protein-coding genes, and their accumulation induces transcriptional silencing of 273 genes, by an unknown RNAi-independent mechanism involving histone modifications (van Dijk et al. 2011).

Xrn1 is a critical exoribonuclease in yeast species, involved in normal and nonsense-mediated decay, hydrolysing RNA from 5' end releasing mononucleotides (Johnson 1997; Muhlrad et al. 1994; Gatfield & Izaurralde 2004), but also in RNAi (Orban & Izaurralde 2005). This enzyme is extremely conserved in all eukaryotes, including Drosophila and C. elegans (Newbury & Woollard 2004; Till et al. 1998), but also human (Chang 2011). In human cells, XRN1 has not been well characterized, but it already has been shown that XRN1 mRNA expression is lost or reduced in primary osteogenic sarcoma (OGS)-derived cell lines. Moreover, a heterozygous missense mutation at a conserved amino acid of XRN1 mRNA has been identified in OGS-derived cell line U2OS (Zhang et al. 2002). Given these observations, and by analogy with the work that has been done in yeast, we examined first XRN1 expression in human cell lines, and then tested whether siRNA-mediated knockdown of XRN1 would induce the accumulation of non-coding transcripts.

a. XRN1 expression in human cell lines

We first examined by western blot XRN1 protein level in several human cell lines, using a specific antibody recognizing the N-terminal sequence of the protein (ab162750, Abcam). We observed that XRN1 protein is present in all tested cell lines, but at different levels. Interestingly, the antibody we used revealed an additional band at approximately 225 kDa, whereas XRN1 protein, translated from the longest variant of XRN1 mRNA, has a molecular weight of 194 kDa (Figure 49a and b).

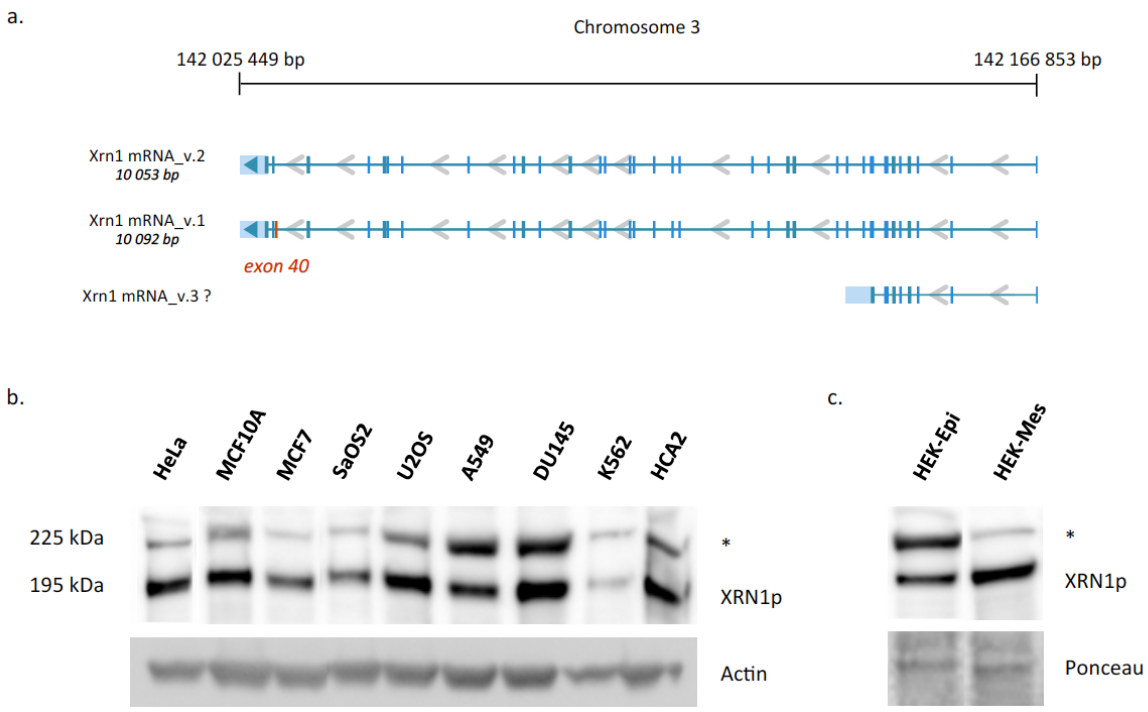

Figure 49 | XRN1 protein levels in human cell lines. **a.** XRN1 annotated isoforms. Quantification of XRN1 protein by western blot in **b.** several human cell lines and **c.** in HEK-Epi and HEK-Mes cells. Actin level was used as a loading control. * Additional protein detected by anti-XRN1 antibody.

In the same way, we analysed by western blot XRN1 protein in HEK-Epi and HEK-Mes cells, to examine whether XRN1 expression is affected, or XRN1 protein stabilized, after EMT. We observed an increase in XRN1 protein level in HEK-Mes compared to HEK-Epi cells, but also a decrease in the level of the additional band detected by the antibody (Figure 49c). These results suggest that XRN1 protein levels are sensitive to cell-type and changes in cell identity. The additional protein of 225 kDa detected in our western blot experiment could

be due to non-specific hybridization of the antibody, but could also be a post-translational modification of XRN1 protein.

b. SiRNA-mediated XRN1 knock-down in HeLa and MCF7 cells

Using three commercial siRNAs directed against XRN1 mRNA sequence (Life Technologies) (References available in Supplementary Table S3), we set up XRN1 knockdown in HeLa and MCF7 cells. Two successive transfections of 48 hours each, using 100 nM of siRNA and in presence of Lipofectamine 2000 (Life Technologies), were necessary to obtain efficient knock down of XRN1 as compared to cells transfected with a scramble, control siRNA (siCTR). We selected the most efficient siRNA (siXRN1). Western blot analysis of cells after transfection showed a complete depletion of XRN1 at a protein level in both HeLa and MCF7 cell lines (Figure 50a). RT-qPCR quantification of XRN1 mRNA abundance after transfection showed a partial 92% and 58% decrease in HeLa and MCF7 cells, respectively (Figure 50b). In parallel, we performed siRNA transfections in HEK-Epi and HEK-Mes cells, but we never succeed to obtain an efficient XRN1 knockdown in these cell lines.

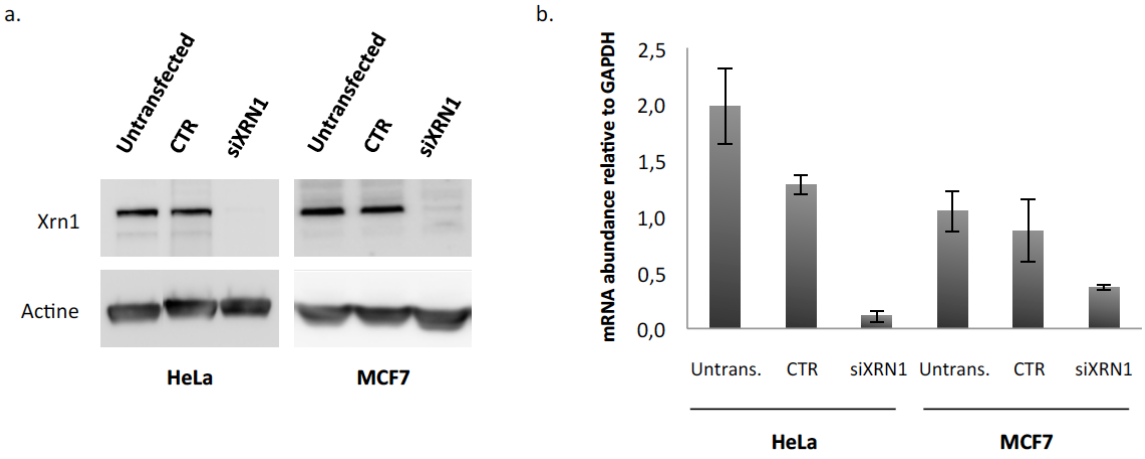

Figure 50 | XRN1 depletion by siRNA in HeLa and MCF7 cells. **a.** XRN1 protein level in cells 96 hours after two successive siRNA transfections. b. Oligo-specific RT-qPCR quantification of XRN1 mRNA level after siXRN1 transfections.

We tested then whether XRN1 knockdown in HeLa and MCF7 induced accumulation of RNA species. We measured by oligo-specific RT-qPCR the abundance of several known

lncRNAs and as-lncRNAs, but none of the selected transcripts showed differential expression in transfected cells compared to control condition *(data not shown)*. This results could indicate that these transcripts are not sensitive to 5'-3' cytoplasmic degradation, or that low levels of XRN1 left after siRNA transfection are sufficient to ensure a full XRN1 activity.

c. Establishment of coding and non-coding transcriptome upon XRN1 knockdown

48 hours after a second round of transfection with control or siXRN1 in HeLa and MCF7 cells, we extracted total RNAs and performed ribosomal RNA depletion. Then, strand-specific cDNA libraries were prepared in duplicates using SOLiD Total RNA-seq kit (Life Technologies). Paired-end sequencing was performed on a SOLiD 5500 sequencer, by the ICGex platform at the Institut Curie. For HeLa cells, RNA-sequencing generated on average 430 600 000 reads per samples, and bioinformatics analysis was performed as described previously. We observed that the quality of reads was very low, and succeed to map on average 39 000 000 of unique reads. 21 to 24% of read pairs were correct. For MCF7 cells, the quality of the sequencing was even lower, with very few uniquely mapped reads. Thus, we focused our analysis on HeLa cells.

We first analysed the full transcriptome of HeLa cells upon XRN1 knockdown, by comparison with cells transfected with siCTR. Unfortunately, the poor quality of sequenced reads, and the low coverage of the human genome due to a very low number of mapped reads, did not allow the assembly of novel lncRNAs and as-lncRNAs. We established density plots of mRNAs, as well as already annotated lncRNAs and as-lncRNAs in siXRN1 versus siCTR. We observed a slight down regulation of mRNAs in siXRN1 cells. LncRNAs and as-lncRNAs were even more affected, with distinct groups of transcripts expressed specifically in each condition (Figure 51a). These results were also observed by heat map representation of the three classes of transcripts (Figure 51b). We performed DEseq experiments to identify differentially expressed non-coding transcripts between siXRN1 and siCTR, but we did not succeed to find candidates responding to our criteria of p-value ≤ 0.05 and fold change ≥ 2 folds.

If the low number of uniquely mapped reads didn't allow the assembly of novel lncRNAs and as-lncRNAs, we identified several loci showing enriched clusters of mapped reads

between siXRN1 and siCTR. We selected two of them, antisense to fizzy/cell division cycle 20 related 1 (FZR1) (Figure 52a) and pregnancy-zone protein (PZP) coding genes (Figure 52b). FZR1 has been described to be an important regulator of G1 phase and is required for efficient DNA replication in human and mouse somatic cells (Sigl et al. 2009). PZP is a member of the alpha 2-macroglobulin plasma proteins, produced predominantly by the mammalian liver. PZP binds a wide variety of important pregnancy-associated molecules such as vascular endothelial growth factor and placenta growth factor (Tayade et al. 2005). We measured the abundance of antisense transcripts to these genes by oligo-specific RT-qPCR in RNA extracts from siXRN1 and siCTR cells. FZR1 as-lncRNA showed equivalent levels between HeLa cells transfected with siXRN1 and siCTR. PZP as-lncRNA exhibited a 3-fold increase in cells upon XRN1 knockdown, unfortunately not reproducible in independent siXRN1 experiments.

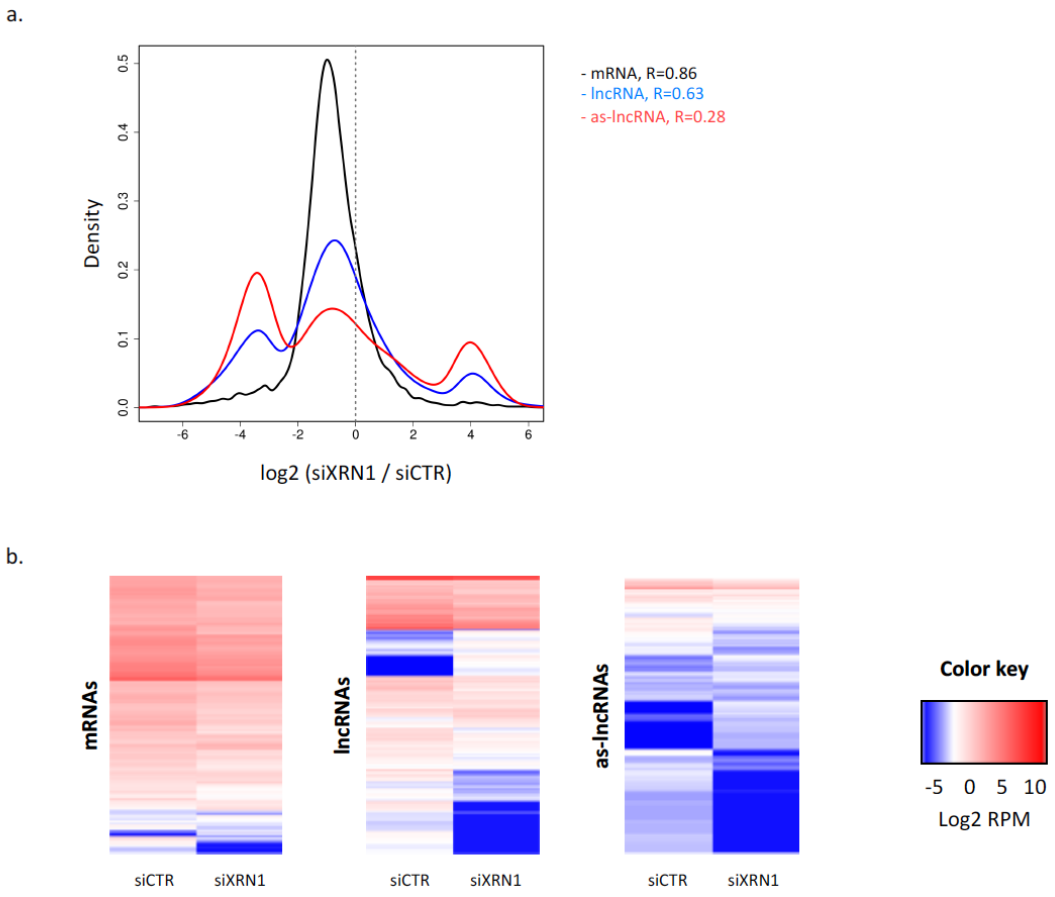

Figure 51 | XRN1-depleted HeLa cells full transcriptome. **a.** Density plots of mRNAs, lncRNAs, and as-lncRNAs in siXRN1 versus siCTR transfected HeLa cells. **b.** Heatmaps of mRNAs, lncRNAs and as-lncRNAs full catalogue.

