Functional characterization of long non-coding RNAs associated with the epithelial-to-mesenchymal transition

Julien Jarroux

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Institut Curie
UMR 3244 « Dynamique de l'information génétique »

Functional characterization of long non-coding RNAs in the epithelial-to-mesenchymal transition

Caractérisation fonctionnelle des longs ARN non-codants dans la transition épithélio-mésenchymateuse

Soutenue par
Julien JARROUX
Le 23 septembre 2019

Ecole doctorale n° ED 515
Complexité du vivant

Spécialité
Génomique

Composition du jury :

M. Alain PUISIEUX PU-PH, Institut Curie-CRCL	*Président*
Mme Maite HUARTE PI, Cima Universidad de Navarra	*Rapporteure*
Mme Eleonora LEUCCI PI, KU Leuven	*Rapporteure*
M. Jean-Christophe ANDRAU DR2, IGMM-CNRS	*Examinateur*
M. Antonin MORILLON DR1, Institut Curie-CNRS	*Directeur de thèse*
Mme Marina PINSKAYA MCU, Institut Curie-SU	*Examinateure*

institutCurie
Ensemble, prenons le cancer de vitesse.

Acknowledgments

First of all, I would like to thank the members of the jury for accepting to review my work: Maite Huarte and Eleonora Leucci for their expertise in reading and reviewing my thesis manuscript, and François Radvanyi and Jean-Christophe Andrau for examining my thesis defense. I would also like to thank Alain Puisieux for agreeing to be president of the jury.

I would also like to thank Charlotte Proudhon, Claire Torchet and Pablo Navarro-Gil for being part of my thesis committee during these four years, their expertise, advices, encouragements and expectations made the work present in this manuscript much more complete.

I want to thank Antonin Morillon for meeting me five years ago, seeing how passionate I was to study long non-coding RNAs and inviting me to join his lab and then stay for a PhD. During these five years together, your mentoring has helped me tremendously see results in a positive way, taking a step back to look at them in a new way, always keeping the greater picture of the project in mind despite daily troubles at the bench. All these things will always stay with me and made me a better scientist, always remembering to stay passionate, to always try to be excited even by the small results and that it's never too late to start a genome-wide screen! Thank you for letting me work on this fascinating topic that I hopefully made a little bit mine, even if when I didn't manage to think I could.

I also want to deeply thank Marina Pinskaya for her close mentorship over the last five years, for a relationship which always grew as experiments went right or wrong. You taught me so many things that I don't even know how to summarize them here, from daily things at the bench to managing a big project and funding it, from being a good scientist to being a good person. I want to thank you for always supporting me, for always sticking together, you taught me the value of hard work, of doing experiments again and again, of reading if I don't know something, of patience when I don't have it anymore. You expected so much of me it made me better. Your experience, knowledge and skills as a researcher, a teacher and a strong woman were priceless to me and I will forever be thankful.

I also want to thank the other members of the Morillon team, especially Maxime who helped me so much on the daily, who answered my ever-lasting questions, kept

scientific conversations alive for our projects with precious insight, and who was always here to make jokes or watch silly youtube videos. I especially thank Marc for all the tremendous work he did on bioinformatics analysis of all my data, truly the best bioinfo I know, you can analyze any data! Thank you to Dominika for daily conversations and Polish chocolate, Anna for always having a kind word in all situations, Nicolas for your help with literally everything in the lab, Ugo for being my communist reference and talking politics when nobody else cares about it. I also want to thank former members of the lab Marc D, Mayuko, Zohra, Claire and Angéline for their help in the beginning of the project. Finally, I want to thank Rocco for his help and for picking up the torch to continue this work and more.

I want to thank Arturo Londono for welcoming me in UMR3244 and being of great help discussing results regarding EMT and this very tricky system of ours. I want to thank all the other people at UMR3244 who helped me daily, from a talk with coffee to doing experiments together: Emilia, Win-Yan, Anne G, Irena, Emilie, Karina, Céline D, Mylène, Mike, Sophie, Alexandre, Fatemeh, Irina, Céline A, Aurore, Alexandra, Arnaud, Dipti, Vincent, Boris, Rachida, Romane, Anne F. I also want to thank the people from the NGS plateform for all their help with my sequencing projects, especially Sylvain Baulande, Sonia Lameiras and Virginie Raynal.

Out of everyone in the lab I especially want to thank Marie-France and Olivia who helped me every step of the anytime I needed anything, going through all the administrative loops from institutes and universities.

Je voudrais aussi remercier ma mère pour tout son soutien tout au long de mes études et avant, pour avoir toujours tout fait pour que nous puissions avancer et faire des études même quand les chances n'étaient pas de notre côté. Merci Maman ! Merci aussi à mon frère Nicolas et à ma tante Martine pour leur aide, psychologique ou pratique pendant ces 4 longues années.

Je remercie particulièrement Eleonore, la meilleure rencontre de l'UMR3244. J'étais content d'arriver dans ce laboratoire mais je n'aurai jamais pensé y rencontrer ma meilleure amie. Entre les cafés, les joies, les pleurs, les gâteaux (les nombreux gâteaux), les paillettes, la construction et la déconstruction, les trajets épuisés en métro où je m'endormais sur ton épaule, Neely, les poulets rôtis du dimanche : être greffé à la hanche avec toi n'aura été que pur bonheur. Merci pour ton soutien

indéfectible, d'avoir su me dire quand j'étais trop méchant avec moi-même parce-que la thèse c'est dur, d'être restée forte face à toute l'adversité : tu m'as inspiré plus que les mots peuvent le dire.

Je remercie aussi Julien, mon fournisseur officiel de thé pour maintenir un taux permettant de resté éveillé quand le redbull ce serait peut-être un peu trop. Merci d'avoir été là, juste à côté, tout le temps.

Merci aussi à Jean-Baptiste et Clément, les meilleurs. Jean-Baptiste pour ton optimisme aussi ensoleillé que la Vendée et le fait de pouvoir accepter mes critiques sur tes présentations comme jamais je pourrai le faire moi-même. Clément merci pour ta présence, ta musique, ton (vrai) soleil de Montpellier dans le cœur et les tâches de vin blanc, les vraies amitiés sont celles qui ne se brisent jamais.

Merci à tous les meilleur.e.s ami.e.s que j'aurai pu avoir et qui m'ont aidé à rester sain d'esprit, conscient de moi et de nous tous. Ces quatre années auraient été bien plus dures et beaucoup moins fun (et politiques) sans vous : Benjamin, Maxime, Aymeric, Roxane, David, Adrien, Hugo, Jonathan, Julie, Alice, Pierre, Sylvain, Damien, Fabrice, Nicolas. (Désolé si j'oublie des gens)

Et toutes les queens de Paris mais surtout Poulette qui était la seule à comprendre et Minima pour les bingos du dimanche soir tout le long de la rédaction.

Cette thèse est dédiée à la mémoire de mon grand-père Gérard Dauga, décédé juste avant mon arrivée au laboratoire, et à me grande-tante Henriette Hillou, décédée juste avant de terminer. J'espère qu'il et elle auraient été fièr.e.

This thesis manuscript is dedicated to the memory of my grand-father Gérard Dauga who passed away right before I started working on all of this, and to my great-aunt Henriette Hillou who passed away as I was finishing it. Hopefully they would be proud.

Abbreviations

ancRNA, asRNA, ANT: antisense non-coding RNA

ATAC-: Assay for Transposage-Accessible Chromatin

CAGE-: Cap Analysis of Gene expression

cDNA: complementary DNA

ceRNA: competing endogenous RNA

cheRNA: chromatin-enriched RNA

ChIP: Chromatin ImmunoPrecipitation

Chro-Seq: Chromatin-associated RNA sequencing

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

CRISPRa: CRISPR-based transcriptional activation

CRISPRi: CRISPR-based transcriptional inhibition

CUT: Cryptic Unstable Transcript

Cyto-Seq: Cytoplasm-associarted RNA sequencing

dCas9: dead Cas9

DE-: differentially-expressed

DNA: Desoxyribonucleic acid

EAL-/MAL-ncRNA: Epithelial/Mesenchymal identity-Associated LncRNA

EMT: Epithelial-to-Mesenchymal Transition

ENCODE: Encyclopedia of DNA Elements

eRNA: enhancer RNA

FBS: Fetal Bovine Serum

GRC-RN: GAA-repeat containing RNA

GRE: Glucocorticoid Response Element

GRO-: Global Run On

HAR: Human Accelerated Region

HEK: Human Epithelial Kidney

HGP: Human Genome Project

Kb: kilobase

lncRNA: long non-coding RNA

MET: Mesenchymal to Epithelial Transition

mRNA: messenger RNA

ncmtRNA: mitochondrial ncRNA

ncRNA: non-coding RNA

NET-: Nascent Elongating Transcript

PALR: Promoter Associated LncRna

PBS: Phosphate Buffered Saline

PCG: Protein Coding Gene

PCR: Polymerase Chain Reaction

PD: Population Division
PI: Propidium Iodine

PROMPT: Promoter upstream transcript

RNA: Ribonucleic acid

RNAi: RNA interference

rRNA: ribosomal RNA

RT: Room Temperature

RTqPCR: Retro-transcription quantitative PCR

SAL: Senescence Associated LncRNA

sgRNA: single guide RNA

snoRNA: small nucleolar RNA

SNPs: Single Nucleotide Polymorphism

snRNA: small nuclear RNA

sORF: short open reading frame

ssRNA: single-stranded RNA
dsRNA: double-stranded RNA

tRNA: transfer RNA

TSS: transcription start site

uaRNA: upstream antisense transcript

UCR: ultraconserved region

UTR: untranslated region

vlincRNA: very long intergenic ncRNA

XUT: Xrn1-Unstable Transcript

Table of content

INTRODUCTION

CHAPTER 1
Generalities on long non-coding RNAs

CHAPTER 2
LncRNAs as regulators of the epithelial-to-mesenchymal transition.

Chapter 1. Generalities on long non-coding RNAs

Note : the following chapter was adapted from a review on the history, discovery and classification of lncRNAs published as part of the book "Long Non Coding RNA Biology" (Jarroux et al., 2017).

The deep complexity of eukaryotic transcriptomes and the rapid development of high-throughput sequencing technologies led to an explosion in the number of newly identified and uncharacterized lncRNAs. Many challenges in lncRNA biology remain, including accurate annotation, functional characterization, and clinical relevance. The long journey for the biological characterization of non-coding RNAs is summed-up in figure C1-1 and this history will be described over the first part of this chapter, from the discovery of RNA to the genomic era. Then, we will discuss the general features of lncRNA genes and transcripts as well as their role in biodiversity and biological complexity. Next, we will consider the specificities associated with subcellular localization of lncRNAs in the cell. Finally, through the different lncRNA classifications which have been proposed, we will discuss their length, genomic location, biogenesis, and overall functions.

1. History and discovery of lncRNAs

1.1. A role for RNA in the cell: the central dogma of molecular biology

Before the ever-expanding catalogues of lncRNAs that we have today, a long experimental and theoretical journey was required to prove the importance of RNA molecules in cell biology. It began in 1869 with the discovery of nucleic acids and it took over a hundred years for researchers to finally identify non-coding transcripts and begin proposing regulatory roles for them (figure C1-1).