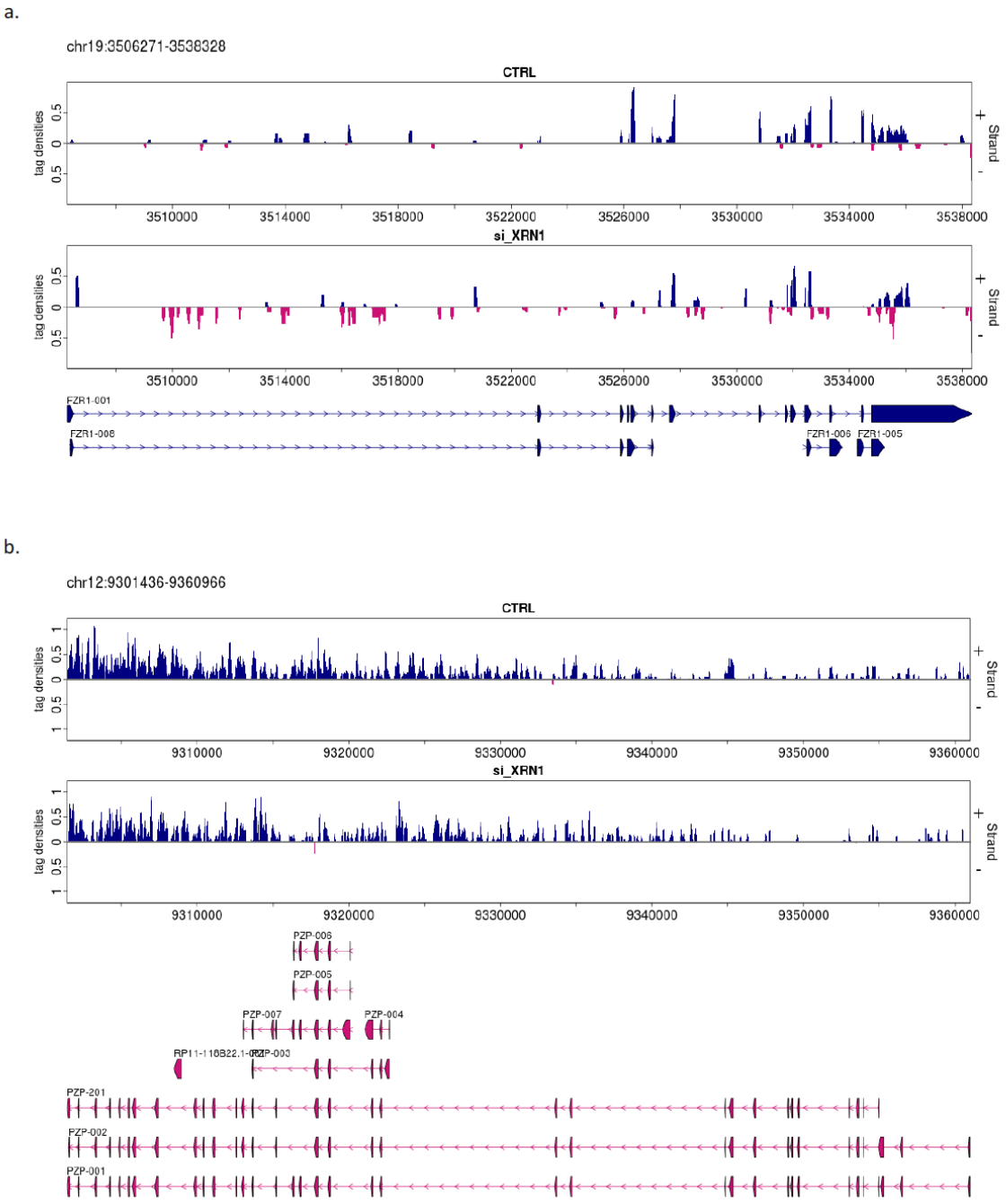

Figure 52 | FZR1 as-lncRNA and PZP as-lncRNA, two potential XRN1-sensitive transcripts in human cells. Observed clusters of mapped reads antisense to **a.** FZR1 protein-coding gene and **b.** PZP protein-coding genes. Snapshots were generated from HeLa siCTR and siXRN1 RNA-seq datasets using Ving bioinformatics tool.

If little is known about XRN1 in human cells, it seems clear that its role in 5'-3' cytoplasmic RNA decay is highly conserved. Decrease of its expression in osteogenic sarcoma-derived cell lines, missense mutation in its amino acids sequence (Zhang et al.

2002), and variability in its protein level in different human cell lines, particularly before and after the EMT, suggest that XRN1 could play an important role in cancer development. Previous studies performed by our laboratory in yeast *Saccharomyces cerevisiae* showed that XRN1 depletion leads to accumulation of cryptic regulatory lncRNAs that induces silencing of several protein-coding genes. Thus, we can speculate that these transcripts are conserved in human cells, with variability in XRN1 protein level inducing differential accumulation of cryptic transcripts and ultimately differential down regulation of target genes.

Unfortunately, the poor quality of RNA-sequencing performed using old SOLiD technology didn't allow us to assembly novel non-coding transcripts upon XRN1 depletion by siRNA in HeLa cells. Moreover, we observed no significant increase in level of known lncRNAs and as-lncRNAs. We can speculate that even low levels of XRN1 after siRNA transfection are sufficient to ensure its activity. On the other hand, we succeed to detect clusters of mapped reads antisense to protein-coding genes and accumulated upon XRN1 depletion, even if their validation by RT-qPCR was only possible in one experiment. More robust XRN1 depletion and high-quality RNA-sequencing are now necessary to examine the existence of XUTs and their potential regulatory role in human cells.

DISCUSSION AND PERSPECTIVES

A. HOTAIR in Epithelial-to-Mesenchymal Transition

HOTAIR, one of most studied lncRNAs in human cells, has been described as aberrantly expressed in numerous cancers, and highly correlated with metastasis and poor prognosis. If HOTAIR has been identified as a scaffold RNA guiding chromatin-modifying complexes PRC2 and LSD1/CoREST/REST to repress gene transcription at the epigenetic level, the link with the metastatic process is still not fully understood. A recent study showed that HOTAIR depletion in HCC1954 cells prevent the induction of EMT by TGF-β1 treatment, suggesting that HOTAIR is a major player in the initiation of the EMT process (Pádua Alves et al. 2013). However, results of HOTAIR loss- and gain-of-function experiments in the HEK model led us to a different hypothesis.

HOTAIR was identified as the most up regulated lncRNA after EMT by comparison between epithelial and mesenchymal immortalized HEK cells. Loss- and gain-of-function in HEK-Mes and HEK-Epi cells, respectively, induced major changes in cells phenotype, showing the importance of HOTAIR in the regulation of cell proliferation, migratory capacity and invasiveness. These results add evidence to a correlation between HOTAIR high expression and the metastatic potential of immortalized cells. HOTAIR depletion by ASOs in mesenchymal HEK cells was probably not efficient enough to induce a complete epigenetic reprogramming and major effects on the transcriptome, and therefore should be optimized. HOTAIR over expression experiments in epithelial HEK cells should be repeated. However, analysis of EMT markers expression suggested that HOTAIR loss- or gain-of-function do not revert or initiate the EMT program, respectively. Given these observations, we can speculate that HOTAIR is not a driver of EMT, but is involved in the regulation and maintenance of the mesenchymal migratory and invasive phenotype.

In mesenchymal HEK cells depleted for HOTAIR, we observed a slight decrease in β-catenin protein level, suggesting a down regulation of the Wnt/β-catenin signalling pathway, and possibly explaining the observed immediate phenotypic changes. As described previously, this down regulation could happen through silencing of WIF-1 gene (Ge et al. 2013). However, no change in the expression of this gene was found by RNA-sequencing data analysis, suggesting that other intermediate genes could be involved and targeted by HOTAIR. Intriguingly, no change was measured for β-catenin protein level in epithelial HEK

cells overexpressing HOTAIR. This observation suggests that HOTAIR-mediated regulation of the Wnt/β-catenin signalling pathway does not occur in epithelial cells where β-catenin is strictly localized at the cytoplasmic side of the membrane and interacts with E-cadherin.

HOTAIR binds and guides PRC2 and LSD1/CoREST/REST complexes to repress gene transcription at the epigenetic level (Tsai et al. 2010). However, its target genes seem to be highly dependent on the cell type. In accordance to this hypothesis, genes identified in previous studies as targets of HOTAIR-mediated regulation (Gupta et al. 2010) do not all show the expected changes in expression in the HEK model. RNA-sequencing approaches allowed the establishment of HOTAIR-targets list in these specific cell types. Consistently with the observed phenotypes, these genes are involved in cell-cell signalling, cell proliferation, but also in Wnt and MAPK signalling pathways.

Over expression of HOTAIR transcripts truncated for PRC2 or LSD1-interacting domains allowed further analysis of HOTAIR-mediated gene regulation mechanism. If it remains to be determined whether PRC2 and LSD1/CoREST/REST complexes can still bind HOTAIR in these conditions, we observed different profiles of phenotypic and transcriptomic changes in HEK-Epi cells. Three types of HOTAIR-mediated gene regulation seem to occur: a first one requiring both PRC2 and LSD1/CoREST/REST interaction domains; a second one requiring only one of the two complexes; a third one involving other factors, possibly interacting with HOTAIR through unknown motifs localized into its sequence. Additional effects on the transcriptome observed for each HOTAIR truncated form suggest that HOTAIR lacking one of the two interaction domains targets genes in a non-specific manner. In this case, the lack of LSD1-interacting domain seems to have dramatic consequences on HEK-Epi cells transcriptome. We could speculate that LSD1-interacting domain negatively regulates the guide function of HOTAIR, impeding its interaction with PRC2 and its activity. Another possibility is that LSD1-interacting domain is important for HOTAIR targeting to a major repressor or activator of transcription. Altogether, our results allow the establishment of a model presented in Figure 53.

Further investigations will be necessary to fully characterize HOTAIR in the HEK cell model and to examine in detail mechanisms associated with HOTAIR for the regulation of gene expression. Notably, different isoforms of HOTAIR have been annotated, but their specific role is not known. Therefore, their respective expression, cellular localisation and function should be examined.

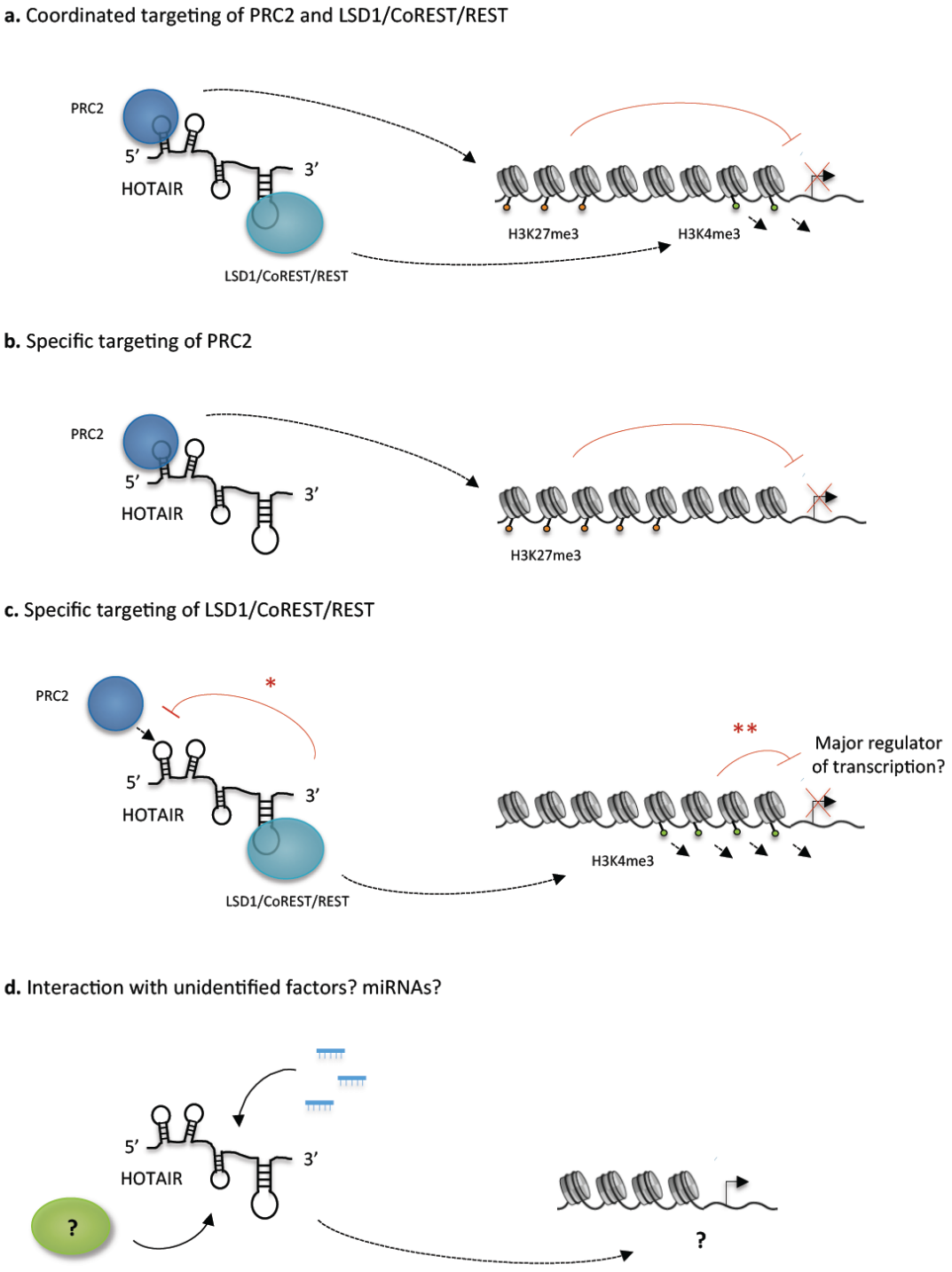

Figure 53 | Potential mechanisms of HOTAIR-mediated gene regulation. **a.** HOTAIR could bind both PRC2 and LS1/CoREST/REST complexes and coordinate their targeting to specific loci, resulting in gene silencing through H3K27 trimethylation and H3K4 demethylation. **b.** and **c.** HOTAIR could also target one of the two complexes to induce gene silencing. Two potential roles of LSD1/CoREST/REST complex are proposed: * modulation of PRC2-mediated gene silencing, or ** regulation of major activator(s) or repressor(s) of transcription. **d.** HOTAIR could bind a third protein factor or complex, yet unidentified, or interact with miRNAs, to regulate gene expression.

A role of PRC2 and LSD1 activity should be addressed in HEK cell lines, and particularly in HEK-Epi cells over expressing HOTAIR forms, using stable knockout of LSD1 and EZH2 component of PRC2. Further experiments should also be performed to examine epigenetic and transcriptional landscapes associated with HOTAIR full-length and truncated forms over expression in HEK cells. To determine whether another mechanism is involved in HOTAIR-mediated gene regulation, its interaction with additional factors should be further examined. Interestingly, a very recent study proposed that HOTAIR could function as a competitive endogenous RNA, sequestering miR331-3p in gastric cancer (Liu et al. 2014).

B. LncRNAs in Epithelial-to-Mesenchymal Transition

If we selected HOTAIR as a promising candidate for loss- and gain-of-function studies, transcriptome analysis revealed that EMT is associated with a massive deregulation of lncRNAs expression. Numerous lncRNAs and as-lncRNAs are significantly deregulated between epithelial and mesenchymal immortalized states of HEK cells, and could be potential inducers or regulators of EMT. Among already annotated transcripts, we identified MALAT1, H19 and GAS5 as significantly up regulated after EMT. The same type of functional studies could be performed in the HEK model for these transcripts. In addition, our results revealed important numbers of novel lncRNAs and as-lncRNAs, deregulated after EMT. We selected several of them, for their localisation, close to or antisense to protein-coding genes previously described in human diseases and cancer. Their structural organization, transcription and function should be examined. However, it is worth to note that the established list of lncRNAs deregulated upon EMT has to be more precisely defined. On one hand, mesenchymal HEK cells display a high genomic instability with numerous chromosome rearrangements, and several transcripts identified as deregulated upon EMT could in reality only be linked to local genomic gains or losses. On the other hand, RNA-sequencing approaches have technical limitations.

C. RNA-sequencing: Technical limitations

Using strand-specific and paired-end RNA-sequencing approaches allowed the detection of annotated transcripts and the assembly of novel lncRNAs and as-lncRNAs, as well as their relative quantification. However, protocol for libraries preparation includes RNA fragmentation, known to introduce sequence content-based biases, and reverse transcription

followed by PCR amplification, which introduces biases due to non-linear amplification. These technical issues can lead to identification of transcripts that are, in reality, only artefacts, and to errors in the estimation of RNAs level of expression. Generation of ChIP-seq data, and their comparison with RNA-seq, could help to define active transcription units, and eliminate sequencing artefacts. Another limit of RNA-seq approaches is the assembly of novel transcripts, which could not be always accurate, particularly for many lncRNAs and as-lncRNAs which abundance is very low. Additional data, such as CAGE tags, could be generated and help in the identification of transcriptional start sites. But several aspects of transcripts will still need to be further examined, such as exons, isoforms, 3' UTRs, and repeated elements. An accurate definition of lncRNAs is notably important in the examination of their coding potential, assessed by *in silico* translation of transcripts, and comparison of obtained amino acid sequences to known protein coding sequences and protein domain families among species. Taking all these considerations into account will be necessary to clearly define the catalogue of lncRNAs deregulated upon EMT, and in general to precisely characterize and classify lncRNAs.

D. Pervasive Transcription

Our results add numerous novel lncRNAs and as-lncRNAs, including INATs, to the increasing catalogues of ncRNAs that continue to be discovered through RNA-seq in different cell and tissue types, and under different environmental conditions. If the fact that the human genome is pervasively transcribed is now well established, the usefulness of such transcription activity remains to be defined. To date, accurate bioinformatics prediction of lncRNA functions is not possible, due to the lack of data on the relationship between sequence, expression and function. Targeted experiments are necessary, but difficult, firstly because most lncRNAs are not conserved, secondly because RNA interference experiments are challenging for nuclear chromatin-associated molecules. Finally, it is not easy to experimentally discriminate between a direct function of a mature lncRNA, its nascent precursor, or the transcription process producing it. This observation could explain why few examples of lncRNAs have been functionally characterized. These examples however show the importance of non-coding transcripts in numerous cellular processes, by regulation of chromatin domains and gene expression.

However, given the high number of identified non-coding transcripts and the few examples serving some function, we can also hypothesize that a part of pervasive transcription is due to uncontrolled transcription initiation. These transcription events need to be contained, as they can disrupt cellular processes, such as transcription or maintenance of genome stability. They might also be toxic for a cell, if interacting with complementary DNA sequences, or sequestering RNA-binding factors. Therefore, transcription termination and degradation are highly controlled by several pathways, such as XRN1-mediated cytoplasmic decay. On another point of view, pervasive transcription initiation might be important in the maintenance of open chromatin domains, for rapid activation of transcription in response to environmental signals. Pervasive transcription *per se* may also results in a large cohort of transcripts which levels are easily regulated and highly dynamic upon specific conditions. Analysis of cryptic transcriptome, upon depletion of transcription termination and RNA degradation pathways, is a way to examine the reality of these hypotheses.

In conclusion, our work identified by RNA-sequencing approaches not only HOTAIR and annotated lncRNAs, but also numerous novel lncRNAs and as-lncRNAs that are highly deregulated in the context of EMT, a critical cellular program involved in the acquisition of a mesenchymal and metastatic phenotype by an epithelial cancer cell. If a complete characterization of these lncRNAs will be necessary to examine if they have a regulatory function in the EMT process, they could be relevant as biomarkers for cancer progression and metastasis.

MATERIAL AND METHODS

1. Cells origin and culture conditions

HEK-Epi and HEK-Mes cells were generated as HEK-HA5-Early and HEK-HA5-Late, respectively, and kindly provided by Arturo Londoño-Vallejo's laboratory (Castro-Vega et al. 2013). HeLa (CCL-2™), MCF7 (HTB-22™) and HEK293T/17 (CRL-11268™) cell lines were obtained from ATCC. Cells were maintained under standard culture conditions in a humidified 5% CO_2 atmosphere at 37°C, in the following media: DMEM high glucose for HeLa and HEK293T cells, and MEM alpha without nucleosides for HEK-Epi, HEK-Mes and MCF7 cells (Life Technologies). Both media were supplemented with 10% foetal bovine serum (FBS), essential amino acids and sodium pyruvate. Cells proliferation rate was monitored by cell counting. Population doublings were calculated using the following formula:

Number of PDs = (Log(final count) – Log(initial count)) / 0.301

2. Migration and invasion assays

Wound healing assays were performed in cells at 80% confluence by generation of a scratch using a 200 µl pipette tip. Images were captured at 0, 12 and 24 hours post-scratch using a Zeiss Axiovert 135 microscope. Wound healing potential was quantified using TScratch software and expressed as a percentage of invaded area.

Migration and invasion assays were performed using ThinCert™ Cell culture inserts (657638, Greiner bio-one), rehydrated for 1 hour in warm serum-free culture medium, and coated with ECM Gel from Engelbreth-Holm-Swarm murine sarcoma (E1270, Sigma-Aldrich). Inserts were then placed in a 6-well culture plate. 200 000 cells suspended in serum-free culture medium were seeded in a coated culture insert, and 2 ml of 10% FBS culture medium was added to the bottom of the well. Migration chambers were incubated at 37°C in a humidified 5% CO_2 atmosphere. After 72 hours, non-migrating cells were removed from the upper side of the chamber with cotton swabs. Migrating cells were fixed with 100% methanol for 15 minutes, then stained with 0.2% Crystal violet for 1 hour. 10 high power field pictures were taken by sample, and number migrating cells was counted using ImageJ software.

3. RNA extraction, processing and analysis

Total RNA extractions were performed directly from 80% confluent cells cultures with Trizol reagent according to the manufactures instructions (15596-026, Life Technologies). Briefly, cells were lysed directly in culture dishes by adding 700 µl of Trizol reagent, scraped and transferred into Eppendorf tubes. Cell suspension was mixed with 200 µl of chloroform, then centrifuged at 12 000 x g for 15 minutes at 4°C. The upper, aqueous phase was recovered, and RNA was precipitated with 350 µl of isopropanol, for 30 minutes, at 12 000 x g and 4°C. After ethanol washes, RNA pellet was dried and resuspended in 50 µl of RNAse-free water. RNA concentration was measured using a Nanodrop.

Ribosomal RNA depletion was performed from 5 μg of total RNA with the RiboMinus™ Eukaryote Kit for RNA-seq (A10837-08, Life Technologies). PolyA+ RNA extraction was carried out from 50 μg of total RNA using the FastTrack MAG Maxi mRNA Isolation Kit (K1580-02, Life Technologies). Cytoplasmic and nuclear RNA were isolated from 10^7 cells using the PARIS™ Kit (AM1921, Life Technologies).

Reverse Transcription (RT) was performed on either 100 ng of RNA with random and oligo(dT) primers mix (iScript™ cDNA Synthesis Kit, 170-8891, BioRad) or 1 μg of RNA with specific oligonucleotides (SuperScript II Reverse Transcriptase, 18064-014, Life Technologies) (see Supplementary Table S1 for details). Reactions without reverse transcriptase were included as negative controls. cDNA was then diluted 10 to 40 times and quantified by real-time PCR with LightCycler® 480 SYBR Green I Master (04707516001, Roche). POLR2A, RPL11 and GAPDH were used as controls to quantify relative abundance of RNA.

Northern blot experiments were performed using total RNA (5-10 µg), resolved on 1% formaldehyde containing agarose gels, transferred to nitrocellulose membranes (Hybond-XL, RPN2020S, GE Healthcare), UV crosslinked and revealed by over-night hybridization with p^{32}-labeled specific oligonucleotide at 42°C in the ULTRAhyb®-Oligo Hybridization Buffer (AM8663, Ambion).

4. Protein extraction and analysis

Proteins were extracted from 80% confluent cells, washed twice and scraped in ice-cold PBS. Cells were pelleted by centrifugation at 1 500 rpm for 10 minutes at 4°C, and lysed in 100 µl of RIPA buffer (R0278, Thermo Scientific) supplemented with 0.1 mM AEBSF (A8456, Sigma-Aldrich) for 30 minutes at 4°C. Samples were centrifuged at 12 000 x g for 30 minutes at 4°C, to recover a protein-containing supernatant. Protein concentration was measured using the BCA Protein Assay Kit (23225, Thermo Scientific).

1 and 5 μg of whole protein extract were supplemented with 2x Laemmli Buffer (161-0737, BioRad), incubated for 15 minutes at 95°C and separated in a NuPAGE Novex 4-12% Bis-Tris Protein Gel (NP0322BOX, Life Technologies) in a NuPAGE SDS Sample Buffer (NP0007, Life Technologies) at 100 V for 2 hours. Proteins were then transferred on a nitrocellulose membrane using iBlot Dry Blotting System (Life Technologies) for 10 minutes. Prior to blotting, the membrane was coloured by Rouge Ponceau for 10 minutes and photographed, then washed in PBST (1x PBS, 0.1% Tween 20). The membrane was blocked with 5% milk in PBST for at least 30 minutes. Immunoblotting was performed using the Epithelial-to-Mesenchymal Transition Western Blot Cocktail (1:2 500, ab157392, Abcam) or the anti-Xrn1 antibody (1:10 000, ab70259, Abcam), in 5% milk-PBST over night at 4°C. Membranes were washed three times with PBST and incubated with either secondary antibody cocktail (1:2 500, ab157392, Abcam) or anti-rabbit secondary antibody (SAB3700834, Sigma-Aldrich) in 5% milk-PBST for 1 hour at room temperature. Blots were washed three times with PBST, and incubated with the SuperSignal West Pico Chemiluminescent Substrate (34079, Pierce) for 10 minutes. Blots were then developed using the ImageQuant LAS 4000 (GE Healthcare). Protein amounts were quantified using ImageJ software.

5. siRNA transfections and ASO treatments

Sequences of used siRNAs are listed in Supplementary Table S3. 100 nM of siRNA were mixed with 500 μl of Opti-MEM reduced serum medium (31985062, Life Technologies). In parallel, 10 μl of Lipofectamine 2000 transfection reagent (11668-019, Life Technologies) were added to 490 μl of Opti-MEM. After 5 minutes of incubation, siRNA and lipofectamine preparations were mixed, incubated at room temperature for 20 minutes, and added to 60%

confluent cells. 6 hours after transfection, cells were washed with PBS, before addition of fresh complete medium.

Sequences of used ASOs are listed in Supplementary Table S3. 10 μM of ASO were simply added to fresh culture medium used to culture cells, and replaced after 2 days for additional 3 days of treatment.

6. Plasmids, transfections and transductions

All plasmids used in this study are listed in Supplementary Table S4. Plasmids expressing HOTAIR forms and GFP were created using Gateway Cloning Technology (Life Technologies). Briefly, GFP, HOTAIR full-length and short transcripts (HOTΔPRC2 and HOTΔLSD1) cDNA sequences were amplified using as a template pFA6a-GFP(S65T)-His3MX6 (41598, Addgene) and pLZRS-HOTAIR plasmid, kindly provided by Dr. Howard Chang. PCR reactions were performed with Phusion High Fidelity DNA Polymerase (M0530S, New England Biolabs). PCR products were purified in 1% agarose 0.5x TBE gel in presence of Ethidium Bromide, using QIAquick gel extraction kit (28706, Qiagen), and cloned into pDONR201 plasmid (Life Technologies) using BP clonase II enzyme mix (12535-019, Life Technologies). BP reactions were treated with 2 μg of Proteinase K (10 165 921 001, Roche) for 1 hour at 37°C, transformed and amplified into TOP10 chemiocompetent E. coli, according to manufacturer's protocol. Plasmids were extracted from kanamycine resistant clones using QIAprep Spin Miniprep Kit (27106, Qiagen), sequenced (GATC Biotech) and validated by restriction digestion. LR reactions were then performed, using pLenti6.2/V5-DEST plasmid (V368-20, Life Technologies) as a destination vector. LR products were treated with 2 μg of Proteinase K for 1 hour at 37°C transformed and amplified into TOP10 chemiocompetent E. coli. Plasmids were extracted from ampicilline resistant clones, sequenced and validated by restriction digestion.

HEK293T/17 cells were cultured in complete medium, and incubated for 48 hours with a transfection mix containing 20 μg of the generated plasmid (pGFP, pHOT, pHOTΔPRC2 or pHOTΔLSD1), 15 μg of psPAX2 (12260, Addgene), and 6 μg of pCMV-VSV-G (8454, Addgene), using Polythylenimine and in presence of NaCl. The virus-containing supernatant was recovered from transfected HEK293T/17 cells, passed through 0.2 μm filter and added to 80% confluent HEK-Epi cells. Cells were cultured for 24 hours, then washed in PBS. Fresh

complete medium containing 5 μg/ml blasticidine (A11139-03, Life Technologies) was added. Cells were maintained under antibiotic selection for 10 days before growing in normal medium.

7. RNA-seq library preparation

For SOLiD sequencing, 1 μg of RNA depleted for ribosomal RNA was subjected for 3 min fragmentation by Zinc solution (AM8740, Life Technologies), purified using RiboMinus™ Concentration Module (K1550-05, Life Technologies), dephosphorylated by Antartic phosphatase (M0289S, New England Biolabs) and phosphorylated by T4 Polynucleotide Kinase (M0201S, New England Biolabs). RNA fragments were again purified using RiboMinus™ Concentration Module and resolved on 6% TBE-Urea gels to select 100-200 nucleotides fragments. RNA fragments were eluted from gel overnight in 0.3 M NaCl at 4°C and precipitated by Ethanol. Ligation of adapters, RT and PCR library amplification were perfomed according to manufactures procedures with the use of SOLiD™ Total RNA-Seq Kit (4445374, Life Technologies) with SOLiD™ RNA Barcoding Kit, Module 1-16 (4427046, Life Technologies). For adaptation of SOLiD RNA-seq library to Illumina sequencing, 5 cycles of PCR were added with adaptors containing both SOLiD and Illumina barcoding sequences. For Illumina sequencing, RNA libraries were prepared from 500 ng of total RNA, according to manufactures procedures with the use of TruSeq Stranded Total RNA Kit (RS-122-2201, Illumina).

8. Chromatin Immunoprecipitation

Cells were cultured in T300 flasks until reaching 80% confluence, then crosslinked for 10 minutes at 37°C in 30 ml of growth medium containing 1.5 ml of freshly prepared Crosslinking buffer (1% formaldehyde, 0.5 M Tris pH 8, 1 M NaCl, 10 mM EDTA, 5 mM EGTA). The reaction was stopped by adding 1.5 ml of 2.5 M Glycine for 5 minutes at room temperature. After two washes with ice-cold PBS, cells were scraped, collected, washed two more times and pelleted by centrifugation 10 minutes at 4°C, 1 500 rpm. The cell pellet was resuspended in 6 ml of Cell Lysis buffer (5 mM PIPES pH 8, 85 mM KCl, 0.5% NP40, 0.1 mM AEBSF) and incubated on ice for 10 minutes. The lysate was passed through a 20G long needle using a 10 ml syringe, and centrifuged at 3 000 rpm for 12 minutes at 4°C. The nuclei pellet was resuspended in 1 ml of Sonication buffer (Tris-HCl pH 8, 1% SDS, 10 mM EDTA,

0.1 mM AEBSF). Samples were sonicated in Bioruptor Plus (Diagenode) for 30 minutes (30 seconds ON, 30 seconds OFF, "high" mode) at 4°C. Sonicated samples were centrifuged 10 minutes at 1 300 rpm, at 4°C. DNA and protein concentration were quantified using Nanodrop and BCA protein assay kit, respectively. Sonication efficiency was verified on a 1% agarose gel in 0.5x TBE.

20 and 50 μl of Protein A beads (10001D, Life Technologies) were blocked by incubation with 1 800 μl of 0.5% BSA in FA150 (50 mM Hepes pH 7.5, 1% Triton X-100, 1 mM EDTA, 150 mM NaCl, 0.1% Na-deoxycholate, 0.1% SDS), for 2 hours or overnight at 4°C, washed three times with 1 ml of ice-cold FA150 and resuspended in 20 and 50 μl of FA150. 20 μl of blocked beads were used for chromatin pre-clearing, and 50 μl were used for IP. For each chromatin sample, the equivalent of 20 μg of DNA was diluted 10 times with FA150 without SDS and incubated with 20 μl of blocked beads for 3 hours at 4°C. Beads were sedimented in a magnetic device, a 1/10 volume of the pre-cleared chromatin was taken as the INPUT-DNA sample, the rest was incubated with a specific antibody (anti-H3K4me3, 4 μg, C15410003, Diagenode; anti-RNA Pol II 8WG16, 2 μg, mms-126R, Covance) over night at 4°C on a rotating wheel. 50 μl of blocked beads were then added, and samples were incubated for 2 more hours at 4°C. Beads were then washed 10 minutes at 4°C once with FA150, twice with FA500 (50 mM Hepes pH 7.5, 1% Triton X-100, 1 mM EDTA, 500 mM NaCl, 0.1% Na-deoxycholate, 0.1% SDS), and once with TE (10 mM Tris-HCl pH 8, 0.5 mM EDTA). Chromatin was eluted in 100 μl of the Elution Buffer (50 mM Tris-HCl pH 7.5, 1% SDS, 10 mM EDTA) at 65°C for 1 hour.

INPUT and IP chromatin samples were further incubated over night at 65°C to reverse a cross-link, in the presence of 7 μl of 20 mg/ml Pronase (10 165 921 001, Roche). Samples were treated with 2 μg of RNAse A for 1 hour at 37°C, and purified by QIAquick PCR Purification Kit. DNA was then analysed by qPCR using the same protocol as previously described.