The link between DNA and RNA was established in the late 1950s as Elliot Volkin and Lawrence Astrachan thoroughly described RNA as a DNA-like molecule synthesized from DNA. This discovery was then further elaborated into a molecular concept of RNA and DNA synthesis (Ochoa, 1980), (Griffiths et al., 2000). Indeed, following the x-ray crystallographic studies of Rosalind Franklin and the establishment of the double helix structure of DNA by James Watson and Francis Crick in 1953, it was proposed in 1961 that RNA could be an intermediate molecule in the information flow from DNA to proteins (Cobb, 2015). First devised in 1958 by Francis Crick and then

by François Jacob and Jacques Monod, the Central Dogma of Molecular Biology comprised transcription of a DNA gene into RNA in the nucleus followed by protein synthesis in the cytoplasm. It was also stated that the information flow can only proceed from DNA to RNA and then from RNA to protein, but never from protein to nucleic acids (Cobb, 2015). The mediating role of RNA became a new focus of research which has been pivotal for the development of modern molecular biology.

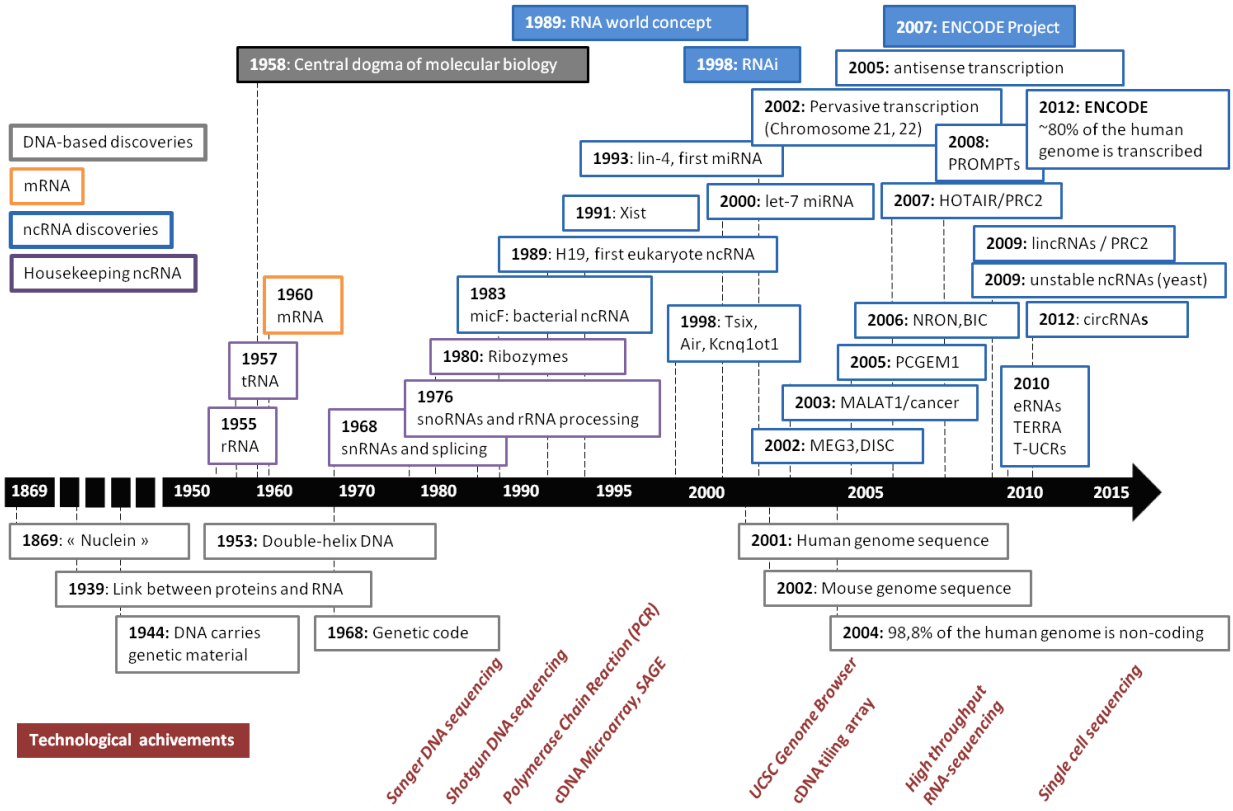


Figure C1-1. The timeline of main discoveries in nucleic acids biology and, in particular, eukaryotic ncRNAs, from the discovery of "nuclein" in 1869 to today. DNA-based discoveries are represented in grey, mRNA in orange, housekeeping RNAs in purple and non-coding RNAs in blue. The appearance of new technologies is noted below in burgundy.

In 1939, Torbjörn Caspersson and Jean Brachet showed independently that the cytoplasm is very rich in RNA. They also showed that cells producing high amount of proteins seemed to have high amounts of RNA as well (Cobb, 2015). This was a first hint for the requirement of RNA during protein synthesis and its role as a link between DNA and proteins. In 1955, Georges Palade identified the very first non-coding (nc)RNA that makes part of the very abundant cytoplasmic ribonucleoprotein (RNP) complex: the ribosome. In his "Central Dogma" Crick also theorized that there

was an "adapter" molecule for the translation of RNA to amino acids. This second class of ncRNAs was discovered in 1957 by Mahlon Hoagland and Paul Zamecnik: the transfer (t)RNA. In 1960, François Jacob and Jacques Monod first coined the term "messenger RNA" (mRNA) as part of their study of inducible enzymes in Escherichia coli. Indeed, they showed the existence of an intermediate molecule carrying the genetic information leading to protein synthesis. Shortly after, the work of Crick helped establish that the genetic code is a comma-less, non-overlapping triplet code in which three nucleotides code for one amino acid. It was later deciphered in vitro as well as in vivo and shown to be universal across all living organisms (Crick, 1968). In the late 1960s, rather different from mRNAs, a new class of short-lived nuclear RNAs was found: heterogeneous nuclear (hn)RNAs. These long RNA molecules, which were in fact precursors for mature rRNAs and mRNAs, led to the study of rRNA processing and the discovery of splicing (Lewis et al., 1975), (Berk, 2016). During that period, small nuclear (sn)RNAs which are part of the spliceosome, the RNP machinery responsible for intron-splicing from pre-mRNAs, were discovered (Weinberg and Penman, 1968); as well, small nucleolar (sno)RNAs, which are involved in the processing and maturation of ribosomal RNAs in the nucleolus were also identified (Zieve and Penman, 1976).

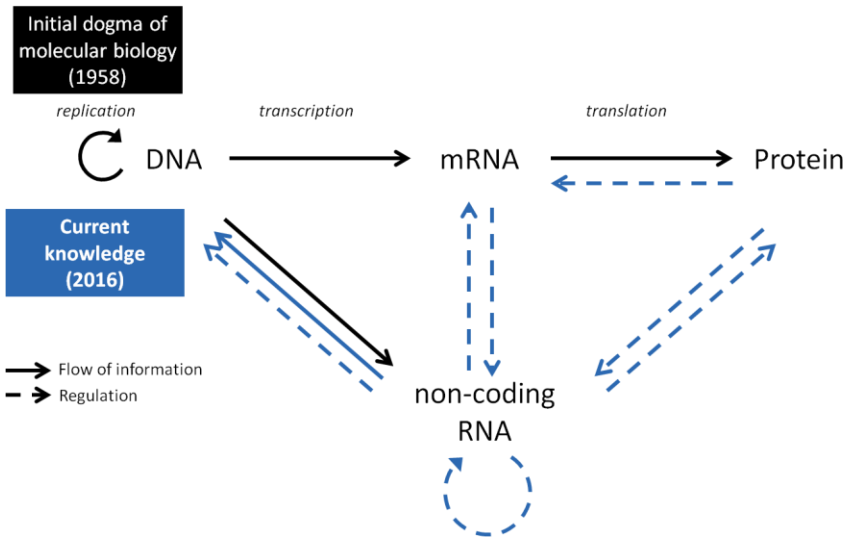


Figure C1-2. Initial and current dogma of molecular biology. Initial dogma is represented in black (1958) while our current knowledge is in blue (2016). Full arrows represent the flow of genetic information and dotted arrows represent regulatory interactions.

Although Jacob, Monod and Crick had already mentioned independently that RNA was not just a messenger, many scientists considered it as a mere unstable intermediate molecule, overlooking the active roles of other classes of ncRNAs. However, this view partially changed in 1980 when Thomas Cech and Sidney Altman discovered that RNA molecules could act as catalysts for a chemical reaction. Initially, Cech's group found an intron from an mRNA in Tetrahymena thermophila that is able to perform its own splicing through an RNA-catalyzed cleavage (Kruger et al., 1982). Subsequently, Altman's group showed that the RNA component of the ribonucleoprotein RNase P is responsible for its activity in degrading RNA (Guerrier-Takada et al., 1983). These RNA-enzymes were called ribozymes and have been shown since then to be key actors of the genetic information flow and are part of both the ribosome and the spliceosome (Cech, 2000), (Butcher, 2009).

The discovery of catalytic RNA also led scientists to develop the RNA World theory, which states that prebiotic life revolved around RNA, since it appeared before DNA and protein. Indeed, the extensive studies of its roles in cell biology revealed that RNA is necessary for DNA replication and that its ribonucleotides are precursors for DNA's deoxyribonucleotides. Moreover, as it was previously mentioned, RNA plays an important role in every step of protein synthesis, both as scripts (mRNAs) and actors (ncRNAs: rRNAs, tRNAs etc) (Figure C1-2) (Bernhardt, 2012). Remarkably, the latter ones are constitutively expressed in the cell and are necessary for vital cellular functions, constituting a class of housekeeping ncRNAs.

More recently in the early 1990s, other classes of regulatory ncRNAs have been described: they are characterized by very specific expression during certain developmental stages, in certain tissues or disease states and play multiple roles in gene expression regulation.

1.2. The first regulatory non-coding RNAs

- **MicroRNAs and RNA interference**

In the early 1990s several scientists observed independently and in different eukaryotic organisms, through experiments of transgene co-expression or viral infection, an intriguing phenomenon of RNA-mediated inhibition of protein synthesis. The regulatory effects of these RNA molecules reshaped the views of RNA as a mere messenger. The very first studies, described the phenomenon as "co-suppression" in plants, "post-transcriptional gene silencing" in nematodes or as

“quelling” in fungi, but none of them suspected RNA to be the key actor until the identification of the first micro (mi)RNA in the nematode Caenorhabditis (C.) elegans in 1993 by Victor Ambros and coworkers. Ambros discovered that the lin-4 gene produces small RNAs of 22 and 61 nt from a longer non-protein-coding precursor. The longer RNA forms a stem-loop structure, which is cut to generate the shorter RNA with antisense complementarity to the 3’ untranslated region (UTR) of the lin-14 transcript (Lee et al., 1993). The lin-4 RNA pairing to lin-14 mRNA was proposed as a molecular mechanism of “post-transcriptional gene silencing”, thus decreasing LIN-14 protein levels at first larval stages of nematode development (Wightman et al., 1993). Michael Wassenegger observed s similar phenomenon occurred in plants which he described as “homology-dependent gene silencing” or “transcriptional gene silencing”; this process is mediated by the incorporation of viroid RNA which induces the methylation of the viroid cDNA and gene silencing (Wassenegger et al., 1994). Ultimately the entire process of RNA-mediated gene silencing was elucidated in 1998 by Andrew Fire and Craig Mello in similar experiments with the unc-22 gene of C. elegans.