REFERENCES

Adams, J.M. & Cory, S., 2007. The Bcl-2 apoptotic switch in cancer development and therapy. *Oncogene*, 26(9), pp.1324–1337.

Agbulut, O. et al., 2007. Green fluorescent protein impairs actin-myosin interactions by binding to the actin-binding site of myosin. *The Journal of biological chemistry*, 282(14), pp.10465–71.

Anastas, J.N. & Moon, R.T., 2013. WNT signalling pathways as therapeutic targets in cancer. *Nature reviews. Cancer*, 13(1), pp.11–26.

Ansieau, S. et al., 2008. Induction of EMT by twist proteins as a collateral effect of tumor-promoting inactivation of premature senescence. *Cancer cell*, 14(1), pp.79–89.

Aravin, A. et al., 2006. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature*, 442(7099), pp.203–7.

Armitage, P. & Doll, R., 1954. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British journal of cancer*, VIII(I), pp.1–12.

Artandi, S.E. & DePinho, R. a, 2010. Telomeres and telomerase in cancer. *Carcinogenesis*, 31(1), pp.9–18.

Azzalin, C.M. et al., 2007. Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. *Science (New York, N.Y.)*, 318(5851), pp.798–801.

Van Bakel, H. et al., 2010. Most "dark matter" transcripts are associated with known genes. *PLoS biology*, 8(5), p.e1000371.

Barrallo-Gimeno, A. & Nieto, M.A., 2005. The Snail genes as inducers of cell movement and survival: implications in development and cancer. *Development (Cambridge, England)*, 132(14), pp.3151–61.

Barrett, J.C., 1993. Mechanisms of multistep carcinogenesis and carcinogen risk assessment. *Environmental health perspectives*, 100(12), pp.9–20.

Barsyte-Lovejoy, D. et al., 2006. The c-Myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis. *Cancer research*, 66(10), pp.5330–7.

Bartel, B., 2005. MicroRNAs directing siRNA biogenesis. *Nature structural & molecular biology*, 12(7), pp.569–71.

Bartel, D.P., Lee, R. & Feinbaum, R., 2004. MicroRNAs : Genomics , Biogenesis , Mechanism , and Function Genomics : The miRNA Genes. , 116, pp.281–297.

Batista, P.J. & Chang, H.Y., 2013. Long noncoding RNAs: cellular address codes in development and disease. *Cell*, 152(6), pp.1298–307.

Beltran, M. et al., 2008. A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes & development*, 22(6), pp.756–69.

Bernard, D. et al., 2010. A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. *The EMBO journal*, 29(18), pp.3082–93.

Berx, G. & van Roy, F., 2009. Involvement of members of the cadherin superfamily in cancer. *Cold Spring Harbor perspectives in biology*, 1(6), p.a003129.

Bhowmick, N. a, Zent, R., et al., 2001. Integrin beta 1 signaling is necessary for transforming growth factor-beta activation of p38MAPK and epithelial plasticity. *The Journal of biological chemistry*, 276(50), pp.46707–13.

Bhowmick, N. a, Ghiassi, M., et al., 2001. Transforming growth factor-beta1 mediates epithelial to mesenchymal transdifferentiation through a RhoA-dependent mechanism. *Molecular biology of the cell*, 12(1), pp.27–36.

Bierie, B. & Moses, H.L., 2006. Tumour microenvironment: TGFbeta: the molecular Jekyll and Hyde of cancer. *Nature reviews. Cancer*, 6(7), pp.506–20.

Bissell, M.J. et al., 2002. The organizing principle: microenvironmental influences in the normal and malignant breast. *Differentiation; research in biological diversity*, 70, pp.537–546.

Blasco, M. a, 2007. The epigenetic regulation of mammalian telomeres. *Nature reviews. Genetics*, 8(4), pp.299–309.

Bodnar, a. G., 1998. Extension of Life-Span by Introduction of Telomerase into Normal Human Cells. *Science*, 279(5349), pp.349–352.

Bonaldo, M.F., Lennon, G. & Soares, M.B., 1996. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Research*, 6(9), pp.791–806.

Bovolenta, M. et al., 2012. The DMD locus harbours multiple long non-coding RNAs which orchestrate and control transcription of muscle dystrophin mRNA isoforms. *PloS one*, 7(9), p.e45328.

Braconi, C. et al., 2011. Expression and functional role of a transcribed noncoding RNA with an ultraconserved element in hepatocellular carcinoma. *Proceedings of the National Academy of Sciences of the United States of America*, 108(2), pp.786–91.

Brase, J.C. et al., 2011. Circulating miRNAs are correlated with tumor progression in prostate cancer. *International journal of cancer. Journal international du cancer*, 128(3), pp.608–16.

Burkhart, D.L. & Sage, J., 2008. Cellular mechanisms of tumour suppression by the retinoblastoma gene. *Nature reviews. Cancer*, 8(9), pp.671–82.

Cabili, M.N. et al., 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development*, 25(18), pp.1915–27.

Calin, G. a et al., 2007. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer cell*, 12(3), pp.215–29.

Cano, C.E. et al., 2012. Homotypic cell cannibalism, a cell-death process regulated by the nuclear protein 1, opposes to metastasis in pancreatic cancer. *EMBO molecular medicine*, 4(9), pp.964–79.

Carlile, M. et al., 2009. Strand selective generation of endo-siRNAs from the Na/phosphate transporter gene Slc34a1 in murine tissues. *Nucleic acids research*, 37(7), pp.2274–82.

Carrieri, C. et al., 2012. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature*, 491(7424), pp.454–7.

Castro-Vega, L.J. et al., 2013. Telomere crisis in kidney epithelial cells promotes the acquisition of a microRNA signature retrieved in aggressive renal cell carcinomas. *Carcinogenesis*, 34(5), pp.1173–80.

Cavallaro, U. & Christofori, G., 2004. Multitasking in Tumor Progression: Signaling Functions of Cell Adhesion Molecules. *Annals of the New York Academy of Sciences*, 1014(1), pp.58–66.

Chan, K.C.A. et al., 2013. Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. *Clinical chemistry*, 59(1), pp.211–24.

Chang, J.H., 2011. NIH Public Access. , 18(3), pp.270–276.

Cheetham, S.W. et al., 2013. Long noncoding RNAs and the genetics of cancer. *British journal of cancer*, 108(12), pp.2419–25.

Cheng, J. et al., 2012. piR-823, a novel non-coding small RNA, demonstrates in vitro and in vivo tumor suppressive activity in human gastric cancer cells. *Cancer letters*, 315(1), pp.12–7.

Cheng, J. et al., 2011. piRNA, the new non-coding RNA, is aberrantly expressed in human cancer cells. *Clinica chimica acta; international journal of clinical chemistry*, 412(17-18), pp.1621–5.

Chin, H.G. et al., 2007. Automethylation of G9a and its implication in wider substrate specificity and HP1 binding. *Nucleic acids research*, 35(21), pp.7313–23.

Chu, C. et al., 2011. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Molecular cell*, 44(4), pp.667–78.

Churchman, L.S. & Weissman, J.S., 2011. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, 469(7330), pp.368–73.

Ciccia, A. & Elledge, S.J., 2011. The DNA Damage Response : Making it safe to play with knives. , 40(2), pp.179–204.

Clark, M.B. et al., 2011. The reality of pervasive transcription. *PLoS biology*, 9(7), p.e1000625; discussion e1001102.

Clemson, C.M. et al., 2010. is Essential for the Structure of Paraspeckles. , 33(6), pp.717–726.

Clevers, H. & Nusse, R., 2012. Wnt/β-catenin signaling and disease. *Cell*, 149(6), pp.1192–205.

Collins, R. & Cheng, X., 2010. A case study in cross-talk: the histone lysine methyltransferases G9a and GLP. *Nucleic acids research*, 38(11), pp.3503–11.

Colotta, F. et al., 2009. Cancer-related inflammation, the seventh hallmark of cancer: links to genetic instability. *Carcinogenesis*, 30(7), pp.1073–81.

Conacci-Sorrell, M. et al., 2003. Autoregulation of E-cadherin expression by cadherin-cadherin interactions: the roles of beta-catenin signaling, Slug, and MAPK. *The Journal of cell biology*, 163(4), pp.847–57.

Correa, P., 2003. Helicobacter Pylori Infection and Gastric Cancer Helicobacter Pylori Infection and Gastric Cancer. *Cancer Epidemiology, Biomarkers and Prevention*, 12.

Counter, C.M. et al., 1992. Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity. *The EMBO journal*, 11(5), pp.1921–9.

Coussens, L.M. & Werb, Z., 2002. Inflammation and cancer. *Nature*, 420(6917), pp.860–867.

Cunnington, M.S. et al., 2010. Chromosome 9p21 SNPs Associated with Multiple Disease Phenotypes Correlate with ANRIL Expression. *PLoS genetics*, 6(4), p.e1000899.

Davis, I.J. et al., 2003. Cloning of an Alpha-TFEB fusion in renal tumors harboring the t(6;11)(p21;q13) chromosome translocation. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10), pp.6051–6.

DeBerardinis, R.J. et al., 2008. The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. *Cell metabolism*, 7(1), pp.11–20.

DeNardo, D.G., Andreu, P. & Coussens, L.M., 2010. Interactions between lymphocytes and myeloid cells regulate pro- versus anti-tumor immunity. *Cancer metastasis reviews*, 29(2), pp.309–16.

Depinho, R.A., 2000. The age of cancer. , 408(November).

der-Sarkissian, H. et al., 2004. The shortest telomeres drive karyotype evolution in transformed cells. *Oncogene*, 23(6), pp.1221–8.

Derrien, T. et al., 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research*, 22(9), pp.1775–89.

Derynck, R., Akhurst, R.J. & Balmain, a, 2001. TGF-beta signaling in tumor suppression and cancer progression. *Nature genetics*, 29(2), pp.117–29.

Van Dijk, E.L. et al., 2014. Ten years of next-generation sequencing technology. *Trends in Genetics*, pp.1–9.

Van Dijk, E.L. et al., 2011. XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature*, 475(7354), pp.114–7.

Division, C. et al., 2002. M olecules and Cells . , 14(3), pp.425–430.

Djebali, S. et al., 2012. Landscape of transcription in human cells. *Nature*, 489(7414), pp.101–8.

Dugimont, T. et al., 1998. The H19 TATA-less promoter is efficiently repressed by wild-type tumor suppressor gene product p53. *Oncogene*, 16(18), pp.2395–401.

El-Serag, H.B., 2002. Hepatocellular carcinoma and hepatitis C in the United States. *Hepatology (Baltimore, Md.)*, 36(5 Suppl 1), pp.S74–83.

Esteller, M., 2011. Non-coding RNAs in human disease. *Nature reviews. Genetics*, 12(12), pp.861–74.

Evan, G., 1998. A Matter of Life and Cell Death. *Science*, 281(5381), pp.1317–1322.

Faghihi, M.A. et al., 2008. Expression of a noncoding RNA is elevated in Alzheimer ' s disease and drives rapid feed-forward regulation of b -secretase. , 14(7), pp.723–730.

Farnebo, M., Bykov, V.J.N. & Wiman, K.G., 2010. The p53 tumor suppressor: a master regulator of diverse cellular processes and therapeutic target in cancer. *Biochemical and biophysical research communications*, 396(1), pp.85–9.

Fatica, A. & Bozzoni, I., 2014. Long non-coding RNAs: new players in cell differentiation and development. *Nature reviews. Genetics*, 15(1), pp.7–21.

Fearon, E.F. & Vogelstein, B., 1990. for Colorectal Tumorigenesis. , 61, pp.759–767.

Feijs, K.L.H. et al., 2013. Macrodomain-containing proteins: regulating new intracellular functions of mono(ADP-ribosyl)ation. *Nature reviews. Molecular cell biology*, 14(7), pp.443–51.

Fellig, Y. et al., 2005. H19 expression in hepatic metastases from a range of human carcinomas. *Journal of clinical pathology*, 58(10), pp.1064–8.

Feng, J. et al., 2006. The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes & development*, 20(11), pp.1470–84.

Fidler, I.J., 2003. The pathogenesis of cancer metastasis: the "seed and soil" hypothesis revisited. *Nature reviews. Cancer*, 3(June), pp.1–6.

Frenck, R.W., Blackburn, E.H. & Shannon, K.M., 1998. The rate of telomere sequence loss in human leukocytes varies with age. *Proceedings of the National Academy of Sciences of the United States of America*, 95(10), pp.5607–10.

Fuster, D.G. & Alexander, R.T., 2014. Traditional and emerging roles for the SLC9 Na+/H+ exchangers. *Pflügers Archiv : European journal of physiology*, 466(1), pp.61–76.

Garcia-Carbonero, R., Carnero, A. & Paz-Ares, L., 2013. Inhibition of HSP90 molecular chaperones: moving into the clinic. *The lancet oncology*, 14(9), pp.e358–69.

Gatfield, D. & Izaurralde, E., 2004. Nonsense-mediated messenger RNA decay is initiated by endonucleolytic cleavage in Drosophila. *Nature*, 429(6991), pp.575–8.

Ge, X.-S. et al., 2013. HOTAIR, a prognostic factor in esophageal squamous cell carcinoma, inhibits WIF-1 expression and activates Wnt pathway. *Cancer science*, 104(12), pp.1675–82.

Gebäck, T. et al., 2009. TScratch: a novel and simple software tool for automated analysis of monolayer wound healing assays. *BioTechniques*, 46(4), pp.265–74.

Gehrke, S. et al., 2010. Pathogenic LRRK2 negatively regulates microRNA-mediated translational repression. *Nature*, 466(7306), pp.637–41.

Gene, T. & Consortium, O., 2000. Gene Ontology : tool for the. , 25(may), pp.25–29.

Gibb, E. a, Brown, C.J. & Lam, W.L., 2011. The functional role of long non-coding RNA in human carcinomas. *Molecular cancer*, 10(1), p.38.

Girard, A. et al., 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, 442(7099), pp.199–202.

Gisselsson, D. et al., 2001. Telomere dysfunction triggers extensive DNA fragmentation and evolution of complex chromosome abnormalities in human malignant tumors. *Proceedings of the National Academy of Sciences of the United States of America*, 98(22), pp.12683–8.

Gong, C. & Maquat, L.E., 2011. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature*, 470(7333), pp.284–8.

Goto, H. et al., 2003. Transduction of green fluorescent protein increased oxidative stress and enhanced sensitivity to cytotoxic drugs in neuroblastoma cell lines Transduction of green fluorescent protein increased oxidative stress and enhanced sensitivity to cytotoxic drugs i. , pp.911–917.

Gottardi, C.J., Wong, E. & Gumbiner, B.M., 2001. E-cadherin suppresses cellular transformation by inhibiting beta-catenin signaling in an adhesion-independent manner. *The Journal of cell biology*, 153(5), pp.1049–60.

Gregory, P. a et al., 2011. An autocrine TGF-beta/ZEB/miR-200 signaling network regulates establishment and maintenance of epithelial-mesenchymal transition. *Molecular biology of the cell*, 22(10), pp.1686–98.

Gregory, P. a et al., 2008. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nature cell biology*, 10(5), pp.593–601.

Greider, C.W. & Blackburn, E.H., 1989. A telomeric sequence in the RNA of Tetrahymena telomerase required for telomere repeat synthesis.

Grivennikov, S.I., Greten, F.R. & Karin, M., 2010. Immunity, Inflammation and Cancer. *Cell*, 140(6), pp.883–899.

Grivna, S.T. et al., 2006. A novel class of small RNAs in mouse spermatogenic cells. *Genes & development*, 20(13), pp.1709–14.

Gröger, C.J. et al., 2012. Meta-analysis of gene expression signatures defining the epithelial to mesenchymal transition during cancer progression. *PloS one*, 7(12), p.e51136.

Guffanti, A. et al., 2009. A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. *BMC genomics*, 10, p.163.

Guo, F. et al., 2010. Inhibition of metastasis-associated lung adenocarcinoma transcript 1 in CaSki human cervical cancer cells suppresses cell proliferation and invasion. *Acta. Biochim. Biophys. Sin.*, 0, pp.224–229.

Gupta, P.B. et al., 2005. The evolving portrait of cancer metastasis. *Cold Spring Harbor symposia on quantitative biology*, 70, pp.291–7.

Gupta, R. a et al., 2010. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, 464(7291), pp.1071–6.

Gupta, R.A. et al., 2011. promote cancer metastasis. , 464(7291), pp.1071–1076.

Gutschner, T. et al., 2013. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer research*, 73(3), pp.1180–9.

Gutschner, T., Diederichs, S. & Rna, K., 2012. A long non-coding RNA point of view. *RNA Biology*, (June), pp.703–719.

Guttman, M. et al., 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature biotechnology*, 28(5), pp.503–10.

Guttman, M. et al., 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458(7235), pp.223–7.

Guttman, M. et al., 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, 477(7364), pp.295–300.

Hagan, J.P. et al., 2009. At least ten genes define the imprinted Dlk1-Dio3 cluster on mouse chromosome 12qF1. *PloS one*, 4(2), p.e4352.

Hahn, W.C. et al., 1999. Creation of human tumour cells with defined genetic elements. *Nature*, 400(6743), pp.464–8.

Hanahan, D. & Folkman, J., 1996. Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. *Cell*, 86(3), pp.353–64.

Hanahan, D. & Weinberg, R. a, 2011. Hallmarks of cancer: the next generation. *Cell*, 144(5), pp.646–74. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21376230 [Accessed July 9, 2014].

Hanahan, D., Weinberg, R.A. & Francisco, S., 2000. The Hallmarks of Cancer Review University of California at San Francisco. , 100, pp.57–70.

Harper, J.W. & Elledge, S.J., 2007. The DNA damage response: ten years after. *Molecular cell*, 28(5), pp.739–45.

Harris, M. a et al., 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*, 32(Database issue), pp.D258–61.

Harrow, J. et al., 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*, 22(9), pp.1760–74.

Hassan, O. et al., 2012. Recent updates on the role of microRNAs in prostate cancer. *Journal of hematology & oncology*, 5(1), p.9.

Hayflick, L., 2003. Living forever and dying in the attempt. *Experimental Gerontology*, 38(11-12), pp.1231–1241.

Hayflick, L., 2000. The illusion of cell immortality. *British journal of cancer*, 83(7), pp.841–6.

He, L. & Hannon, G.J., 2004. MicroRNAs: small RNAs with a big role in gene regulation. *Nature reviews. Genetics*, 5(7), pp.522–31.

Heldin, C.H., Miyazono, K. & ten Dijke, P., 1997. TGF-beta signalling from cell membrane to nucleus through SMAD proteins. *Nature*, 390(6659), pp.465–71.

Hibi, K. et al., 1996. Loss of H19 imprinting in esophageal cancer. *Cancer research*, 56(3), pp.480–2. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8564957.

Hinck, L., Nelson, W. james & Papkoff, J., 1994. Wnt-1 modulates Cell-Cell adhesion in mammalian cells by stabilizinh B-catenin binding to the cell adhesion protein cadherin. *The Journal of cell biology*, 124(5), pp.729–741.

Hirohashi, S., 1998. Inactivation of the E-cadherin-mediated cell adhesion system in human cancers. *The American journal of pathology*, 153(2), pp.333–9.

Howe, L.R. et al., 2003. Twist Is Up-Regulated in Response to Wnt1 and Inhibits Mouse Mammary Cell Differentiation Twist Is Up-Regulated in Response to Wnt1 and Inhibits Mouse Mammary. , pp.1906–1913.

Hsu, P.P. & Sabatini, D.M., 2008. Cancer cell metabolism: Warburg and beyond. *Cell*, 134(5), pp.703–7.

Huang, K. et al., 2002. Relationship of XIST Expression and Responses of Ovarian Cancer to Chemotherapy 1, 59920.

Huang, L. et al., 2014. Overexpression of long noncoding RNA HOTAIR predicts a poor prognosis in patients with cervical cancer. *Archives of gynecology and obstetrics*.

Huang, Y. et al., 2011. Biological functions of microRNAs: a review. *Journal of physiology and biochemistry*, 67(1), pp.129–39.

Huang, Y. et al., 2013. Molecular functions of small regulatory noncoding RNA. *Biochemistry. Biokhimiĭa*, 78(3), pp.221–30.

Huarte, M. et al., 2010. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*, 142(3), pp.409–19.

Hung, T. et al., 2011. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nature genetics*, 43(7), pp.621–9.

Hutchinson, J.N. et al., 2007. A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC genomics*, 8, p.39.

Ishizu, H., Nagao, A. & Siomi, H., 2011. Gatekeepers for Piwi-piRNA complexes to enter the nucleus. *Current opinion in genetics & development*, 21(4), pp.484–90.

Jackson, S.P. & Bartek, J., 2010. The DNA-damage response in human biology and disease. *Nature*, 461(7267), pp.1071–1078.

Jalali, S., Jayaraj, G.G. & Scaria, V., 2012. Integrative transcriptome analysis suggest processing of a subset of long non-coding RNAs to small RNAs. *Biology direct*, 7(1), p.25.

Jemal, A. et al., 2008. Cancer statistics, 2008. *CA: a cancer journal for clinicians*, 58(2), pp.71–96.

Jensen, T.H., Jacquier, A. & Libri, D., 2013. Dealing with pervasive transcription. *Molecular cell*, 52(4), pp.473–84.

Jeon, Y. & Lee, J.T., 2011. YY1 tethers Xist RNA to the inactive X nucleation center. *Cell*, 146(1), pp.119–33.

Ji, P. et al., 2003. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene*, 22(39), pp.8031–41.

Jiang, X.R. et al., 1999. Telomerase expression in human somatic cells does not induce changes associated with a transformed phenotype. *Nature genetics*, 21(1), pp.111–4.

Johnson, A.W., 1997. Rat1p and Xrn1p are functionally interchangeable exoribonucleases that are restricted to and required in the nucleus and cytoplasm.

Jones, R.G. & Thompson, C.B., 2009. Tumor suppressors and cell metabolism : a recipe for cancer growth. (514), pp.537–548.

Junn, E. & Mouradian, M.M., 2010. MicroRNAs in neurodegenerative disorders. *Cell cycle (Georgetown, Tex.)*, 9(9), pp.1717–21.

Kalluri, R. & Weinberg, R.A., 2009. Review series The basics of epithelial-mesenchymal transition. , 119(6).

Kamb, a et al., 1994. A cell cycle regulator potentially involved in genesis of many tumor types. *Science (New York, N.Y.)*, 264(5157), pp.436–40.

Kamijo, T. et al., 1997. Tumor suppression at the mouse INK4a locus mediated by the alternative reading frame product p19ARF. *Cell*, 91(5), pp.649–59.

Kapranov, P. et al., 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science (New York, N.Y.)*, 316(5830), pp.1484–8.

Kapranov, P. et al., 2010. The majority of total nuclear-encoded non-ribosomal RNA in a human cell is "dark matter" un-annotated RNA. *BMC biology*, 8(1), p.149.

Karnoub, A.E. & Weinberg, R. a, 2007. Chemokine networks and breast cancer metastasis. *Breast disease*, 26, pp.75–85.

Kastan, M.B., 2008. DNA damage responses: mechanisms and roles in human disease: 2007 G.H.A. Clowes Memorial Award Lecture. *Molecular cancer research : MCR*, 6(4), pp.517–24.

Khalil, A.M. et al., 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 106(28), pp.11667–72.

Khorkova, O. et al., 2014. Natural antisense transcripts. *Human molecular genetics*.

Kim, D.H. et al., 2006. Argonaute-1 directs siRNA-mediated transcriptional gene silencing in human cells. *Nature structural & molecular biology*, 13(9), pp.793–7.

Kim, K. et al., 2013. HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. *Oncogene*, 32(13), pp.1616–25.

Kim, K.K. et al., 2006. Alveolar epithelial cell mesenchymal transition develops in vivo during pulmonary fibrosis and is regulated by the extracellular matrix. *Proceedings of the National Academy of Sciences of the United States of America*, 103(35), pp.13180–5.

Kim, R., Emi, M. & Tanabe, K., 2007. Cancer immunoediting from immune surveillance to immune escape. *Immunology*, 121(1), pp.1–14.

Kim, T.-K. et al., 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295), pp.182–7.

Kimelman, D. & Xu, W., 2006. Beta-Catenin Destruction Complex: Insights and Questions From a Structural Perspective. *Oncogene*, 25(57), pp.7482–91.

Kino, T. et al., 2010. Noncoding RNA Gas5 is a growth arrest and stravation-associated repressor of the glucocorticoid receptor. *Sci Signal*, 3(107).

Klymkowsky, M.W. & Savagner, P., 2009. Epithelial-mesenchymal transition: a cancer researcher's conceptual friend and foe. *The American journal of pathology*, 174(5), pp.1588–93.

Koch, F. et al., 2008. Genome-wide RNA polymerase II: not genes only! *Trends in biochemical sciences*, 33(6), pp.265–73.

Kokudo, T. et al., 2008. Snail is required for TGFbeta-induced endothelial-mesenchymal transition of embryonic stem cell-derived endothelial cells. *Journal of cell science*, 121(Pt 20), pp.3317–24.

Korkola, J. & Gray, J.W., 2010. Breast cancer genomes - form and function. *Current opinion in genetics & development*, 20(1), pp.4–14.

Korpal, M. et al., 2008. The miR-200 family inhibits epithelial-mesenchymal transition and cancer cell migration by direct targeting of E-cadherin transcriptional repressors ZEB1 and ZEB2. *The Journal of biological chemistry*, 283(22), pp.14910–4.

Kouzarides, T., 2007. Chromatin modifications and their function. *Cell*, 128(4), pp.693–705.

Kugel, J.F. & Goodrich, J. a, 2012. Non-coding RNAs: key regulators of mammalian transcription. *Trends in biochemical sciences*, 37(4), pp.144–51.

Kuiper, R.P., 2003. Upregulation of the transcription factor TFEB in t(6;11)(p21;q13)-positive renal cell carcinomas due to promoter substitution. *Human Molecular Genetics*, 12(14), pp.1661–1669.

Lai, F. et al., 2013. Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature*, 494(7438), pp.497–501.

Larue, L. & Bellacosa, A., 2005. Epithelial-mesenchymal transition in development and cancer: role of phosphatidylinositol 3' kinase/AKT pathways. *Oncogene*, 24(50), pp.7443–54.

Latos, P. a et al., 2012. Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science (New York, N.Y.)*, 338(6113), pp.1469–72.

Lee, J.H. et al., 2014. Highly multiplexed subcellular RNA sequencing in situ. *Science (New York, N.Y.)*, 343(6177), pp.1360–3.

Lee, J.M. et al., 2006. The epithelial-mesenchymal transition: new insights in signaling, development, and disease. *The Journal of cell biology*, 172(7), pp.973–81.

Leighton J. Core, Joshua J. Waterfall, J.T.L., 2008. Nascent RNA Sequencing reveals widespread pausing and divergent initiation at human promoters. , 322(DECEMBER), pp.1845–1848.

Leighton, P. a et al., 1995. An enhancer deletion affects both H19 and Igf2 expression. *Genes & Development*, 9(17), pp.2079–2089.

Li, C.H. & Chen, Y., 2013. Targeting long non-coding RNAs in cancers: progress and prospects. *The international journal of biochemistry & cell biology*, 45(8), pp.1895–910.

Li, V.S.W. et al., 2012. Wnt signaling through inhibition of β-catenin degradation in an intact Axin1 complex. *Cell*, 149(6), pp.1245–56.

Li, W. et al., 2013. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, 498(7455), pp.516–20.

Lim, D.H.K. & Maher, E.R., 2010. *Genomic imprinting syndromes and cancer.* 1st ed., Elsevier Inc.

Lin, M.F., Jungreis, I. & Kellis, M., 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics (Oxford, England)*, 27(13), pp.i275–82.

Liu, X. et al., 2013. The long non-coding RNA HOTAIR indicates a poor prognosis and promotes metastasis in non-small cell lung cancer. *BMC cancer*, 13(1), p.464.

Liu, X.-H. et al., 2014. Lnc RNA HOTAIR functions as a competing endogenous RNA to regulate HER2 expression by sponging miR-331-3p in gastric cancer. *Molecular cancer*, 13(1), p.92.

Lowe, S.W., Cepero, E. & Evan, G., 2004. Intrinsic tumour suppression. *Nature*, 432(7015), pp.307–15.

Lunde, B.M., Moore, C. & Varani, G., 2007. RNA-binding proteins: modular design for efficient function. *Nature reviews. Molecular cell biology*, 8(6), pp.479–90.

Luo, J., Solimini, N.L. & Elledge, S.J., 2009. Principles of cancer therapy: Oncogene and non-oncogene addiction. *Cell*, 136(5), pp.823–837.

Lyashenko, N. et al., 2011. Differential requirement for the dual functions of β-catenin in embryonic stem cell self-renewal and germ layer formation. *Nature cell biology*, 13(7), pp.753–61.

Martens, J. a, Laprade, L. & Winston, F., 2004. Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. *Nature*, 429(6991), pp.571–4.

Martianov, I. et al., 2007. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature*, 445(7128), pp.666–70.

Matouk, I.J. et al., 2014. Oncofetal H19 RNA promotes tumor metastasis. *Biochimica et biophysica acta*, 1843(7), pp.1414–26.

Matouk, I.J. et al., 2007. The H19 non-coding RNA is essential for human tumor growth. *PloS one*, 2(9), p.e845.

Matouk, I.J. et al., 2010. The oncofetal H19 RNA connection: hypoxia, p53 and cancer. *Biochimica et biophysica acta*, 1803(4), pp.443–51.

Matthews, A.J. et al., 2014. *Regulation of immunoglobulin class-switch recombination: choreography of noncoding transcription, targeted DNA deamination, and long-range DNA repair.* 1st ed., Elsevier Inc.

Mattick, J.S., 2009. The genetic signatures of noncoding RNAs. *PLoS genetics*, 5(4), p.e1000459.

Medici, D., Hay, E.D. & Olsen, B.R., 2008. Snail and Slug Promote Epithelial-Mesenchymal Transition through □ -Catenin – T-Cell Factor-4-dependent Expression of Transforming Growth Factor- □ 3. , 19(November), pp.4875–4887.

Meng, W. & Takeichi, M., 2009. Adherens Junction : Molecular. , pp.1–13.

Mercer, T.R. & Mattick, J.S., 2013. Understanding the regulatory and transcriptional complexity of the genome through structure. *Genome research*, 23(7), pp.1081–8.

Miyazawa, K. et al., 2002. Two major Smad pathways in TGF-beta superfamily signalling. *Genes to cells : devoted to molecular & cellular mechanisms*, 7(12), pp.1191–204.

Morales, C.P. et al., 1999. Absence of cancer-associated changes in human fibroblasts immortalized with telomerase. *Nature genetics*, 21(1), pp.115–8.

Mortazavi, A. et al., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. , 5(7), pp.1–8.

Mougiakakos, D. et al., 2010. Regulatory T cells in cancer. *Advances in cancer research*, 107(10), pp.57–117.

Muhlrad, D., Decker, C.J. & Parker, R., 1994. Deadenylation of the unstable mRNA encoded by the yeast MFA2 gene leads to decapping followed by 5'-->3' digestion of the transcript. *Genes & Development*, 8(7), pp.855–866.

Muntoni, F., Torelli, S. & Ferlini, A., 2003. Review Dystrophin and mutations : one gene , several proteins , multiple phenotypes. , 44(0), pp.731–740.

Natoli, G. & Andrau, J.-C., 2012. Noncoding transcription at enhancers: general principles and functional models. *Annual review of genetics*, 46, pp.1–19.

Negrini, S., Gorgoulis, V.G. & Halazonetis, T.D., 2010. Genomic instability--an evolving hallmark of cancer. *Nature reviews. Molecular cell biology*, 11(3), pp.220–8.

Newbury, S. & Woollard, A., 2004. enclosure during C . elegans embryogenesis The 5 – 3 exoribonuclease xrn-1 is essential for ventral epithelial enclosure during C . elegans embryogenesis. , pp.59–65.

Niessen, K. et al., 2008. Slug is a direct Notch target required for initiation of cardiac cushion cellularization. *The Journal of cell biology*, 182(2), pp.315–25.