In 2000, another essential miRNA was identified in C. elegans. This miRNA, let-7, was shown to have homologues in several other organisms, including humans (Ameres and Zamore, 2013), (He and Hannon, 2004). The biogenesis as well as the molecular mechanisms of miRNA-mediated gene silencing has been extensively characterized. In 2001 by Thomas Tuschl showed that, in C. elegans, long double-stranded RNA is processed into shorter fragments of 21-25 nts. Since this discovery, it has been demonstrated that premature transcripts in the nucleus are processed into hairpin-structured RNA by the Drosha-containing Microprocessor complex, then exported to the cytoplasm where they are cleaved into a double-stranded RNA by Dicer. One of the strands of this double stranded RNA is loaded to the RISC complex and then targeted to an mRNA molecule by complementarity, thus inducing translational repression (He and Hannon, 2004). This simplified scheme constitutes the mechanistic basis of RNA interference (RNAi) and presently unites all gene silencing phenomena at transcriptional and post-transcriptional levels mediated by small ncRNAs including miRNAs, small interfering (si)RNAs and Piwi-interacting (pi)RNAs, all of which are processed from double-stranded RNA precursors (Montgomery, 2004), (Castel and Martienssen, 2013).

- **lncRNA discovery in the pre-genomic era: H19**

Although the focus on RNAi resulted in a breakthrough for modern biology and biotechnology, as well as providing a deeper understanding of gene regulation, development and disease, the relevance of lncRNAs remained largely unexplored. Nevertheless, some lncRNAs were investigated in the late 1980s, such as H19 the first eukaryotic lncRNAs, and Xist the milestones of dosage compensation in mammals.

In the 1980s, scientists were using differential hybridization screens of cDNA libraries to clone and study genes with tissue-specific and temporal patterns of expression. Initially, efforts were focused on genes producing known proteins; subsequently, a posteriori approach was adopted without regard to the coding potential of RNA. Through this approach, the first non-coding gene was discovered, H19, even though at that time it was first classified as an mRNA (Pachnis et al., 1984). In the late 1980s, elegant genetic and molecular studies discovered a phenomenon of genomic imprinting or parent-of-origin specific expression that constitutes part of the dosage compensation mechanisms that act through epigenetic gene silencing. Independently, two imprinted genes were identified: the paternally expressed protein-coding Igf2 and the maternally expressed H19. Both genes were localized to mouse chromosome 7 in proximity to each other forming the H19/IGF2 cluster (Bartolomei et al., 1991), (Barlow et al., 1991). What made H19 unusual was the absence of translation even though the gene contained small open reading frames. H19 showed high sequence conservation across mammals and the abundant transcript presented features of mRNAs: transcribed by RNA polymerase II, spliced, 3' polyadenylated and localized to the cytoplasm (Brannan et al., 1990). The expression of H19 in transgenic mice revealed to be lethal in prenatal stages, suggesting not only that the dosage of this lncRNA is tightly controlled, but that it has an important role in embryonic development. Since then, H19 has been thoroughly investigated and represents the prototype of a multi-tasking lncRNA. However, the function of H19 as a RNA molecule in its own right remained a mystery until the functional characterization of another lncRNA involved in dosage compensation in mammals, Xist. The pioneering studies of H19 and Xist revolutionized our view of non-coding gene function and on the biological relevance of lncRNAs in general. These examples demonstrated the complexity and versatility of regulatory circuits orchestrated by a single lncRNA. They also stimulated the

discovery and suggested potential mechanisms for other yet uncharacterized non-coding transcripts. A global effort towards lncRNA identification and characterization began in the 2000s, as a plethora of novel non-coding transcripts were uncovered from the sequencing of the complete human genome.

1.3. From non-coding genome to non-coding transcriptome

Our modern view of eukaryotic transcriptomes was preceded by comprehensive investigations of genomic DNA and the discovery that, in addition to sequences coding for proteins (PC) and regulatory elements essential for PC genes (PCG) transcription, the majority of the genome contains sequences that were considered to be useless evolutionary fossils. To differentiate these sequences from PC sequences this DNA was named non-coding and referred to as selfish or junk DNA for almost 20 years (Orgel and Crick, 1980).

By the end of 1990, a worldwide sequencing effort to completely sequence the human genome, The Human Genome Project (HGP), was established by the National Institute of Health (NIH, USA). In parallel, the American biochemist and entrepreneur Craig Venter founded his own company and sought private funding to achieve the same goal. This put pressure on the public groups involved in the HGP and the race to unravel the human genome began. The first bacterial genome was published in 1995 (Fleischmann et al., 1995). It was followed in 1999 by the sequence of the euchromatic portion of human chromosome 22 (Dunham et al., 1999), which covered approximately 65% of what is now known to be the full chromosome 22. This sequence was thought to contain 545 protein-coding genes (whether known or predicted), with PC exons spanning a mere 3% of the full sequence.

Finally, the first draft of the complete human genome was published in Nature in 2001 covering 96% of the euchromatin (Lander et al., 2001), followed the next day by Craig Venter's publication in Science of the whole-genome sequence obtained by the shotgun-cloning method (Venter, 2001). Regular updates completed most of the human genome sequence in 2003. In the meantime, the genomes of several other organisms had already been released, notably yeast (Goffeau et al., 1996), pufferfish (Crollius, 2000), worm (Waterston and Sulston, 1995), fruit fly (Adams et al., 2000) and mouse (Chinwalla et al., 2002), thus allowing comparative studies to be performed.

The first surprise from this comprehensive genomic sequencing effort was the rather low number of PCGs compared to what was initially expected. Indeed, early studies that looked at the repartition of CpG islands predicted 70,000-80,000 genes in the human genome (Antequera and Bird, 1993), a figure close to the well-admitted 100,000 genes from the mid 1980s. However, the HGP predicted around 31,000 PCGs in 2001, reduced to 22,287 PCGs in 2004 (Lander et al., 2001), (Human Genome Sequencing Consortium, 2004). In whole, only 1.2% of the human genome represents PC exons, whereas 24% and 75% were attributed to intronic and intergenic non-coding DNA.

The HGP also revealed that most of the genome is actually transcribed, whether it encodes proteins or not. Indeed, a tiling array with oligonucleotide probes spanning human chromosomes 21 and 22 revealed that 90% of detected cytosolic polyadenylated transcripts map to non-coding genomic regions and not to exons (Kapranov, 2002). Similar results were found by the FANTOM and RIKEN consortia when analyzing the transcriptome in both human (The FANTOM Consortium, 2005a) and mouse (Okazaki et al., 2002). They sequenced more than 60,000 full-length cDNAs from mouse in a standardized manner to generate accurate maps of the 5' and 3' boundaries of all transcripts, thus defining transcription start (TSS) and termination (TTS) sites. Remarkably, Cap Analysis Gene Expression (CAGE)-sequencing, a technique that sequences 5' ends of capped transcripts, revealed over 23,000 ncRNAs originating from both sense and antisense transcription representing approximately two thirds of the mouse genome (Katayama et al., 2005). For the first time, antisense transcription was proposed to contribute to the regulation of gene expression at transcriptional level in mammals.

These results were later confirmed by even larger-scale studies conducted in humans by the ENCODE (Encyclopedia of DNA Elements) consortium. This project compiled over 200 experiments in its pilot phase (ENCODE Project Consortium et al., 2007) and up to 1,640 datasets from 147 different cell lines in its later release (ENCODE Project Consortium, 2012a). Through various sequencing techniques, landscapes of DNase I hypersensitive sites, histone modifications, transcription factor binding sites and the whole transcriptome, were defined. Conclusions from these studies estimated that 93% of the human genome is actively transcribed and associated with at least one primary transcript (i.e. coding and non-coding exons

and introns); among these transcripts approximately 39% of the genome represented PCGs (from promoter to poly-A signal, including introns), 1% protein-coding exons, while the other 54% mapped outside of PCGs (Figure C1-3). However, many lncRNAs overlap with PCG annotations in both sense, coding and antisense strands. More recently, the Mouse counterpart of the ENCODE Consortium confirmed previous reports by publishing a similar analysis which showed that 46% of the mouse genome produces mRNAs while at least 87% of its genome is transcribed (Mouse ENCODE Consortium et al., 2012), (Yue et al., 2014).

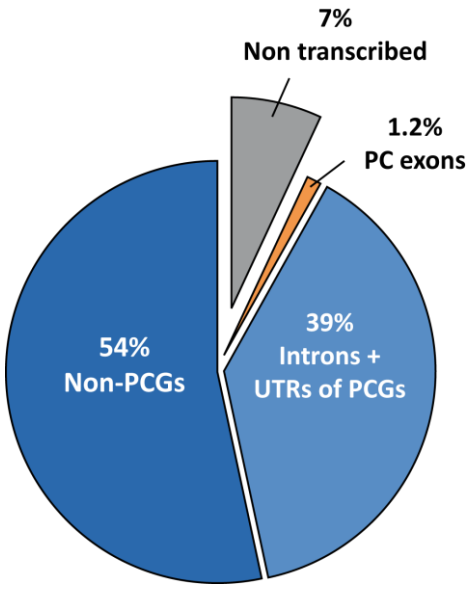


Figure C1-3. Proportion of transcribed protein-coding (PC) and non-coding sequences (introns, UTRs and others) in the human genome according to ENCODE (ENCODE Project Consortium, 2012a). Orange represents all protein-coding exons, light blue represents sequences associated to protein-coding genes which are not coding (UTRs, introns), dark blue represents all the non-coding regions in the genome.

Many studies aiming at the characterization of non-coding transcription were also performed in other eukaryotes, including Saccharomyces cerevisiae. Even in this primitive unicellular eukaryote, about 85% of genome is transcribed (David et al., 2006). The fact that most of the genome is transcribed, a phenomenon often referred to as « pervasive transcription », is widespread among eukaryotes and an expanding body of literature details its function (Dinger et al., 2009), (Berretta and Morillon, 2009). The identification and characterization of non-coding transcripts as unique ncRNAs extended the old definition of a "gene" beyond its coding function.

Furthermore, the discovery of the non-coding genome and transcriptome gave rise to heated debates in the scientific community concerning the biological significance and functional relevance of this junk DNA and RNA, still perceived as a dark matter (Mattick, 2003), (Dinger et al., 2009), (Clark et al., 2013). These debates challenged the Central Dogma, promoting ncRNAs to the epicenter of the cellular processes as a driver of biological complexity through evolution.

2. A general portrait of lncRNA genes and transcripts

LncRNAs have been identified in all species which have been studied at the genomic level, including animals, plants, fungi, prokaryotes and even viruses. Genome-wide studies continue to enlarge the catalogue of lncRNAs continuously reshaping the specific features of lncRNAs as transcription units. Here, we will discuss the biological and functional relevance of lncRNAs through their origin, conservation and diversification across species, and summarize the main features that distinguish them from PCG (Table C1-I).