Ntini, E. et al., 2013. Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nature structural & molecular biology*, 20(8), pp.923–8.

Oft, M., Heider, K.H. & Beug, H., 1998. TGFbeta signaling is necessary for carcinoma cell invasiveness and metastasis. *Current biology : CB*, 8(23), pp.1243–52.

Ogawa, Y., Sun, B.K. & Lee, J.T., 2008. Intersection of the RNA interference and X-inactivation pathways. *Science (New York, N.Y.)*, 320(5881), pp.1336–41.

Okamura, K. & Lai, E.C., 2008. Endogenous small interfering RNAs in animals. *Nature reviews. Molecular cell biology*, 9(9), pp.673–8.

Oom, A.L., Humphries, B. a & Yang, C., 2014. MicroRNAs: Novel Players in Cancer Diagnosis and Therapies. *BioMed research international*, 2014, p.959461.

Orban, T.I. & Izaurralde, E., 2005. Decay of mRNAs targeted by RISC requires XRN1 , the Ski complex , and the exosome. pp.459–469.

Ostrand-Rosenberg, S. & Sinha, P., 2009. Myeloid-derived suppressor cells: linking inflammation and cancer. *J Immunol.*, 182(8), pp.4499–4506.

Pádua Alves, C. et al., 2013. Brief report: The lincRNA Hotair is required for epithelial-to-mesenchymal transition and stemness maintenance of cancer cell lines. *Stem cells (Dayton, Ohio)*, 31(12), pp.2827–32.

Pandey, R.R. et al., 2008. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Molecular cell*, 32(2), pp.232–46.

Pang, K.C., Frith, M.C. & Mattick, J.S., 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends in genetics : TIG*, 22(1), pp.1–5.

Park, S.-M. et al., 2008. The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. *Genes & development*, 22(7), pp.894–907.

Poliseno, L. et al., 2010. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, 465(7301), pp.1033–8.

Polyak, K. & Weinberg, R. a, 2009. Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nature reviews. Cancer*, 9(4), pp.265–73.

Pontier, D.B. & Gribnau, J., 2011. Xist regulation and function explored. *Human genetics*, 130(2), pp.223–36.

Popov, N. & Gil, J., 2010. Epigenetic regulation of the INK4b-ARF-INK4a locus: In sickness and in health. *Epigenetics*, 5(8), pp.685–690.

Prasanth, K. V & Spector, D.L., 2007. Eukaryotic regulatory RNAs: an answer to the "genome complexity" conundrum. *Genes & development*, 21(1), pp.11–42.

Preker, P. et al., 2011. PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic acids research*, 39(16), pp.7179–93.

Preker, P. et al., 2008. RNA Exosome depletion reveals transcription upstream of active human promoters. *Scien*, 322(December), pp.1851–1854.

Rajaram, M. et al., 2013. Two Distinct Categories of Focal Deletions in Cancer Genomes. *PloS one*, 8(6), p.e66264.

Rapicavoli, N. a et al., 2011. The long noncoding RNA Six3OS acts in trans to regulate retinal development by modulating Six3 activity. *Neural development*, 6(1), p.32.

Redon, S., Reichenbach, P. & Lingner, J., 2010. The non-coding RNA TERRA is a natural ligand and direct inhibitor of human telomerase. *Nucleic acids research*, 38(17), pp.5797–806.

Rikitake, Y., Mandai, K. & Takai, Y., 2012. The role of nectins in different types of cell-cell adhesion. *Journal of cell science*, 125(Pt 16), pp.3713–22.

Rinn, J.L. et al., 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129(7), pp.1311–23.

Rinn, J.L. & Chang, H.Y., 2012. Genome regulation by long noncoding RNAs. *Annual review of biochemistry*, 81, pp.145–66.

Rubin, E.M. et al., 2010. Wnt inhibitory factor 1 decreases tumorigenesis and metastasis in osteosarcoma. *Molecular cancer therapeutics*, 9(3), pp.731–41.

Salk, J.J., Fox, E.J. & Loeb, L.A., 2010. Mutational heterogeneity in human cancers: origin and consequences. *Annual review of pathologies*, 5(2), pp.51–75.

Sánchez, Y. & Huarte, M., 2013. Long non-coding RNAs: challenges for diagnosis and therapies. *Nucleic acid therapeutics*, 23(1), pp.15–20.

Sánchez-tilló, E. et al., 2011. ZEB1 to regulate tumor invasiveness.

Santos, J.M. et al., 1999. The stationary-phase morphogene bolA from Escherichia coli is induced by stress during early stages of growth. *Molecular microbiology*, 32(4), pp.789–98.

Sartori, D.A. & Chan, D.W., 2014. Biomarkers in prostate cancer: what's new? *Current opinion in oncology*, 26(3), pp.259–264.

Sauka-Spengler, T. & Bronner-Fraser, M., 2008. A gene regulatory network orchestrates neural crest formation. *Nature reviews. Molecular cell biology*, 9(7), pp.557–68.

Scaruffi, P., 2011. The transcribed-ultraconserved regions: a novel class of long noncoding RNAs involved in cancer susceptibility. *TheScientificWorldJournal*, 11, pp.340–52.

Scaruffi, P. et al., 2009. Transcribed-Ultra Conserved Region expression is associated with outcome in high-risk neuroblastoma. *BMC cancer*, 9, p.441.

Schoeftner, S. & Blasco, M. a, 2010. Chromatin regulation and non-coding RNAs at mammalian telomeres. *Seminars in cell & developmental biology*, 21(2), pp.186–93.

Serrano, M., Hannon, G.J. & Beach, D., 1993. A new regulatory motif in cell-cycle control causing specific inhibition of cyclin D/CDK4. *Nature Publishing Group*, 366.

Shay, J.W. & Bacchetti, S., 1997. A survey of telomerase activity in human cancer. *European journal of cancer (Oxford, England : 1990)*, 33(5), pp.787–91.

Shi, Y. & Massagué, J., 2003. Mechanisms of TGF-beta signaling from cell membrane to the nucleus. *Cell*, 113(6), pp.685–700.

Shields, J.D. et al., 2010. Induction of lymphoidlike stroma and immune escape by tumors that express the chemokine CCL21. *Science (New York, N.Y.)*, 328(5979), pp.749–52.

Shigunov, P. et al., 2012. PUMILIO-2 is involved in the positive regulation of cellular proliferation in human adipose-derived stem cells. *Stem cells and development*, 21(2), pp.217–27.

Sigl, R. et al., 2009. Loss of the mammalian APC/C activator FZR1 shortens G1 and lengthens S phase but has little effect on exit from mitosis. *Journal of cell science*, 122(Pt 22), pp.4208–17.

Sirchia, S.M. et al., 2005. Loss of the inactive X chromosome and replication of the active X in BRCA1-defective and wild-type breast cancer cells. *Cancer research*, 65(6), pp.2139–46.

Sirchia, S.M. et al., 2009. Misbehaviour of XIST RNA in breast cancer cells. *PloS one*, 4(5), p.e5559.

Sleutels, F., Zwart, R. & Barlow, D.P., 2002. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, 415(6873), pp.810–3.

Smit, M. a & Peeper, D.S., 2008. Deregulating EMT and senescence: double impact by a single twist. *Cancer cell*, 14(1), pp.5–7.

Song, J., 2007. EMT or apoptosis: a decision for TGF-beta. *Cell research*, 17(4), pp.289–90.

Spano, D. et al., 2012. Molecular networks that regulate cancer metastasis. *Seminars in cancer biology*, 22(3), pp.234–49.

Stefani, G. & Slack, F.J., 2008. Small non-coding RNAs in animal development. *Nature reviews. Molecular cell biology*, 9(3), pp.219–30.

Stockinger, a et al., 2001. E-cadherin regulates cell growth by modulating proliferation-dependent beta-catenin transcriptional activity. *The Journal of cell biology*, 154(6), pp.1185–96.

Su, W.-Y., Xiong, H. & Fang, J.-Y., 2010. Natural antisense transcripts regulate gene expression in an epigenetic manner. *Biochemical and biophysical research communications*, 396(2), pp.177–81.

Tachibana, M. et al., 2005. Histone methyltransferases G9a and GLP form heteromeric complexes and are both crucial for methylation of euchromatin at H3-K9. *Genes & development*, 19(7), pp.815–26.

Talmadge, J.E. & Fidler, I.J., 2010. AACR Centennial series: the biology of cancer metastasis: historical perspective. *Cancer research*, 70(14), pp.5649–5669.

Tam, O.H. et al., 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, 453(7194), pp.534–8.

Tang, F., 2010. Small RNAs in mammalian germline: Tiny for immortal. *Differentiation; research in biological diversity*, 79(3), pp.141–6.

Tano, K. et al., 2010. MALAT-1 enhances cell motility of lung adenocarcinoma cells by influencing the expression of motility-related genes. *FEBS letters*, 584(22), pp.4575–80.

Tarin, D., 2005. The Fallacy of Epithelial Mesenchymal Transition in Neoplasia. *Cancer research*, pp.5996–6001.

Tayade, C. et al., 2005. Functions of alpha 2 macroglobulins in pregnancy. *Molecular and cellular endocrinology*, 245(1-2), pp.60–6.

Tellez, C.S. et al., 2011. EMT and stem cell-like properties associated with miR-205 and miR-200 epigenetic silencing are early manifestations during carcinogen-induced transformation of human lung epithelial cells. *Cancer research*, 71(8), pp.3087–97.

Teng, G. & Papavasiliou, F.N., 2009. *Long noncoding RNAs: implications for antigen receptor diversification.* First edit., Elsevier.

Teng, M.W.L. et al., 2008. Immune-mediated dormancy: an equilibrium with cancer. *Journal of leukocyte biology*, 84(4), pp.988–93.

Tepass, U. et al., 2000. Cadherins in embryonic and neural morphogenesis. *Nature reviews. Molecular cell biology*, 1(November), pp.1–10.

Thiery, J.P., 2002. Epithelial-mesenchymal transitions in tumour progression. *Nature reviews. Cancer*, 2(6), pp.442–54.

Thomson, T. & Lin, H., 2009. The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. *Annual review of cell and developmental biology*, 25, pp.355–76.

Till, D.D. et al., 1998. Identification and developmental expression of a 5'-3' exoribonuclease from Drosophila melanogaster. *Mechanisms of development*, 79(1-2), pp.51–5.

Ting, A.H. et al., 2005. Short double-stranded RNA induces transcriptional gene silencing in human cancer cells in the absence of DNA methylation. *Nature genetics*, 37(8), pp.906–10.

Tisseur, M., Kwapisz, M. & Morillon, A., 2011. Pervasive transcription - Lessons from yeast. *Biochimie*, 93(11), pp.1889–96.

Tong, Y.-K. & Lo, Y.M.D., 2006. Diagnostic developments involving cell-free (circulating) nucleic acids. *Clinica chimica acta; international journal of clinical chemistry*, 363(1-2), pp.187–96.

Triantafillidis, J.K., Nasioulas, G. & Kosmidis, P. a, 2009. Colorectal cancer and inflammatory bowel disease: epidemiology, risk factors, mechanisms of carcinogenesis and prevention strategies. *Anticancer research*, 29(7), pp.2727–37.

Tripathi, V. et al., 2010. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular cell*, 39(6), pp.925–38.

Tsai, M.-C. et al., 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science (New York, N.Y.)*, 329(5992), pp.689–93.

Tsang, K.M. et al., 2013. A genome-wide survey of transgenerational genetic effects in autism. *PloS one*, 8(10), p.e76978.

Tsang, W.P. et al., 2010. Oncofetal H19-derived miR-675 regulates tumor suppressor RB in human colorectal cancer. *Carcinogenesis*, 31(3), pp.350–8.

Tufo, G. et al., 2014. The protein disulfide isomerases PDIA4 and PDIA6 mediate resistance to cisplatin-induced cell death in lung adenocarcinoma. *Cell death and differentiation*, 21(5), pp.685–95.

Ulitsky, I. & Bartel, D.P., 2013. lincRNAs: genomics, evolution, and mechanisms. *Cell*, 154(1), pp.26–46.

Valenta, T., Hausmann, G. & Basler, K., 2012. The many faces and functions of β-catenin. *The EMBO journal*, 31(12), pp.2714–36.

Vincent-Salomon, A. et al., 2007. X inactive-specific transcript RNA coating and genetic instability of the X chromosome in BRCA1 breast tumors. *Cancer research*, 67(11), pp.5134–40.

Vogelstein, B. & Kinzler, K.W., 2004. Cancer genes and the pathways they control. *Nature medicine*, 10(8), pp.789–99.

Wang, J. et al., 2010. CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. *Nucleic acids research*, 38(16), pp.5366–83.

Wang, X. et al., 2008. Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature*, 454(7200), pp.126–30.

Wang, Z., Gerstein, M. & Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1), pp.57–63.

Watanabe, T. et al., 2008. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, 453(7194), pp.539–43.

Weinberg, R. a, 1995. The retinoblastoma protein and cell cycle control. *Cell*, 81(3), pp.323–30.

Weinberg, R. a, 2008. Twisted epithelial-mesenchymal transition blocks senescence. *Nature cell biology*, 10(9), pp.1021–3.

Wilusz, J.E., Sunwoo, H. & Spector, D.L., 2009. Long noncoding RNAs: functional surprises from the RNA world. *Genes & development*, 23(13), pp.1494–504.

Wright, W.E., Pereira-Smith, O.M. & Shay, J.W., 1989. Reversible cellular senescence: implications for immortalization of normal human diploid fibroblasts. *Molecular and cellular biology*, 9(7), pp.3088–92.

Yang, F. et al., 2014. Long non-coding RNA GHET1 promotes gastric carcinoma cell proliferation by increasing c-Myc mRNA stability. *The FEBS journal*, 281(3), pp.802–13.

Yang, J. et al., 2004. Twist , a Master Regulator of Morphogenesis , Plays an Essential Role in Tumor Metastasis Ben Gurion University of the Negev. , 117, pp.927–939.

Yang, J., Mani, S. a & Weinberg, R. a, 2006. Exploring a new twist on tumor metastasis. *Cancer research*, 66(9), pp.4549–52.

Yang, J. & Weinberg, R. a, 2008. Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis. *Developmental cell*, 14(6), pp.818–29.

Yang, L., Lin, C. & Liu, Z.-R., 2006. P68 RNA helicase mediates PDGF-induced epithelial mesenchymal transition by displacing Axin from beta-catenin. *Cell*, 127(1), pp.139–55.

Yang, L., Pang, Y. & Moses, H.L., 2010. TGF-B and immune cells: an important regulatory axis in the tumor microenvironment and progression. *Trends Immunol.*, 31(6), pp.220–227.

Yang, Z. et al., 2011. Overexpression of long non-coding RNA HOTAIR predicts tumor recurrence in hepatocellular carcinoma patients following liver transplantation. *Annals of surgical oncology*, 18(5), pp.1243–50.

Yang, Z. et al., 2007. Up-regulation of gastric cancer cell invasion by Twist is accompanied by N-cadherin and fibronectin expression. *Biochemical and biophysical research communications*, 358(3), pp.925–30.

Yap, K.L. et al., 2010. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Molecular cell*, 38(5), pp.662–74.

Yilmaz, M. & Christofori, G., 2009. EMT, the cytoskeleton, and cancer cell invasion. *Cancer metastasis reviews*, 28(1-2), pp.15–33.

Yoon, J.-H. et al., 2013. Scaffold function of long non-coding RNA HOTAIR in protein ubiquitination. *Nature communications*, 4, p.2939.

Yu, X. et al., 2013. Homocysteine inhibits hepatocyte proliferation via endoplasmic reticulum stress. *PloS one*, 8(1), p.e54265.

Yuan, J. et al., 2014. A long noncoding RNA activated by TGF-β promotes the invasion-metastasis cascade in hepatocellular carcinoma. *Cancer cell*, 25(5), pp.666–81.

Zaidi, S.K. et al., 2010. Architectural epigenetics: mitotic retention of mammalian transcriptional regulatory information. *Molecular and cellular biology*, 30(20), pp.4758–66.

Zeisberg, E.M., Potenta, S., et al., 2007. Discovery of endothelial to mesenchymal transition as a source for carcinoma-associated fibroblasts. *Cancer research*, 67(21), pp.10123–8.

Zeisberg, E.M., Tarnavski, O., et al., 2007. Endothelial-to-mesenchymal transition contributes to cardiac fibrosis. *Nature medicine*, 13(8), pp.952–61.

Zeisberg, M. et al., 2007. Fibroblasts derive from hepatocytes in liver fibrosis via epithelial to mesenchymal transition. *The Journal of biological chemistry*, 282(32), pp.23337–47.

Zeisberg, M., Shah, A. a & Kalluri, R., 2005. Bone morphogenic protein-7 induces mesenchymal to epithelial transition in adult renal fibroblasts and facilitates regeneration of injured kidney. *The Journal of biological chemistry*, 280(9), pp.8094–100.

Zentner, G.E. & Henikoff, S., 2013. Regulation of nucleosome dynamics by histone modifications. *Nature structural & molecular biology*, 20(3), pp.259–66.

Zhang, K. et al., 2002. The human homolog of yeast SEP1 is a novel candidate tumor suppressor gene in osteogenic sarcoma. *Gene*, 298(2), pp.121–7.

Zhang, X. et al., 2003. A pituitary-derived MEG3 isoform functions as a growth suppressor in tumor cells. *The Journal of clinical endocrinology and metabolism*, 88(11), pp.5119–26.

Zhao, Y. et al., 2014. Role of long non-coding RNA HULC in cell proliferation, apoptosis and tumor metastasis of gastric cancer: a clinical and in vitro investigation. *Oncology reports*, 31(1), pp.358–64.

Zhou, Y. et al., 2007. Activation of p53 by MEG3 non-coding RNA. *The Journal of biological chemistry*, 282(34), pp.24731–42.

ANNEXES

Supplementary Table S1 | Primer sequences.

Target	Direction	Sequence 5'->3'	Taken from
7SL	Frw	GGAGTTCTGGGCTGTAGTGC	
	Rev	ATCAGCACGGGAGTTTTGAC	
ABL2	Frw	GGACACTTCACTTTGCTGCC	
	Rev	TAGTGCCTGGGGTTCAACAT	
ACTA2	Frw	TCAATGTCCCAGCCATGTAT	Lee et al, Clinical Cancer Res. 2013
	Rev	CAGCACGATGCCAGTTGT	
AL589743.1	Rev	CTCTGGCAGGTGATGAGAT	
	Frw	TCTGGATCTTGTCATTGTGA	
attB1-GFP	Frw	GGGGACAAGTTTGTACAAAAAAGCAGGCTAAAGGAGAAGAACTTTTCAC	
attB2-GFP	Rev	GGGGACCACTTTGTACAAGAAAGCTGGGTTTTGTATAGTTCATCCATGC	
attB1-HOTAIR 5'	Frw	GGGGACAAGTTTGTACAAAAAAGCAGGCTGACTCGCCTGTGCTCTGGAGC TTGATCCGA	
attB2-HOTAIR 3'	Rev	GGGGACCACTTTGTACAAGAAAGCTGGGTTTTTTTTTTTGAAAATGCATCCA GATATTA	
attB1-HOTAIRΔPRC2	Frw	GGGGACAAGTTTGTACAAAAAAGCAGGCTTCTTTATTTTTTTAAGGCC	
attB2-HOTAIRΔPRC2	Rev	GGGGACCACTTTGTACAAGAAAGCTGGGTTTATATTCACCACATGTAAAA	
attB1-HOTAIRΔLSD1	Frw	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCAGTTCTCAGGCGAG	
attB2-HOTAIRΔLSD1	Rev	GGGGACCACTTTGTACAAGAAAGCTGGGTTGGTTTCACTTTTAAAAATTT	
B2MG	Frw	TGCTGTCTCCATGTTTGATGTATCT	
	Rev	TCTCTGCTCCCCACCTCTAAGT	
BDNF	Frw	TATGAATCGCCAGCCAATTCT	
	Rev	AGTGCCGAACTACCCAGTCGTA	
BTRC as-lncRNA	Frw	TAGATTTTTGGAGCAGGGAG	
	Rev	GAGGTACTGCAATACCAGG	
CDHR1	Frw	TGAGTAAGGCTGTGGTGCTG	
	Rev	TATTCCATCACCCATGCAGA	
CLDN2	Frw	TCCCACTGACTGACCCTCTGT	
	Rev	GCCACTGCTTCTCCTTCCCAT	
CTD-2314B22.3	Frw	ACTTCTTCTGCCACTGTCAC	
	Rev	GATGGATTTTGGGTATTTCA	
CTNNB1	Frw	TGCAGTTCGCCTTCACTATG	
	Rev	ACTAGTCGTGGAATGGCACC	
DCAF6	Rev	ATTGTCCTGGAGTTGCGATG	
	Frw	AATGAGACGCTCTGCTGTTG	
DCAF6 as-lncRNA	Frw	TGCGTCTTTTTCACAATATCC	
	Rev	TTCATCCCAAATCCAGGTATG	
DCAF6 exon 15	Frw	ATATCAGGAAGGAGTATCTG	
DCAF6 exon 16	Frw	AGCCTCAGTTCCAAACAGAA	
DCAF6 exon 16	Rev	ACAAGAGCGATCAAGATTGA	
DCAF6 exon 16	Frw	GTGCAAGGTATCGAGCAGGA	
DCAF6 exon 17	Rev	CTATCACCTATTGTTGTTCC	
DCAF6 exon 17	Frw	GGAACAACAATAGGTGATAG	
DCAF6 exon 18	Rev	TGGTACCATTGTCCTGGAGT	
DCAF6 exon 18	Frw	GCCCTATAATATATTGTG	
DCAF6 exon 19	Rev	CAGCTTCCAGAAGCATCAAA	
DCAF6 intron 17-18 +12362	Frw	GTAGTAAAATTGCACATGAA	
	Rev	AACGCAGACACAACAGTGAGA	
DCAF6 intron 17-18 +143	Frw	GGTCGCTTTCCACAAACTAT	
	Rev	GCATTCTTCAACAAAATATC	
DCAF6 intron 17-18 +2286	Frw	CACAGTGCATTGTATATTTG	
	Rev	TGCAGGGAGATATGAATAAA	
DCAF6 intron 17-18 +4202	Frw	TGCGTCTTTAACACATACCCA	
	Rev	AGATTGAAGATCTAGGTTCT	

DCAF6 intron 17-18 +5646	Frw	CTGTTGTCCAAAGTGTTCTG	
	Rev	TGATAAAGTAAGTGTGGAAA	
DCAF6 intron 17-18 +7650	Frw	TGTTTTGAGATGGAGTCTT	
	Rev	TGAAACCCCATCTCTACTA	
DCAF6 intron 17-18 +9546	Frw	CTGGCCTAGAGCATACTTTT	
	Rev	ACCTCAGCTTCCCAAAGT	
DCAF6 promoter	Frw	TTGCAGGAGAGAGGGTGAAAT	
	Rev	GTTGTACAACCATCCAGCAG	
DMD as-lncRNA	Frw	GATGCGGTGGCTCACACCTG	
	Rev	GCTACAGGCATGCACCAC	
FAT3	Frw	GAGACACTGTCCCCATTAGA	
	Rev	AAGTACACTGCATCCTCACC	
FERMT1 as-lncRNA	Frw	AAGCAAAGTTACTGGCAGAG	
	Rev	CAAGCACCTTCAGATATTCC	
FN1	Frw	GGTTTCCCATTATGCCATTG	
	Rev	TTCCAAGACATGTGCAGCTC	
FZR1 as-lncRNA	Frw	TGGGGTTGACACCATCAGAC	
	Rev	GACCGTCCTTGTTGCTCTGT	
GAPDH	Frw	TCCTGCACCACCAACTGCTTAGC	
	Rev	TGATGTCATCATACTTGGCAG	
GATA2	Frw	AAGCTGCTGTGACTGTGTCC	
	Rev	TTACAGGGTCCACCTGACAA	
GPR1	Frw	AATGCCATCGTCATTTGGTT	
	Rev	CAACTGGGCAGTGAAGGAAT	
HAVCR1	Frw	GGAAGGACACACGCTATAAG	
	Rev	CTCCAATGATACGGTGATTT	
HOTAIR	Frw	GGTAGAAAAAGCAACCACGAAGC	Gupta et al., Nature 2010
	Rev	ACATAAACCTCTGTCTGTGAGTGCC	
HOXD10	Frw	GCTGAGGCGCTTTAATGAAC	Gupta et al., Nature 2010
	Rev	GGTCCCAGAAACTCTGACCA	
HSP90B1	Frw	ACCTGCTGCATGTCACAGAC	
	Rev	AGAAACCGACACCAAACTGG	
IL6R as-lncRNA	Frw	GTAAAAGCCCCAATCTTAGC	
	Rev	CCCGATCTTCAGTATGTCAC	
IL7R	Frw	CCCTCGTGGAGGTAAAGT	
	Rev	TTTGCAGGTTAGACTCTTTTC	
INHBE	Frw	CTGAGCACCACATCACCAAC	
	Rev	TGTCCTGCTGTGTCCAAGAG	
JUB	Frw	TATCCATGGACCGGGATTAT	
	Rev	CAGCCTTCCTCATCACTCAG	
KRT19	Frw	TGAGACGGAACAGGCTCT	
	Rev	CTCATGGTTCTTCTTCAGGT	
LAMB3	Frw	GCCACATTCTCTACTCGGTGA	
	Rev	CCAAGCCTGAGACCTACTGC	
LAMC2	Frw	CTCTGCTTCTCGCTCCTCC	
	Rev	TCTGTGAAGTTCCCGATCAA	
MACROD2 as-lncRNA	Frw	GAGCTGTACGTGGATAGGAG	
	Rev	CTGAGGAAGGTCAGTTGAAG	
MALAT1	Frw	GAATTGCGTCATTTAAAGCCTAGTT	Ji et al., Oncogene 2003
	Rev	GTTTCATCCTACCACTCCCAATTAAT	
MYO1D	Frw	GTCGTTTCTGTGAACCCTTA	
	Rev	ATAGCCTTGTAAGCAGCATC	
NUPR1	Frw	CCAGATGAGGCCAGACCAT	
	Rev	GAGACTCAGTCAGCGGGAAT	
OCLN	Frw	CAGGACGTGCCTTCACCCCC	
	Rev	CCACCGCTGCTGTAACGAGGC	
PCDH10	Frw	CCCGTCTACACTGTGTCCCT	
	Rev	GGAGTACACGACCTCACCGT	
PCDHB5	Frw	AGGTGTGTTTGACCGGAGAC	
	Rev	TCCCTATTTCTTCACCAGCG	
PCNP as-lncRNA	Frw	CACCGTTATGCAGAATCCAC	
	Rev	AGAAGTGACGCAGCCCTCTA	
PDIA4	Frw	CTTGGCAGGAAAACTGAAGC	
	Rev	TTGGCCCTTGTACTTCTTGG	
POLR2F	Frw	ATGTCGAGATCCTCCCCTCTG	
	Rev	GGCCTTGAGTTCCTTCATGG	

PVRL3 AS1	Frw	CTCTTAGGCCAAGAAGAACA	
	Rev	CAAAAGATGTACTGATCTCAGATTT	
PZP as-lncRNA	Frw	TTGCTTTTGTTGCAGTTGCT	
	Rev	ATTGAAACTGGACCCCTTCC	
RAB3GAP2 as-lncRNA	Frw	GGTTTAATTGATCAAAGACT	
	Rev	CCTTATATGAGGTTGCACTA	
RPL11	Frw	AGCAGCCAAGGTGTTGGAG	
	Rev	TACTCCCGCACCTTTAGACC	
SIRT2	Frw	CCTCGCCAAGGAACTCTATC	
	Rev	GTGTAGCAGCGCAGGAGTAG	
SLC9A3 as-lncRNA	Frw	TCTCCTGAATTTTGTTTGCT	
	Rev	TGCAGTTTCATATGCGTCTA	
SNAIL	Frw	TGACCTGTCTGCAAATGCTC	
	Rev	CAGACCCTGGTTGCTTCAA	
TFAP2A	Frw	AGGGACTTTGGGTACGTGTG	
	Rev	TTGTAGCCAGGAGCATGTTTT	
TJP3	Frw	GCCAGTTTCAAGCGCCCGGT	
	Rev	TCTGCAATCACCCGCACGGTG	
TRIB3	Frw	ATGATTCCCTGTGGGACAAG	
	Rev	AGTCCTGGAAGGGGTAGTGG	
U1	Frw	CCATGATCACGAAGGTGGTTT	
	Rev	ATGCAGTCGAGTTTCCCACAT	
U6	Frw	CTCGCTTCGGCAGCACA	
	Rev	AACGCTTCACGAATTTGCGT	
VIM	Frw	TTCCAAGCCTGACCTCACGGCTG	
	Rev	TTCCGGTTGGCAGCCTCAGAGA	
XRN1	Frw	AGTCCCCAAGTTTTACAGATG	
	Rev	GAAATTCTAAAGTGAACATCATC	
ZEB2	Frw	TATGGCCTACACCTACCCAAC	Beltran et al., Genes Dev 2008
	Rev	AGGCCTGACATGTAGTCTTGTG	
ZFHX4 AS1	Frw	TTGATTTTTCTCATGCCTCT	
	Rev	TATTGATTCCCCAATTTGTC	

Supplementary Table S2 | List of genes identified as differentially expressed between HEK-Epi and HEK-Mes cells, and present in the EMT core list of genes established by Gröger *et al.*

	Number of genes	Names
Upregulated in HEK-Mes cells, in line with Gröger list	47	ACTA2, ADM, ARID5B, BAMBI, CD44, COL5A1, COL6A1, CTGF, ELL2, F3, FBLN1, FBN1, FN1, FST, HBEGF, HMGA2, HTRA1, IGFBP7, IL18, IL7R, INHBA, JUP, KRT17, KRT81, LOX, MYL9, MYO10, NAV3, NID2, NR2F1, PMP22, POLR3G, PPAP2B, PTX3, RGS4, SERPINE2, SFRP1, SMAD3, SRCAP, SRGN, SRPX2, TBX3, THBS1, TNC, TRIM29, SERPINE2, SFRP1
Downregulated in HEK-Mes cells, in line with Gröger list	62	ANK3, BDKRB2, CDS1, CNTN1, COBLL1, CXCL1, DAPK1, DPYSL3, DSC2, ELF3, EMP3, ENPP1, EPCAM, FERMT1, FGFR2, GLS, HS3ST3A1, IFI44, IFI44L, IFI6, IFIT1, IFITM1, IGFBP2, ITGA2, KCNK1, LAD1, LAMC2, LSR, LTBP1, MAN1A1, MAP7, MBP, MFAP2, MME, MTUS1, OAS1, PEG10, PKP2, PLA2G16, PLA2G4A, PLS1, PPL, PRUNE2, RHOD, S100A2, SEMA3A, SERPINA1, SFN, SLC39A8, SLC6A15, SMPDL3B, SYT11, SYTL2, TCF4, TGFA, TLR3, TM4SF1, TPD52L1, TUBA1A, VCAN, DAPK1, LSR

Supplementary Table S3 | siRNA and ASO sequences.

Name	Reference	From	Sequence
siXRN1-1	HSS122909	Life Technologies	-
siXRN1-2	HSS182510	Life Technologies	-
siXRN1-3	HSS182511	Life Technologies	-
siCTR	AM4611	Life Technologies	-
siHOT-1	-	Sigma-Aldrich	GAACGGGAGUACAGAGAGAUU
siHOT-2	-	Sigma-Aldrich	CCACAUGAACGCCCAGAGAUU
siHOT-3	-	Sigma-Aldrich	UAACAAGACCAGAGAGCUGUU
siGFP-CTR	-	Sigma-Aldrich	CUACAACAGCCACAACGUCDTDT
ASO HOT-1	520121	ISIS Pharmaceuticals	CATCACTTATTTAAGTGTTC
ASO HOT-2	520131	ISIS Pharmaceuticals	GCACAGAAAATGCATCCAGA
ASO HOT-3	520167	ISIS Pharmaceuticals	CCGCTCAGGTTTTTCCAGCG
ASO HOT-4	520174	ISIS Pharmaceuticals	CATGGGTTCCGTGTAGACGC
ASO HOT-5	520196	ISIS Pharmaceuticals	GTCCCACTGCATAATCACTC
ASO CTR	141923	ISIS Pharmaceuticals	CCTTCCCTGAAGGTTCCTCC

Supplementary Table S4 | Plasmids references.