Table C1-I. Comparison of lncRNA and mRNA features

Feature	lncRNA	mRNA
Transcription		
	RNA polymerase II RNA polymerase III (B2-SINE; NDM29 (Massone et al., 2012), (Espinoza et al., 2007)) RNA polymerase IV and V (plants, (Ariel et al., 2015))	RNA polymerase II
Chromatin modifications		
H3K4me3	Low (eRNA, PROMPTs) High (others)	High
H3K4me1	High (eRNA, PROMPTs) Low (others)	Low
H3K27ac	High	Low
H3K36me	Moderate / High	High
H3K79me2	Enriched (bidirectional lncRNAs)	Low
H3K27me3	Present at bivalent and repressed promoters	Present at bivalent and repressed promoters
Transcript features		
5'-Cap	Present (7-methylguanosine, m7G)	Present (m7G)

Poly(A) tail	Present or not Bimorphic	Present
Length	200 - > 100 kb (10 kb mean)	5 kb mean
Exon-intron composition	Yes Exons are longer	Yes
Splicing	Yes or less efficient No (macro lcnRNA, vlincRNAs)	Yes
RNA Stability	Variable, globally lower than mRNA Highly unstable (eRNA, XUTs, CUTs, PROMPTs)	Variable
Evolutionary conservation	High (lincRNAs) Low or not conserved (others)	High
Protein-coding potential	Non or very low (sORFs)	Yes
Structure	Versatile, multi-modular	Kozac hairpin at the 5'-end
Subcellular localization	Nucleus Cytosol Mitochondria	Cytosol
Expression specificity	Very high, including inter-individual variability of expression	Moderate or low
Transcript abundance	Very low or low High (for few)	Moderate to high

2.1. Origin and evolutionary conservation

Non-coding genes were proposed to arise through various mechanisms including DNA-based or RNA-based duplications of existing genomic sequences, the metamorphosis of PCGs by loss-of-protein-coding potential, transposable elements exaptation, or non-coding DNA exaptation (Marques and Ponting, 2014). Homologous non-coding genes arise from duplications of already existing lncRNA genes. Pseudogenes are an example of PCGs metamorphosis during which the duplicated ancestral open-reading frame had accumulated disruptions destroying its potential to be translated. Once transcribed, pseudogenes often produce lncRNAs, as in the case of PTENP1. Pseudogenization of a PCG, due to mutations deleterious to translation, can also produce lncRNA genes that do not have an apparent protein-coding "homologue". An example is Xist which is derived from an ancestral Lnx3 gene and which has acquired several frame-shifting mutations during early evolution of placental mammals (Duret, 2006). Exaptation or co-option of RNA-derived transposable elements (TE) into non-coding genes is another frequent mechanism of lncRNA origination. In humans TEs constitute a large portion of the genome (40-45%) (Lander et al., 2001). Most of them are genomic remnants that are

currently defunct, but are often embedded into non-coding transcripts. TEs are considered as major contributors to the origin and diversification of lncRNAs in vertebrates (Ganesh and Svoboda, 2016). Together with local repeats, they provide lncRNA genes with TSS, splicing, polyadenylation, RNA editing and RNA binding sites, as well as nuclear retention signals or particular secondary structures for protein binding (Kapusta et al., 2013), (Johnson and Guigo, 2014), (Hacisuleyman et al., 2016).

Finally, pervasive transcription of the genome may generate cryptic RNAs that, if maintained through evolution, can give rise to lncRNA genes with novel functions. In particular, exaptation of non-coding sequences into lncRNAs can occur through the acquisition of regulatory elements within a silent region, thereby promoting transcription. However, the de novo origin of lncRNAs remains difficult to prove and is represented by few examples, such as the testis-specific lncRNA Poldi (Heinen et al., 2009). Interestingly in humans the testis and cerebral cortex are the most enriched tissues for the expression of PCGs and non-coding genes of de novo origin. This particularity was suggested to contribute to phenotypic traits that are unique to humans, such as an improved cognitive ability (Wu et al., 2011), (Durruthy-Durruthy et al., 2015).

Genomic and transcriptomic studies across the eukaryotic kingdom also allowed for the analysis of the primary sequence conservation of protein-coding and non-coding loci. These studies revealed that the human genome is highly dynamic, and only 2.2% of its DNA sequence is subjected to conservation constraints (Rands et al., 2014). Remarkably, non-coding genes are among the least conserved with more than 80% of lncRNA families being of primate origin (Necsulea et al., 2014). This finding raised skepticism regarding the functionality and biological relevance of lncRNAs and initiated a search for other conservation constrains (Young and Ponting, 2013), (Ponting et al., 2009). If the criterion of primary sequence conservation is too restrictive in regard to lncRNA genes, other features such as structure, function, and expression from syntenic loci, constitute multidimensional factors that are more applicable for evolutionary studies of lncRNAs (Diederichs, 2014). The study of the non-coding transcriptome of 17 different species (16 vertebrates and the sea urchin) also showed that although the body of non-coding genes tends not to be conserved, short patches of conserved sequences could be found at their 5' ends. This confirmed

a higher conservation of TSS and synteny, as well as expression patterns in different tissues, especially in those involved in development (Hezroni et al., 2015). Indeed, the most conserved are developmentally regulated lncRNA of the lincRNAs subfamily. These lncRNAs have a remarkably strong conservation of spatio-temporal and syntenic loci expression, suggesting that it is selectively maintained and crucial for developmental processes (Necsulea et al., 2014), (Washietl et al., 2014), (Ulitsky et al., 2011).

2.2. Role of lncRNAs in biological diversity

The diversity of the non-coding transcriptome is considered as an argument to explain the remarkable phenotypic differences observed among species given a relatively similar numbers of protein-coding genes among fruit fly (13,985; BDGP release 4), nematode worm (21,009; Wormbase release 150), and human (23,341; NCBI release 36) (Willingham and Gingeras, 2006). In 2001, John Mattick and Michael Gagen proposed, for the very first time, that non-coding transcripts named "efference" RNA, together with introns, constitute an endogenous network enabling dynamic gene-gene communications and the multitasking of eukaryotic genomes. In contrast to core proteomic circuits, this higher-order regulatory system is based on RNA and operates through RNA-DNA, RNA-RNA, and RNA-protein interactions to promote the evolution of developmentally sophisticated multicellular organisms and the rapid expansion of phenotypic complexity. A direct correlation between the portion of non-coding sequences in the genome and organism complexity was hypothesized (Mattick and Gagen, 2001), (Mattick, 2001). Interestingly comparative genomics allowed the identification of a few regions in the human genome that have high divergence when compared to other species (Pollard et al., 2006a), (Bird et al., 2007). These Human Accelerated Regions (HAR) contain many lncRNA genes and have been suggested to be involved in the acquisition of human-specific traits during evolution. In 2006, a first lncRNA from these regions was shown to be expressed during cortical brain development (Pollard et al., 2006b). Since then, many mutations involved in diseases were identified in these non-coding regions and shown to be associated with regulatory elements in the brain (Bae et al., 2014). A more recent study showed that mutations of HAR enhancer elements could be involved in the development of autism, thus supporting the hypothesis that some HAR could be involved in human-specific behavioral traits, and cognitive or social

disorders when mutated (Doan et al., 2016). However, the functionality of non-coding transcripts was and still remains hotly debated. Nevertheless, the concept of developmental and evolutionary significance has stimulated an exhaustive molecular characterization of lncRNA genes and transcripts.

2.3. Coding potential of lncRNA transcripts

As dictated by the acronym, lncRNA genes do not encode for proteins. Cytosol localized lncRNAs were found associated with mono- or poly-ribosomal complexes (van Heesch et al., 2014). However, this association is not necessarily linked to translation but rather proposed to determine lncRNA decay (Carlevaro-Fita et al., 2015), (Wery et al., 2016). Some lncRNAs include short open reading frames (sORFs) and undergo translation, though only a minority of such translation events results in stable and functional peptides (Housman and Ulitsky, 2016), (Andrews and Rothnagel, 2014). This is the case of DWORF, a muscle-specific lncRNA which encodes a functional peptide of 34 amino acids (Banfai et al., 2012), (Ruiz-Orera et al., 2014), (Ji et al., 2015), (Nelson et al., 2016). More recently, some examples of lncRNA-encoded peptides have been shown to be functionally relevant. This is the case of the SPAR polypeptide encoded by LINC00961 in humans, which was identified in both human and murine muscle cells and shown to repress the activation of translation regulator complex mTORC1 (Matsumoto et al., 2017). Upon muscle injury *in vivo* in mice, Matsumoto and colleagues showed the SPAR-encoding lncRNA is downregulated, thus allowing the activation of mTORC1 and proper activation of muscle regeneration. Another example is a peptide encoded by HOXB-AS3 which acts as a tumor suppressor in colon cancer by reprograming cell metabolism and thus inhibiting tumor growth (Huang et al., 2017). Interestingly, other variants of the HOXB-AS3 lncRNA which do not encode this peptide were shown to be upregulated in acute myeloid leukemia and ovarian cancer where they supposedly promote tumorigenesis (Huang et al., 2019; Zhuang et al., 2019). Proteomic studies will undoubtedly introduce a new “coding” aspect to lncRNAs, expanding our conception of “coding” and leading to a possible concept of bifunctionality.

2.4. LncRNA transcription and the associated chromatin signature

The majority of eukaryotic lncRNAs are produced by RNA polymerase II, with some exceptions such as the murine heat-shock induced B2-SINE RNAs (Espinoza et al., 2007), or the human neuroblastoma associated NDM29 (Massone et al., 2012), which are synthesized by RNA polymerase III. However, the last two examples are not strictly considered as lncRNAs because the transcript length is below the arbitrary threshold of 200 nt. In plants, two specialized RNA polymerases, Pol IV and Pol V, transcribe some lncRNA genes (Ariel et al., 2015). Many lncRNAs are capped at the 5' end, except those processed from longer precursors (intronic lncRNAs or circRNAs). However, some ambiguities exist concerning the presence of a cap, especially for highly unstable and low abundant transcripts, since they cannot all be captured by the CAGE-seq technique. LncRNAs may or not be 3' end polyadenylated; in addition they may also be present as both forms, such as bimorphic transcripts like NEAT1 and MALAT1 (Yang et al., 2011), (Djebali et al., 2012). LncRNAs with a polyadenylation signal have higher stability than those that are poorly polyadenylated or not polyadenylated, with the exception of lncRNAs bearing specific 3' end structures as in case of MALAT1 (Wilusz et al., 2012).

LncRNA genes can have a multi-exonic composition with similar splicing signals as PCG, and therefore could undergo splicing into several different isoforms with distinct functional outcomes and clinical relevance (Spurlock et al., 2015), (Hoffmann et al., 2015), (Meseure et al., 2016). However, they usually comprise fewer and slightly longer exons than PCGs (Derrien et al., 2012a), (Bogu et al., 2016).

As RNA polymerase II transcribes most of the lncRNA genes, their genomic regions present a chromatin organization that resembles that of PCGs, with a few differences. This could be due to the globally low expression of lncRNAs, which is a consequence of either low rate of transcription, lower stability or both. Globally, lncRNAs TSS reside within DNase I hypersensitive sites suggesting nucleosomes are depleted from this region. LncRNA promoters have lower levels of histone H3 K4 trimethylation (H3K4me3), which is in accordance with their low transcription rate. lncRNas associated to regulatory elements such as enhancers (eRNAs) and promoters (PROMPTs) present high levels of histone H3 K4 monomethylation (H3K4me1) and K27 acetylation (H3K27ac) at promoters, which is considered as a specific signature of enhancer and promoter associated unstable transcripts

(Marques et al., 2013). Over the body of most lncRNAs with the exception of eRNAs and PROMPTs, histone H3 K36 trimethylation (H3K36me3) can be found and is a mark of the elongating phase of transcription. In mouse, bidirectional transcription which is often associated with developmental genes and genes involved in transcription regulation, was found to harbor high H3K79 dimethylation (H3K79me2) and elevated RNA polymerase II levels. This signature is characteristic of intensified rates of early transcriptional elongation within a region transcribed in both directions (Lepoivre et al., 2013).

2.5. Expression pattern of lncRNAs: stability, specificity, and abundance

- **Stability**

Several genome-wide studies addressed lncRNA stability and, depending on the employed experimental approach, revealed some discrepancy for different species of lncRNAs. In mouse, the measurements of the lncRNA half-life showed they are less stable than mRNAs (Clark et al., 2012). Comparison of the stability of different lncRNA species revealed that intronic or promoter-associated lncRNAs are less stable than either intergenic, antisense, or 3' UTR-associated lncRNAs. Single exon transcripts, a class of nuclear-localised lncRNAs, are overrepresented among unstable transcripts. Circular RNAs are an example of highly stable lncRNAs compared to their linear counterparts (Enuka et al., 2016).

- **Specificity**

Multiple transcriptome profiling globally highlighted a highly specific spatio-temporal, lineage, tissue and cell-type expression patterns for lncRNAs compared to PCGs; only a minority are ubiquitously present across all tissues or cell-types, such as TUG1 or MALAT1 (Djebali et al., 2012), (Ward et al., 2015), (Li et al., 2015a). As previously mentioned, brain and testis represent a very rich source of uniquely expressed lncRNAs supporting the hypothesis that such transcripts are important for the acquisition of specific phenotypic traits (Ward et al., 2015), (Washietl et al., 2014). The ubiquitously expressed lncRNAs are often highly abundant, whereas specific lncRNAs present in one tissue or cell-type tend to be expressed at low levels (Jiang et al., 2016). Moreover, inter-individual expression analysis in normal human primary granulocytes revealed increased variability in lncRNA abundance compared to mRNAs (Kornienko et al., 2016). Some disease-associated single-nucleotide

polymorphisms (SNPs) within lncRNA genes and their promoters were linked to altered lncRNA expression, thus supporting their functional relevance in pathologies (Kumar et al., 2013). The high specificity of lncRNAs expression argues in favor of important regulatory roles that these molecules can act in different biological contexts, including normal and pathological development.

2.6. Subcellular localization of lncRNAs

Globally, unlike mRNAs, many lncRNAs have nuclear residence with focal or dispersed localization pattern (NEAT1) (Cabili et al., 2015). However, others were also found both in the nucleus and in the cytosol (TUG1, HOTAIR), or in the cytosol exclusively (DANCR) (Djebali et al., 2012). Multiple determinants, such as a specific RNA motif (BORG) or RNA-protein assemblies may dictate the subcellular localization of lncRNAs and define their function (Chen, 2016; Shukla et al., 2018; Zhang et al., 2014). Remarkably, environmental changes or infection can induce lncRNA delocalization (or active trafficking) from one cellular compartment to another, as in the case of stress-induced lncRNAs (Giannakakis et al., 2015). HuR and GRSF1 modulate nuclear export and mitochondrial localization of the nuclear-encoded RMRP lncRNA (Noh et al., 2016).

Knowing the subcellular localization of a particular lncRNA provides important insights into its biogenesis and function. LncRNAs could be exclusively cytosolic (DANCR and OIP5-AS1) or nuclear (NEAT1) or have a dual localization (HOTAIR) (Ayupe et al., 2015). Several subgroups of lncRNAs with a precise subcellular localization have been defined, such as **chromatin enriched (che)RNAs** (Werner and Ruthenburg, 2015a), and **chromatin associated lncRNAs, CARs** (Mondal et al., 2010). cheRNAs were later confirmed to act as activators of transcription for nearby genes and Werner suggested chromatin-enriched RNAs are the most effective chromatin-signature in a very cell-type specific manner (Werner et al., 2017). Many nuclear and chromatin functions have been proposed for such lncRNAs, including the assembly of subnuclear domains or RNP complexes, the guiding of chromatin modifications, and the activation or repression of protein activity (Singh and Prasanth, 2013). **GAA repeat-containing RNAs, GRC-RNAs**, represent a subclass of nuclear lncRNAs that show focal localization in the mammalian interphase nucleus, where they are a part of the nuclear matrix. They have been suggested to play a role in the organization of

the nucleus by assembling various nuclear matrix-associated proteins (Zheng et al., 2010).

The mitochondrial genome is also transcribed into **mitochondrial ncRNAs, ncmtRNAs** (Rackham et al., 2011), (Burzio et al., 2009), (Anandakumar et al., 2015). Their biogenesis is dependent on nuclear-encoded mitochondrial processing proteins. After synthesis, some ncmtRNAs are exported from mitochondria to the nucleus (Landerer et al., 2011). Importantly, expression of ncmtRNAs is altered in cancers promoting them as potential targets for cancer therapy (Vidaurre et al., 2014), (Lobos-González et al., 2016).

3. Classification of lncRNAs

As mentioned earlier, advances in deep sequencing technologies gave rise to a plethora of novel transcripts requiring a universal standardized system for lncRNA classification and functional annotation. The state of lncRNA annotations is still ongoing and different classifications have been proposed, based on their length, location in respect to known genomic annotations or regulatory elements, and on biogenesis pathways or function.

3.1. Classification according to length

By convention, a length of 200 nt constitutes a bottom line for discrimination of long or large ncRNAs from small or short ncRNAs. However, lncRNAs vary significantly in size, and those that exceed the length of 10 kb belong to the groups of very long intergenic (vlinc)RNAs and macro lncRNAs. These transcripts possess some particular features that distinguish them from other lncRNAs: they are poorly or not spliced, weakly polyadenylated at 3' end, and are produced by particular genomic loci. The majority of vlincRNAs are localized in close proximity or within PCG promoters on the same or opposite strand and function in cis as positive regulators of the transcription of nearby genes. Interestingly, some vlincRNA promoters harbor LTR sequences that are highly regulated by three major pluripotency-associated transcription factors, suggesting a possible role in early embryonic development (St Laurent et al., 2013). Others are specifically induced by senescence and are required for the maintenance of senescent features that in turn control the transcriptional response to environmental changes (Lazorthes et al., 2015). Macro lncRNAs are often antisense to PCGs and are produced from imprinted clusters in a parent-of-origin

specific manner. Macro lncRNAs silence nearby imprinted genes either through their lncRNA product triggering epigenetic chromatin modifications or by a transcriptional interference mechanism (Guenzl and Barlow, 2012).

3.2. Classification according to genomic location in respect to PCGs

- **Intergenic lncRNAs**

This attribute is commonly used by the GENCODE/Ensembl portal in transcript biotype annotations, but it is also employed on an individual scale by consortia and laboratories for newly assembled lncRNA transcripts. Initially transcripts are classified as either intergenic or intragenic (Figure C1-4). Long or large intergenic non-coding (linc)RNAs do not intersect with any protein-coding and ncRNA gene annotations. This category also includes the adopted GENCODE and homonymous biotype of long or large intervening ncRNAs that were originally defined by specific histone H3 K4-K36 chromatin signatures within evolutionary conserved genomic loci (Khalil et al., 2009), (Guttman et al., 2009a). LincRNAs are usually shorter than PCGs, are transcribed by RNA polymerase II, contain 5'-caps, are 3'-polyadenylated, and are spliced. Although several highly conserved lincRNAs exists, the majority possess modest sequence conservation comprising short, 5' biased patches of conserved sequence nested in exons (Hezroni et al., 2015). Highly conserved lincRNAs are believed to contribute to biological processes that are common to many lineages, such as embryonic development (Necsulea et al., 2014), while others are proposed to assure phenotypic and functional variations at individual and interspecies levels. Many, if not most, lincRNA are localized in the nucleus where they exercise their regulatory functions. One such example is lincRNA-p21 which is induced by p53 upon DNA damage (Huarte et al., 2010). LincRNA-p21 physically associates with and recruits the nuclear factor hnRNP-K to specific promoters mediating p53-dependent transcriptional responses.

Intragenic lncRNAs overlap with PCG annotations and can be further classified into antisense, bidirectional, intronic and overlapping sense lncRNAs.

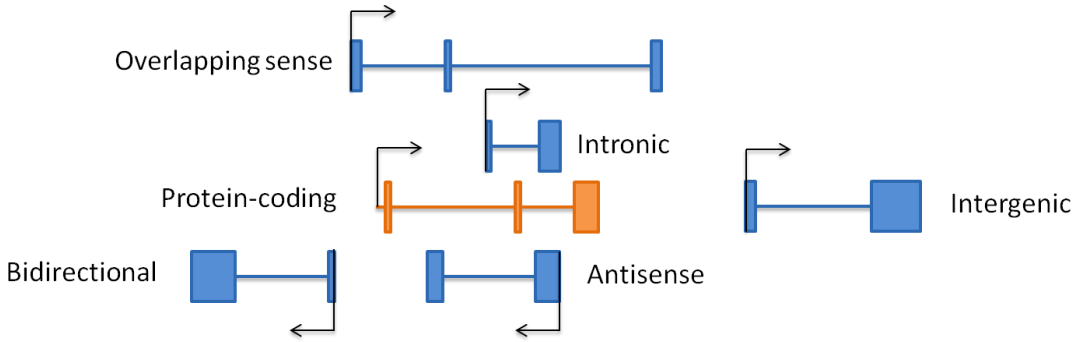


Fig. C1-4 Annotation of non-coding transcripts according to their genomic position relative to a protein-coding gene (blue box – protein-coding exon, pink box – non-coding exon).

- **Antisense lncRNAs**

Antisense lncRNAs, asRNAs or ancRNAs, were first discovered in single gene studies, but the development of stranded tiling and RNA-seq technologies has identified them as a common genome-wide feature of eukaryotic transcriptomes (Goodman et al., 2013), (Kapranov, 2005), (Wood et al., 2013). This group encompasses so-called natural antisense transcripts, NATs, which are in turn subdivided into cis-NATs, which affect the expression of the corresponding sense transcripts, and into trans-NATs, which regulate expression of non-paired genes from other genomic locations (Magistri et al., 2012), (Su et al., 2010), (Yuan et al., 2015a). A recent study has pointed to a higher specificity of expression and an increased stability of asRNAs compared to lincRNAs and sense intragenic lncRNAs (Ayupe et al., 2015). Due to sequence complementarity to sense-paired mRNAs or pre-mRNAs, asRNAs can act through RNA-RNA pairing, thereby ensuring specific targeting of the asRNA regulatory activity. This is the case of BACE1-AS that is highly expressed in Alzheimer's disease and stabilizes the BACE1 mRNA, which results in an increased expression of the BACE1 encoded beta-secretase and the accumulation of amyloid-beta peptides in the brain (Faghihi et al., 2008). Antisense transcription across intron regions has been shown to regulate the local chromatin organization and environment, thus affecting co-transcriptional splicing of sense-paired pre-mRNAs (Gonzalez et al., 2015). Some NATs contain the inverted short interspersed nuclear element B2 (SINEB2), such as AS-Uchl1 (Carrieri et al., 2012). These NATs, called **SINEUPs**, are able to stimulate sense mRNA translation through lncRNA-mRNA pairing thanks to a complementary 5' overlapping sequence to the paired-sense protein coding gene. Recently, SINEUPs were proposed as a synthetic reagent for biotechnological

applications and in therapy of haploinsufficiencies (Zucchelli et al., 2015), (Indrieri et al., 2016). In spite of the poor evolutionary conservation of sense-antisense transcription, some subgroups of lncRNAs, such as senescence-associated vlincRNAs and macro lncRNAs in mammals, or XUTs in yeast, are mostly constituted of antisense transcripts, which suggests potential antisense mediated regulatory pathways in control of cellular homeostasis, stress response and disease (Wood et al., 2013).