Plasmid name	Application	References
pLZRS-HOTAIR	Gateway cloning	Gupta et al. 2008
psPAX2	Cell transduction	12260 - Addgene
pCMV-VSV-G	Cell transduction	8454 - Addgene
pLenti6.2/V5-DEST	Gateway cloning	Life Technologies
pDONR201	Gateway cloning	Life Technologies
pFA6a-GFP(S65T)-His3MX6	Gateway cloning	Longtine et al. 1998
pHOT	HOTAIR overexpression	This study
pHOTΔPRC2	HOTAIR overexpression	This study
pHOTΔLSD1	HOTAIR overexpression	This study
pGFP	HOTAIR overexpression	This study

Supplementary Table S5 | Complete list of GO Terms and KEGG pathways, HOTAIR full-length vs CTR.

term ID	description	pval
	GO TERMS	
GO:0007267	cell-cell signaling	6,33E-06
GO:0042127	regulation of cell proliferation	5,70E-05
GO:0030334	regulation of cell migration	1,33E-04
GO:0040012	regulation of locomotion	1,68E-04
GO:0045787	positive regulation of cell cycle	2,36E-04
GO:0006954	inflammatory response	5,95E-04
GO:0007610	behavior	6,55E-04
GO:0044057	regulation of system process	7,82E-04
GO:0008361	regulation of cell size	1,08E-03
GO:0001525	angiogenesis	1,36E-03
GO:0009611	response to wounding	1,86E-03
GO:0031667	response to nutrient levels	1,98E-03
GO:0010033	response to organic substance	2,29E-03
GO:0040008	regulation of growth	2,58E-03
GO:0007584	response to nutrient	2,73E-03
GO:0006812	cation transport	3,45E-03
GO:0035556	intracellular signal transduction	4,14E-03
GO:0015849	organic acid transport	4,18E-03
GO:0001932	regulation of protein phosphorylation	4,95E-03
GO:0007611	learning or memory	5,26E-03
GO:0009991	response to extracellular stimulus	5,36E-03
GO:0007167	enzyme linked receptor protein signaling pathway	5,77E-03
GO:0006811	ion transport	6,16E-03
GO:0006952	defense response	7,73E-03
GO:0015837	amine transport	7,80E-03
GO:0000165	MAPK cascade	7,99E-03
GO:0031960	response to corticosteroid stimulus	1,22E-02
GO:0002237	response to molecule of bacterial origin	1,29E-02
GO:0042592	homeostatic process	1,29E-02
GO:0009968	negative regulation of signal transduction	1,33E-02
GO:0016055	Wnt receptor signaling pathway	1,63E-02
GO:0034976	response to endoplasmic reticulum stress	1,68E-02
GO:0010827	regulation of glucose transport	1,68E-02
GO:0006984	ER-nucleus signaling pathway	1,85E-02
GO:0006813	potassium ion transport	1,93E-02
GO:0001503	ossification	1,97E-02
GO:0008037	cell recognition	2,20E-02
GO:0030968	endoplasmic reticulum unfolded protein response	2,24E-02
GO:0006928	cellular component movement	2,25E-02
GO:0007243	intracellular protein kinase cascade	2,45E-02
GO:0042493	response to drug	2,52E-02
GO:0060348	bone development	2,81E-02
GO:0001890	placenta development	3,07E-02
GO:0045785	positive regulation of cell adhesion	3,07E-02
GO:0002790	peptide secretion	3,12E-02
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	3,21E-02
GO:0010817	regulation of hormone levels	3,35E-02
GO:0007265	Ras protein signal transduction	3,44E-02
GO:0006836	neurotransmitter transport	3,46E-02
	KEGG Pathways	
hsa04060	Cytokine-cytokine receptor interaction	2,32E-02
hsa04010	MAPK signaling pathway	9,51E-02
hsa04810	Regulation of actin cytoskeleton	9,17E-02
hsa04916	Melanogenesis	1,50E-02
hsa04514	Cell adhesion molecules (CAMs)	6,58E-02
hsa05217	Basal cell carcinoma	8,53E-03
hsa05213	Endometrial cancer	2,63E-02
hsa05214	Glioma	5,37E-02
hsa04610	Complement and coagulation cascades	7,37E-02
hsa05218	Melanoma	8,11E-02

Supplementary Table S6 | Complete list of GO Terms and KEGG pathways, truncated HOTAIR forms.

a. HOTΔLSD1

term ID	description	pval
	GO terms	
GO:0040008	regulation of growth	1,06E-05
GO:0008285	negative regulation of cell proliferation	1,66E-05
GO:0001666	response to hypoxia	1,95E-05
GO:0030193	regulation of blood coagulation	2,53E-05
GO:0050880	regulation of blood vessel size	3,99E-05
GO:0050878	regulation of body fluid levels	5,43E-05
GO:0030199	collagen fibril organization	6,31E-05
GO:0042060	wound healing	6,72E-05
GO:0044275	cellular carbohydrate catabolic process	7,51E-05
GO:0006928	cellular component movement	8,03E-05
GO:0022602	ovulation cycle process	8,95E-05
GO:0046164	alcohol catabolic process	1,12E-04
GO:0051789	response to protein stimulus	1,25E-04
GO:0008544	epidermis development	1,52E-04
GO:0007398	ectoderm development	1,63E-04
GO:0030182	neuron differentiation	1,70E-04
GO:0001974	blood vessel remodeling	2,24E-04
GO:0005996	monosaccharide metabolic process	3,70E-04
GO:0010243	response to organic nitrogen	3,75E-04
GO:0003013	circulatory system process	4,08E-04
GO:0016052	carbohydrate catabolic process	4,39E-04
GO:0001822	kidney development	4,62E-04
GO:0030155	regulation of cell adhesion	4,66E-04
GO:0031400	negative regulation of protein modification process	6,66E-04
GO:0045137	development of primary sexual characteristics	7,56E-04
GO:0006000	fructose metabolic process	7,98E-04
GO:0009719	response to endogenous stimulus	8,05E-04
GO:0051258	protein polymerization	8,33E-04
GO:0044092	negative regulation of molecular function	8,92E-04
GO:0042981	regulation of apoptotic process	9,19E-04
GO:0031399	regulation of protein modification process	9,53E-04
GO:0048771	tissue remodeling	1,01E-03
GO:0050817	coagulation	1,07E-03
GO:0031099	regeneration	1,10E-03
GO:0010811	positive regulation of cell-substrate adhesion	1,19E-03
GO:0006936	muscle contraction	1,25E-03
GO:0003012	muscle system process	1,28E-03
GO:0006986	response to unfolded protein	1,52E-03
GO:0003006	developmental process involved in reproduction	1,53E-03
GO:0006690	icosanoid metabolic process	1,87E-03
GO:0009968	negative regulation of signal transduction	2,43E-03
GO:0040007	growth	2,55E-03
GO:0009628	response to abiotic stimulus	2,78E-03
GO:0009069	serine family amino acid metabolic process	3,04E-03
GO:0043062	extracellular structure organization	3,27E-03
GO:0009409	response to cold	3,33E-03
GO:0042592	homeostatic process	3,74E-03
GO:0033559	unsaturated fatty acid metabolic process	3,91E-03
GO:0006775	fat-soluble vitamin metabolic process	4,21E-03
GO:0051781	positive regulation of cell division	4,25E-03
GO:0006575	cellular modified amino acid metabolic process	4,26E-03
GO:0010817	regulation of hormone levels	4,34E-03
GO:0009064	glutamine family amino acid metabolic process	4,63E-03
GO:0048872	homeostasis of number of cells	4,66E-03
GO:0001503	ossification	4,91E-03
GO:0060348	bone development	5,24E-03
GO:0051351	positive regulation of ligase activity	5,35E-03
GO:0051091	positive regulation of sequence-specific DNA binding transcription factor activity	5,79E-03
GO:0050886	endocrine process	5,87E-03
GO:0046942	carboxylic acid transport	5,99E-03
GO:0007584	response to nutrient	6,32E-03

GO:0007050	cell cycle arrest	6,48E-03
GO:0051090	regulation of sequence-specific DNA binding transcription factor activity	6,48E-03
GO:0015849	organic acid transport	6,54E-03
GO:0000302	response to reactive oxygen species	6,94E-03
GO:0007267	cell-cell signaling	7,08E-03
GO:0009991	response to extracellular stimulus	7,25E-03
GO:0031667	response to nutrient levels	7,76E-03
GO:0010035	response to inorganic substance	7,78E-03
GO:0046425	regulation of JAK-STAT cascade	7,82E-03
GO:0048511	rhythmic process	8,43E-03
GO:0042325	regulation of phosphorylation	8,52E-03
GO:0042445	hormone metabolic process	8,85E-03
GO:0045765	regulation of angiogenesis	8,86E-03
GO:0030218	erythrocyte differentiation	8,86E-03
GO:0043388	positive regulation of DNA binding	8,94E-03
GO:0044236	multicellular organismal metabolic process	9,43E-03
GO:0046364	monosaccharide biosynthetic process	9,43E-03
GO:0044259	multicellular organismal macromolecule metabolic process	9,76E-03
GO:0019319	hexose biosynthetic process	9,76E-03
GO:0007167	enzyme linked receptor protein signaling pathway	1,05E-02
GO:0002443	leukocyte mediated immunity	1,10E-02
GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	1,15E-02
GO:0018149	peptide cross-linking	1,21E-02
GO:0051094	positive regulation of developmental process	1,26E-02
GO:0006544	glycine metabolic process	1,29E-02
GO:0051340	regulation of ligase activity	1,41E-02
GO:0006800	oxygen and reactive oxygen species metabolic process	1,48E-02
GO:0032504	multicellular organism reproduction	1,48E-02
	KEGG pathways	
hsa04115	p53 signaling pathway	8,83E-06
hsa05130	Pathogenic Escherichia coli infection	3,57E-04
hsa04610	Complement and coagulation cascades	4,21E-04
hsa00010	Glycolysis / Gluconeogenesis	1,99E-03
hsa00051	Fructose and mannose metabolism	3,11E-03
hsa03050	Proteasome	4,56E-03
hsa00330	Arginine and proline metabolism	4,66E-03
hsa04350	TGF-beta signaling pathway	1,49E-02
hsa00250	Alanine, aspartate and glutamate metabolism	1,81E-02
hsa04540	Gap junction	1,85E-02
hsa00590	Arachidonic acid metabolism	1,91E-02
hsa04514	Cell adhesion molecules (CAMs)	2,00E-02
hsa04512	ECM-receptor interaction	2,24E-02
hsa04510	Focal adhesion	2,42E-02
hsa04360	Axon guidance	2,82E-02
hsa00230	Purine metabolism	3,26E-02
hsa04614	Renin-angiotensin system	3,42E-02
hsa00910	Nitrogen metabolism	3,69E-02
hsa05217	Basal cell carcinoma	3,88E-02
hsa05219	Bladder cancer	3,97E-02
hsa05214	Glioma	4,45E-02

b. HOTΔPRC2

term ID	description	p-value
	GO terms	
GO:0006954	inflammatory response	3,60E-05
GO:0006690	icosanoid metabolic process	7,59E-06
GO:0006875	cellular metal ion homeostasis	1,14E-02
GO:0044057	regulation of system process	7,07E-05
GO:0060585	positive regulation of prostaglandin-endoperoxide synthase activity	3,16E-02
GO:0030005	cellular di-, tri-valent inorganic cation homeostasis	2,24E-02
GO:0045884	regulation of survival gene product expression	1,52E-02
GO:0006631	fatty acid metabolic process	3,43E-03
GO:0006636	unsaturated fatty acid biosynthetic process	3,72E-02
GO:0007584	response to nutrient	2,66E-02
GO:0006928	cellular component movement	1,81E-03
GO:0040008	regulation of growth	3,35E-02
GO:0042127	regulation of cell proliferation	5,30E-03
GO:0043449	cellular alkene metabolic process	1,48E-02
GO:0040012	regulation of locomotion	2,14E-03
GO:0001775	cell activation	2,27E-03
GO:0055066	di-, tri-valent inorganic cation homeostasis	2,51E-02
GO:0006937	regulation of muscle contraction	1,64E-02
GO:0001932	regulation of protein phosphorylation	4,33E-03
GO:0007610	behavior	8,19E-05
GO:0031667	response to nutrient levels	5,77E-05
GO:0000038	very long-chain fatty acid metabolic process	1,29E-02
GO:0051674	localization of cell	1,16E-03
GO:0050817	coagulation	1,60E-03
GO:0007267	cell-cell signaling	4,59E-02
GO:0050878	regulation of body fluid levels	4,76E-02
GO:0008037	cell recognition	2,43E-02
GO:0006836	neurotransmitter transport	1,55E-02
GO:0045321	leukocyte activation	4,32E-03
GO:0043408	regulation of MAPK cascade	2,62E-02
GO:0080135	regulation of cellular response to stress	5,03E-03
GO:0010941	regulation of cell death	1,04E-02
GO:0001525	angiogenesis	1,65E-02
GO:0046942	carboxylic acid transport	3,72E-02
GO:0006575	cellular modified amino acid metabolic process	4,80E-02
GO:0033559	unsaturated fatty acid metabolic process	2,16E-04
GO:0006952	defense response	1,17E-02
GO:0016265	death	1,61E-02
GO:0030593	neutrophil chemotaxis	2,69E-04
GO:0015849	organic acid transport	4,87E-02
GO:0009719	response to endogenous stimulus	3,33E-04
GO:0009611	response to wounding	4,42E-03
GO:0007626	locomotory behavior	4,87E-02
GO:0035556	intracellular signal transduction	1,13E-02
GO:0046888	negative regulation of hormone secretion	4,94E-02
GO:0030203	glycosaminoglycan metabolic process	4,69E-02
GO:0051270	regulation of cellular component movement	1,57E-02
	KEGG Pathways	
hsa04060	Cytokine-cytokine receptor interaction	1,11E-03
hsa04514	Cell adhesion molecules (CAMs)	2,13E-02
hsa00030	Pentose phosphate pathway	6,33E-02
hsa00590	Arachidonic acid metabolism	6,45E-02

Supplementary Table S7 | Common genes between the catalogue of differentially expressed transcripts in HEK-Epi cells over expressing full-length HOTAIR, and catalogues established by Gupta *et al.* and Gröger *et al.*

	number	genes
pHOT_down	**792**	
commons Gupta	**13**	ABCA12, ACE2, CYS1, GATM, IRF5, KIAA1324L, SCN3B, SIPA1L2, SMPD3, TLR2, TMEM27, TMEM47, VPS37D
commons Groger	**29**	ACTA2, ARHGAP26, BAMBI, BDKRB2, CDH3, COBLL1, CTH, CXCL16, CXCL2, FLRT2, FST, HTRA1, IFI44L, IFITM1, KRT17, LCN2, MAP1B, NRCAM, OAS1, PRRG4, PRUNE2, RHOD, SERPINA3, SLC27A2, SLC6A15, ST14, ST6GALNAC2, SYNE1, VEGFA
pHOT_up	**772**	
commons Gupta	**17**	ABCC9, ANKFN1, CISH, CRABP2, DOK1, ENPP2, FGFR4, GMPPA, IRF6, ITGB2, NPR3, PCDH18, PDE4B, PIAS3, PLCB4, SEMA4A, STON1
commons Groger	**58**	AKR1C1, AKR1C3, ANGPTL4, ANXA6, APOBEC3B, CITED2, CNTN1, COL4A1, CPM, CRABP2, CTGF, CTSL2, CXCL1, ENPP1, ENPP2, ETV1, FERMT1, FGFR3, FRMD4A, FZD7, GADD45B, GPX3, HAS2, HMGA2, ID2, IGFBP3, IGFBP5, IL18, IL1A, IL1B, IL7R, INHBA, INSIG1, LAD1, LRP8, MARCKS, MGLL, MYLK, NAV3, NFE2L3, PDGFRL, PLAUR, PLCB4, PRSS23, PRSS8, RGS4, SCNN1A, SEMA3A, SERPINB2, SERPINB7, SRGN, SRPX2, STC1, TGM2, THBS1, TM4SF1, TMEM158, TSPAN1