- **Bidirectional lncRNAs**

The discovery of bidirectional transcription as an intrinsic feature of the eukaryotic transcriptional machinery has also given rise to the identification of bidirectional lncRNAs (Xu et al., 2009), (Kapranov, 2005), (Scruggs et al., 2015), (Wei et al., 2011), (Seila et al., 2008). Originating from the opposite strand of a PCG, these transcripts do not overlap, or only partially overlap with the 5' region of paired PCGs, as is the case of promoter associated (pa)ncRNAs, long upstream antisense transcripts (LUATs) and upstream antisense transcripts (uaRNA) (Hamazaki et al., 2015), (Hung et al., 2011), (Lepoivre et al., 2013), (Uesaka et al., 2014), (Flynn et al., 2011). Presently, the number of bidirectional lncRNAs is largely underestimated not only because of the inaccurate annotation of transcriptional start sites (TSS) and promoters in the genome, but also because of the highly unstable nature of these ncRNAs and the corresponding difficulty to detect them. Genomic studies have revealed that bidirectional promoters display distinct sequences and epigenetic features; moreover, they can be found near genes involved in specific biological processes such as developmental transcription factors or cell-cycle regulation (Hung et al., 2011), (Hu et al., 2014a), (Lepoivre et al., 2013), (Uesaka et al., 2014), (Sigova et al., 2013). An imbalance in bidirectional transcription constitutes an endogenous fine–tuning mechanism that is particularly operative when facultative gene activation or repression is required (Morris et al., 2008), (Kambara et al., 2015).

- **Intronic lncRNAs**

Intronic lncRNAs are restricted to PCG introns and could be either true unique transcripts or byproducts of pre-mRNA processing. Examples of pre-mRNA derived intronic transcripts are circular intronic (ci)RNAs produced from lariat introns which have escaped from debranching (Zhang et al., 2013) and sno-lncRNAs

produced from introns with two imbedded snoRNA genes (Yin et al., 2012). Such lncRNAs are proposed to positively regulate the transcription of the host PCG or its splicing by accumulating near the transcription locus. Another example of intronic lncRNAs of lariat origin, named switch RNAs, are produced by transcription through the immunoglobulin switch regions. They are folded into G-quadruplex structures to bind and recruit the activation-induced cytidine deaminase AID to DNA in a sequence-specific manner thereby ensuring proper class switch recombination in the germline (Zheng et al., 2015). Standalone intronic transcripts, expressed independently of the PCG hosts, are believed to be the most prevalent class of intronic lncRNAs, including so-called totally intronic ncRNAs (TIN) (Nakaya et al., 2007), (Louro et al., 2008). Expression of a certain TIN is activated during inflammation but the exact function of these lncRNAs is still poorly understood (St Laurent et al., 2012).

- **Overlapping sense lncRNAs**

Overlapping sense transcripts encompass exons or the whole PCGs within its introns without any exon overlap and are transcribed in the same sense direction. One such example is SOX2-OT that harbors in its intron one of the major pluripotency regulators, the SOX2 gene. SOX2-OT is dynamically expressed and is alternatively spliced not only during differentiation but also in cancer cells where it was proposed to regulate SOX2 (Shahryari et al., 2015).

Intronic and overlapping sense lncRNAs could form circular lncRNAs (circRNAs) due to head-to-tail non-canonical splicing (Memczak et al., 2013), (Hansen et al., 2013). Some sequence features such as the presence of repetitive elements within introns could be decisive for activation of non-canonical splicing and the generation of a circular RNA molecule (Kramer et al., 2015). For example, Alu elements within introns are proposed to participate in RNA circularization via RNA-RNA pairing (Hadjiargyrou and Delihas, 2013). Remarkably, such events seem to be tissue or cell-type specific, or restricted to a certain developmental stage, as well as a characteristic of certain pathological contexts (Rybak-Wolf et al., 2015), (Peng et al., 2015). More generally, circRNAs function in the cytosol as miRNA sponges, as the case of CDR1as/ciRS-7 which is an RNA sponge of miR-7 (Memczak et al., 2013), (Hansen et al., 2013). Some circRNAs termed exon-intron circRNAs (EIciRNAs), still contain unspliced introns which ensures they will be retained in the nucleus, where

they are able to interact with U1 snRNP as well as promote transcription of their parental genes (Li et al., 2015d). The most remarkable property of circRNAs is their high stability which makes them eligible as potent diagnostic markers and therapeutic agents (Li et al., 2015b).

3.3. Classification according to genomic location within specific DNA regulatory elements

- **Pseudogenes**

In addition to PCGs, mammalian genomes contain tens of thousands of pseudogenes, which are genomic remnants of ancient protein-coding genes that have lost their coding potential through evolution. Importantly, many of them are transcribed in both sense and antisense directions into lncRNAs. Given high sequence similarity with parental genes, **pseudogene-derived lncRNAs** can regulate PCG expression via RNA-RNA pairing by acting as miRNA sponges, by producing endogenous siRNAs or by interacting with mRNAs (Milligan and Lipovich, 2015), (Zheng et al., 2007), (Grandér and Johnsson, 2015). PTENP1, a lncRNA pseudogene-derived from the tumor-suppressor gene PTEN, was among the first reported non-coding miRNA sponges with a function in cancer (Poliseno et al., 2010).

- **Ultra-conserved regions**

Ultra-conserved regions (UCRs) are genome segments (≥ 200 bp) that exhibit 100% DNA sequence conservation between human, mouse and rat. The human genome contains 481 UCRs within intragenic (39%), intronic (43%) and exonic (15%) sequences (Bejerano, 2004). These regions are extensively transcribed into T-UCR lncRNAs (Mestdagh et al., 2010), (Watters et al., 2013). Remarkably, expression of T-UCRs is induced by cancer-related stresses such as retinoid treatment or hypoxia. They are aberrantly expressed in different cancers and some are associated with poor-prognosis (Ferdin et al., 2013), (Watters et al., 2013), (Fassan et al., 2014). Given high specificity of expression, T-UCRs were proposed as molecular markers for cancer diagnosis and prognosis (Scaruffi et al., 2009). The function of T-UCRs is still poorly understood. Evf2 (or Dlx6as) is an example of a T-UCR with "decoy" function. It interacts with the transcription activator DLX1 increasing its association with key DNA enhancers, but also with the SWI/SNF-like chromatin remodeler brahma-related gene 1 (BRG1) inhibiting its ATPase activity. As a result, Evf2 provokes

chromatin remodeling and Dlx5/6 enhancers decommissioning with a final repression of transcription (Feng, 2006), (Cajigas et al., 2015).

- **Telomeres**

Telomeres, which are protective nucleo-protein structures at the ends of chromosomes, are transcribed into non-coding telomeric repeat-containing RNAs, TERRA, in all eukaryotes. This family of transcripts is generated from both Watson and Crick strands in a cell-cycle dependent manner (Feuerhahn et al., 2010)(Porro et al., 2010). Formation of RNA-DNA hybrids by TERRA at chromosome ends promotes recombination and, hence, delays senescence. However, in cells lacking telomerase and homology directed repair TERRA expression induces telomere shortening and accelerates senescence (Balk et al., 2013), (Balk et al., 2014). Subtelomeric regions are also actively transcribed (Greenwood and Cooper, 2012), (Trofimova et al., 2015), (Broadbent et al., 2015). In budding yeast this heterogeneous population of lncRNAs, named subTERRA, is transiently accumulating in late G2/M and G1 phases of the cell cycle in wild-type cells or in asynchronous cells deleted for the Xrn1 exoribonuclease (Kwapisz et al., 2015). The exact function of subTERRA is not yet clear though it has been proposed to have a regulatory role in telomere homeostasis.

- **Centromeres**

Recent findings in different eukaryotes including human revealed that centromeric repeats are actively transcribed into lncRNAs during the progression from late mitosis into early G1 (Wong et al., 2007), (Quénet and Dalal, 2014), (Blower, 2016), (Chan et al., 2012), (Rošić et al., 2014). These lncRNAs physically interact with different centromere-specific nucleoprotein components, such as CENP-A/-C and HJURP, and are required for correct kinetochore assembly and the maintenance of centromere integrity.

- **Ribosomal DNA loci**

Ribosomal (r)DNA loci were shown to be transcribed by RNA polymerase II, antisense to the rRNA genes, into a heterogeneous population of lncRNAs, called **PAPAS** (promoter and pre-rRNA antisense). Their expression is induced in quiescent cells and triggers the recruitment of histone H4K20 methyltransferase Suv4-20h2 to ribosomal RNA genes for histone modification and transcriptional silencing

(Bierhoff et al., 2014). PAPAS also allow heterochromatin formation and gene silencing in growth-arrested cells.

- **Promoter and enhancers**

Promoters and enhancers constitute fundamental cis-regulatory elements for the control of PCG expression, serving as platforms for the recruitment of transcription factors, transcription machinery and the establishment of particular chromatin organization. Remarkably, many, if not all, functional enhancers and promoters are pervasively transcribed, respectively into eRNAs and PALRs, in both sense and antisense directions. Transcribed enhancer and promoter regions possess particular histone modification signatures that distinguish them from other transcription units. Such signatures include increased histone H3 K27ac and K4me1 as compared with other lncRNA and PCGs.

The termination of enhancer-derived lncRNAs, eRNAs, depends on the Integrator complex which ensures 3' end transcript cleavage. The result is that eRNAs are poorly polyadenylated or not polyadenylated and are highly unstable. Their expression is specific to cell-type, tissue, or stages of development and can be activated by external or internal stimuli. Enhancer transcription was proposed to mark functional, active enhancer elements. However, eRNAs function as unique transcripts is still controversial and the function of only few eRNAs, such as FOXC1e or NRIP1e (Li et al., 2013b) has been demonstrated. Specifically, it is proposed that these eRNAs control promoter chromatin environment, enhancer-promoter looping, RNA polymerase II loading and pausing, and "transcription factor trapping"; all these events contribute to a robust transcription activation of nearby and distant genes (Li et al., 2016a).

Promoter-associated lncRNAs or PALRs are transcribed in sense and antisense directions at promoter regions and can partially overlap the 5'-end of PCGs (Kapranov et al., 2007). This class of transcripts includes highly unstable PROMPTs (promoter upstream transcripts) and upstream antisense RNAs (uaRNAs) that are more easily detectable in a context where the nuclear exosome has been depleted (Preker et al., 2011), (Preker et al., 2008), (Flynn et al., 2011). Polyadenylation-dependent degradation of PROMPTs was proposed to ensure directional RNA production from otherwise bidirectional promoters (Ntini et al., 2013). The presence

of a splicing competent intron within uaRNAs was shown to facilitate gene looping placing termination factors at the vicinity of a bidirectional promoter for termination and thereby ensuring RNA polymerase II directionality towards a PCG (Agarwal and Ansari, 2016). Some PARLs were shown to negatively regulate transcription of the nearby genes. One such example is a PALR from the CCND1 gene promoter which represses transcription by recruiting TLS and locally inhibiting CBP/p300 histone acetyltransferase activity on the downstream target gene, cyclin D1 (Wang et al., 2008), (Song et al., 2012).

- **Untranslated regions**

The 3'-untranslated regions (UTRs) of eukaryotic genes can be transcribed into independent transcription units or **UTR associated (ua)RNAs** (Mercer et al., 2011). They are generated either by an independent transcriptional event from the upstream PCG, or by post-transcriptional processing of a pre-mRNA. Expression of uaRNAs is regulated in a developmental stage- and tissue-specific fashion and is evolutionarily conserved. Recently, the GALNT5-uaRNA has been shown to be independently upregulated in gastric cancer patients (Guo et al., 2018). In gastric cancer cells, it was shown to promote tumor progression by inhibiting HSP90-mediated ubiquitination.

3.4. Classification according to lncRNA mechanism of action

To highlight a regulatory role, lncRNAs are often classified based on their function. Several archetypal activities of lncRNAs are used for classification: scaffolds, guides, decoys or ribo-repressors, ribo-activators and sponges, precursors of small ncRNAs. Here we present examples of functional lncRNA classifications that regroup several lncRNAs into subclasses with a common operating mode.

- **Scaffolds**

LncRNA scaffolds function in the assembly of RNP complexes. The structural plasticity of lncRNAs allows them to adopt complex and dynamic three-dimensional structures with high affinity to proteins (Guo et al., 2015). LncRNA scaffolds are often actors of epigenetic and transcriptional control of gene expression regulation. In this case a lncRNA can act in trans or in cis in respect to its transcription site (Quinn and Chang, 2015). They are known to associate with a multitude of histone- or DNA-

modifying and nucleosome remodeling complexes (Han and Chang, 2015), (Davidovich and Cech, 2015). LncRNA-mediated assembly of these complexes reshapes the epigenetic landscape and the organization of chromatin domains, thus allowing the modulation of all DNA-based processes including transcription, recombination, DNA repair, as well as RNA processing (Yoon et al., 2013a)(Zheng et al., 2015), (Lee et al., 2016), (Gonzalez et al., 2015).

Guide lncRNA can recruit RNP complexes to specific chromatin loci. Remarkably, a guide function of one and the same lncRNA depends on the biological context (cell-/tissue-type, developmental stage, pathology) and often cannot be explained by a simple RNA/DNA sequence complementarity. For some lncRNA guides the formation of a triple helix structure between DNA and the lncRNA was experimentally proven, as in the case of Khps1 which anchors the CBP/p300 complex to the proto-oncogene SPHK1 (Postepska-Igielska et al., 2015). Another example is MEG3 which guides the EZH2 subunit of PRC2 to TGFβ-regulated genes (Mondal et al., 2015).

- **Ribo-repressors or lncRNA decoys**

LncRNA decoys tend to repress protein activities through the induction of allosteric modifications, the inhibition of catalytic activity, or by blocking binding sites. One classical example of a ribo-repressor lncRNA is GAS5 (growth arrest specific 5), which acts as a decoy for a glucocorticoid receptor (GR) by mimicking its genomic DNA glucocorticoid response element (GRE). The interaction of GAS5 with GR prevents it from binding to the GRE and ultimately represses GR-regulated genes, thus influencing many cellular functions including metabolism, cell survival, and the response to apoptotic stimuli (Kino et al., 2010).

- **Ribo-activators and lncRNA sponges**

LncRNAs can also act as ribo-activators essential for or enhancing protein activities. One such example is the lnc-DC lncRNA which promotes the phosphorylation and activation of the STAT3 transcription factor (Wang et al., 2014). Another subclass is the lncRNA transcriptional co-activators, also called activating ncRNAs (ncRNA-a), which possess enhancer-like properties (Ørom et al., 2010). They were shown to interact with and regulate the kinase activity of Mediator, hence facilitating chromatin looping and transcription (Lai et al., 2013). In addition to Mediator-interacting RNAs other lncRNAs are able to upregulate transcription and could also

be considered as ncRNA-a, among them the steroid receptor RNA activator SRA which interacts with and enhances the function of the insulator protein CTCF (Yao et al., 2010), and NeST, which binds to and stimulates the activity of a subunit of the histone H3 Lysine 4 methyltransferase complex (Gomez et al., 2013).
Competing endogenous RNAs (ceRNAs), also known as lncRNA sponges, are represented by lncRNAs and circRNAs that share partial sequence similarity to PCG transcripts; they function by competing for miRNA binding and post-transcriptional control (Szcześniak and Makałowska, 2016). Pseudogene-derived lncRNAs represent an important source of ceRNAs as they are particularly enriched in miRNA response elements, as in case of the already mentioned PTENP1 (An et al., 2016). The subcellular balance between ceRNA, one or multiple miRNAs and mRNA targets constitutes a complex network allowing a fine-tuning of the regulation of gene expression during adaptation, stress response and development (Thomson and Dinger, 2016), (Tay et al., 2014).

- **Small ncRNA precursors**

Many lncRNAs host small RNA genes and serve as precursor lncRNA for shorter regulatory RNAs, in particular, those involved in the RNAi pathway (mi/si/piRNAs). Many lncRNAs were identified and functionally studied before their precursor function was known. Such is the case for H19, one of the first discovered lncRNA genes, and which contains two conserved microRNAs, miR-675-3p and miR-675-5p. In undifferentiated cells, H19 acts as a ribo-activator interacting with and promoting the activity of the ssRNA-binding protein KSRP (K homology-type splicing regulatory protein) to prevent myogenic differentiation (Giovarelli et al., 2014). During development, and in particular, during skeletal muscle differentiation, H19 is processed into miRNAs ensuring the post-transcriptional control of the anti-differentiation transcription Smad factors (Dey et al., 2014). Some piRNA clusters were found to map to lncRNA genes, mostly in exonic but also in non-exonic regions enriched in mobile elements thereby constituting putative pi-lncRNA precursors (Ha et al., 2014). Putative endo-siRNAs map to predicted hairpin RNA inverted repeats within lncRNA genes, but could also originate from any double-stranded lncRNA-RNA precursors that could be produced by sense-antisense convergent transcription (Carlile et al., 2009), (Werner, 2013). Endo-siRNAs have been documented in many eukaryotes, including fly, nematode and mouse. Overlapping

and bidirectional transcription is an abundant and conserved phenomenon among eukaryotes (Kapranov et al., 2007), (Wood et al., 2013). However, in mammals processing of sense-antisense paired transcripts into siRNA and their functional relevance is still controversial and requires experimental evidence, specifically at the single cell level. LncRNA processing into small RNA molecules could depend on different cellular machineries such as RNase P and RNase Z mediated cleavage of the small cytoplasmic mascRNA from MALAT1 (Wilusz et al., 2008) or Drosha-DGCR8 driven termination and 3'-end formation for lnc-pri-miRNAs (Dhir et al., 2015).

3.5. Classification according to associated biological processes

The examination of the non-coding transcriptome in different biological contexts of normal and pathological development has resulted in the discovery of lncRNAs specifically associated with particular biological states or pathologies. LncRNAs differentially expressed during replicative senescence represent senescence-associated lncRNAs, or SAL (Abdelmohsen et al., 2013). One such example, SALNR, is able to delay oncogene induced senescence by its interaction with and inhibition of the NF90 post-transcriptional repressor (Wu et al., 2015). Hypoxia, one of the classic features of the tumor microenvironment, induces the expression of many lncRNAs, in particular those from UCRs, named HINCUTs (Ferdin et al., 2013), (Choudhry et al., 2016). Oxidative stress induces the production of stress-induced lncRNAs, si-lncRNAs, that accumulate at polysomes in contrast to mRNAs, which are depleted (Giannakakis et al., 2015). Deep sequencing transcriptome analysis of mammalian stem cells identified non-annotated stem transcripts, or NASTs, that appear to be important for maintaining pluripotency (Fort et al., 2014). Finally, with the progression of clinical and diagnostic studies, a growing number of specific disease-associated lncRNAs have been detected. An example is the prostate cancer associated transcripts (PCATs), such as PCAT1 that were shown to have a role in cancer biology, but also as a potent prognostic marker (Prensner et al., 2011).

Chapter 2. Long non-coding RNAs as regulators of the epithelial-to-mesenchymal transition

LncRNAs have been shown to be differentially expressed in many different context and to have their expression actually be highly specific to biological processes or pathological variations. One of the pathologies in which lncRNAs are highly upregulated is cancer where they can promote tumor progression and metastasis formation, especially through the epithelial to-mesenchymal transition (EMT). In this chapter, we will first discuss the specific features of cancer and EMT as a driver of tumor progression. Then, we will briefly look at a few lncRNAs which have been associated with EMT and see how they can either activate or hinder the transition.

1. EMT as a driver of metastasis, drug resistance and tumor recurrence

1.1. Generalities on cancer

Cancer is a general term given to a collection of related genetic diseases represented by a tumor resulting from uncontrolled division of cells. It develops due to progressive transformation of somatic cells to neoplastic ones endowed with abnormal functions. Over 120 types of cancers are diagnosed in the human population across almost all tissues and organs. Even if each cancer type possesses its unique clinical, cellular and molecular traits, all cancer cells share ten common features, conceptualized in 2000 by D. Hanahan and R. Weinberg and updated in 2011: these processes allow cells to acquire the ability to grow and divide in an unrestrained way evading cell death and immune destruction, in addition to promote inflammation and invasion of surrounding tissues (figure C2-1). This process is accompanied by big chromosomal rearrangements and changes in the biochemical architecture of cells (Hanahan and Weinberg, 2011).

Cancer results from genomic instability with age-related incidence. Together with inherited genetic variations predisposing to cancer, it gives rise to genetic mutations and epigenetic alterations, which foster cancer hallmarks during all stages of tumorigenesis. The exceptional genetic complexity renders cancer difficult to diagnose at early stages of disease and to cure.

In cancer, it has been estimated that more than 90% of all cancer deaths are associated with metastasis (Chaffer and Weinberg, 2011). Investigation of the

molecular and biochemical basis of the metastatic process is, thus, fundamental for understanding of cancer biology, prognostic and treatment development.

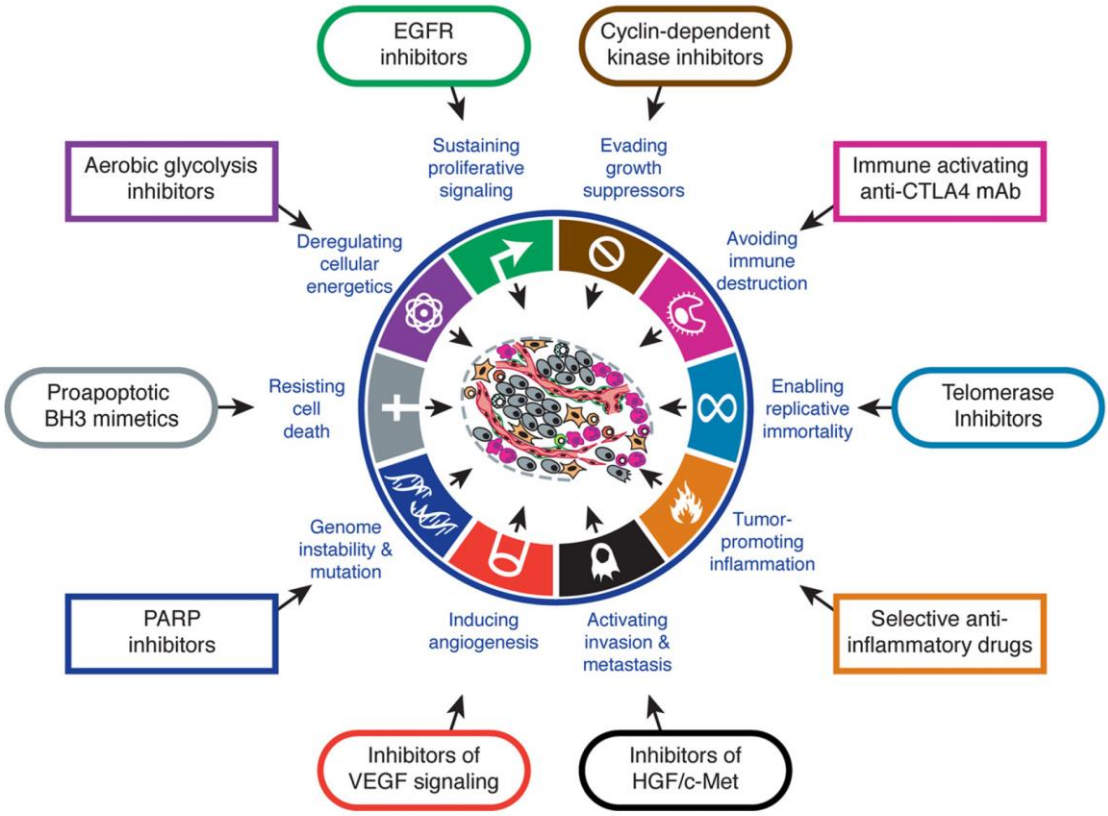


Figure C2-1. The ten hallmarks of cancer (Hanahan and Weinberg, 2011).

As a tumor grows in size, it stimulates the formation of new blood vessels that provide it with oxygen and nutrients, a process called angiogenesis. Eventually, cells of the original "primary" tumor can change their phenotypic and migratory properties to spread in the surrounding tissue, especially around nearby vessels. Tumor cells then enter the lymphatic and circulatory systems to colonize new tissues, forming "secondary" tumors (Kalluri and Weinberg, 2009). To successfully disseminate, cancer cells thus acquire particular properties of motility and invasion, ability to modulate the secondary site or local microenvironments to colonize secondary tissues.

Hence, cells of high metastatic potential are characterized by high plasticity and capable of changing their identity and communicate with other cells through complex gene expression reprogramming. In particular, activation of the epithelial-to-mesenchymal transition (EMT) has been described as a strategy adopted by

epithelial cancer cells to promote local invasion and dissemination at distant organs (Cano et al., 2000; Thiery, 2002).

1.2. EMT as a driver of metastasis and tumor progression

EMT is a highly dynamic biological process which consists of the reprograming of normal or neoplastic epithelial cells to gradually lose their differentiated epithelial characteristics including cell adhesion and polarity and to acquire mesenchymal traits enabling cytoskeleton reorganization and motility (Jolly et al., 2015; Lamouille et al., 2014a) (Figure C2-2).

As mentioned, distant metastasis is tightly associated to the capacity of cells to migrate and invade tissues, notably by breaking cell-cell junctions, remodel the cell matrix adhesion sites and the extracellular matrix as a whole. So far, a direct link between the transition itself and metastasis has yet to be fully demonstrated and actually remains debated (Jolly et al., 2017).

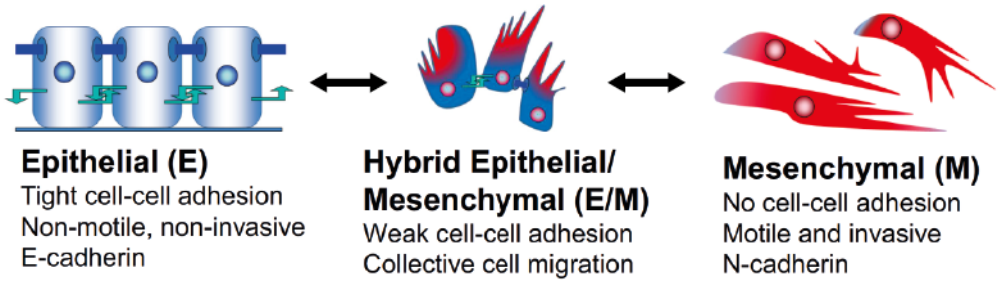


Figure C2-2. Epithelial to mesenchymal transition is a dynamic process with canonical morphological and phenotypic changes. Adapted from Jolly et al., 2015. Upon EMT, epithelial cells lose their specific features such as cell-cell adhesion and gain high motility and invasiveness.

The EMT defines the process through which epithelial cells become mesenchymal but its counter-process already exist as the mesenchymal-to-epithelial transition (MET). In a way, these two are essentially the two sides of the same coin and although they are often considered as mutually exclusive phenotypes, they are actually two ends of a very dynamic process. Full EMT is rarely achieved during the transition and there is a growing body of evidence regarding the role of hybrid states of EMT in cellular plasticity, tumor progression, stemness properties, drug resistance and tumor recurrence (Polyak and Weinberg, 2009; Santamaría et al., 2019; Ye and Weinberg, 2015).

In particular, epigenetic landscape and balance in expression of EMT genes may contribute to the residency of cells in an epithelial state, preserving or maintaining epithelial identity or, in the contrast, allowing partial or full transition to a mesenchymal state. According to the "EMT gradient model" (Ombrato and Malanchi, 2014), at the very early steps of transition it promotes stemness, while at advanced steps it allows to acquire fully mesenchymal features such as high migratory and invasive traits (Jolly et al., 2015, 2017). Hence, EMT/MET balance defines cell fate. Recently, these hybrid states have been observed in tumor models and confirmed to have strong phenotypic differences, especially regarding plasticity, stemness and metastatic potential (Pastushenko et al., 2018).

1.3. Molecular basis of the EMT

Under normal conditions, epithelial cells are characterized by their tight organization through the presence of many junction proteins such as EpCAM, β-Catenin, E-cadherin, Zonula-Occludens proteins, Occludins or Claudins, that link cells together through strong interactions. Also, a rigid cytoskeleton network of keratin and actin fibers maintains cell morphology as well as junction integrity.

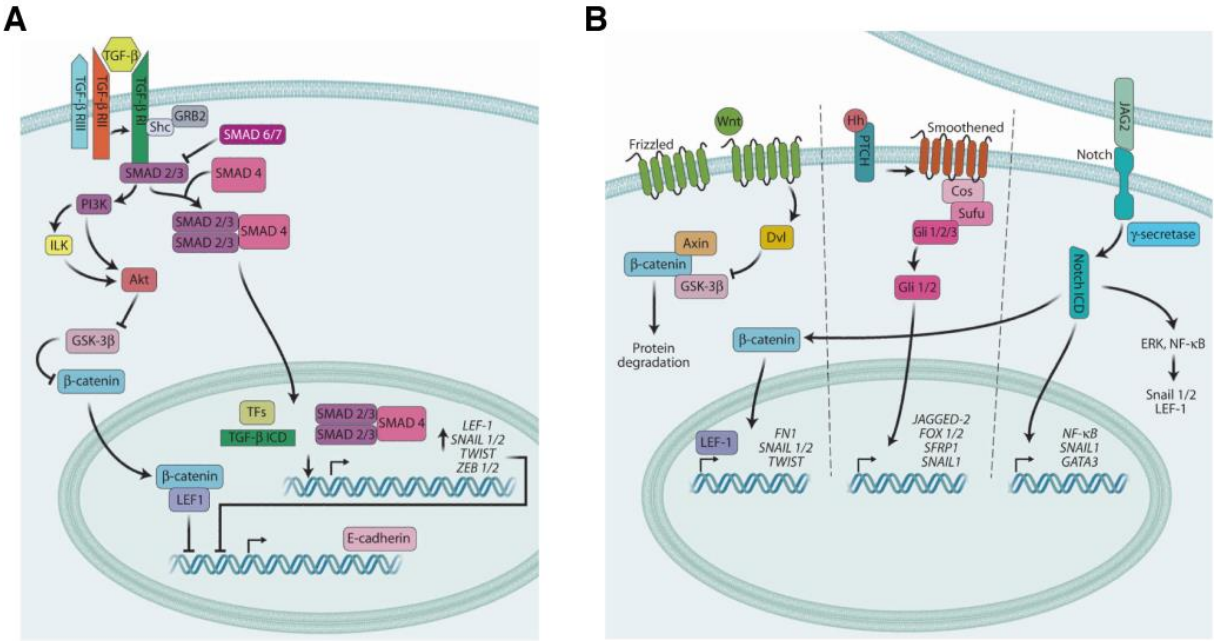


Figure C2-3. Signaling pathways regulating EMT. (A) Activation of EMT by TGFβ the SMAD2/3 cascade, thus inducing the expression of EMT-TF. **(B)** Wnt, Notch and Hedgehog pathways also induce the expression of EMT-TF. Adapted from Gonzalez and Medici, 2014.

The major player which orchestrates EMT is TGFβ. It activates SMAD-dependent or independent pathways (figure C2-3A), followed by several intracellular pathways including MAPK, PI3K, Hedgehog, Notch and Wnt (figure C2-3B). This triggers the expression of specific EMT transcription factors (EMT-TF), non-coding RNAs as well as various epigenetic modifiers.

Transcriptional factors such as zinc-finger E-box-binding (ZEB), Slug, Snail and Twist have been identified as master regulators of EMT, coordinating repression of epithelial genes and activation of mesenchymal genes (Lamouille et al., 2014a). The expression and action of these transcription factors can, in turn, be regulated at post-transcriptional level by RNAi pathway via mir200 and mir34 for example (Lamouille et al., 2013), but also through alternative splicing or epigenetically (De Craene & Berx, 2013). In the latter case, chromatin-modifying complexes, such as histone lysine methyltransferases (Polycomb), histone deacetylases (NuRD) and demethylases (Lsd1, PHF2) may determine the transcriptional activity of a genomic locus through covalent chromatin modifications and, as a consequence, may govern the epithelial-mesenchymal plasticity (Tam and Weinberg, 2013).

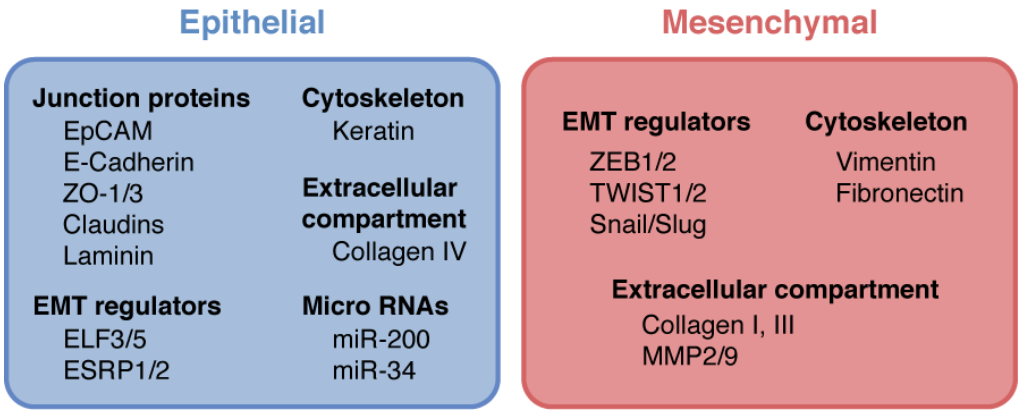


Figure C2-4. Key players and markers of the epithelial-to-mesenchymal transition in epithelial and mesenchymal cells. This includes: junction proteins EpCAM, E-Cadherin, ZO-1/3, Claudins or Laminin which are lost upon EMT; specific cytoskeleton components such as keratin in epithelial cells, vimentin and fibronectin in mesenchymal cells; components of the extracellular matrix such as collagens or metalloproteases MMP2/9; some key EMT regulators such as ELF3/5 and ESRP1/2 in epithelial cells and they EMT-TFs ZEB1/2, TWIST1/2 and Snail/Slug in mesenchymal cells; as well as microRNAs miR-200 and -34 in epithelial cells.

In addition to changes in gene expression, cells undergo drastic changes in morphology as a direct result of the reorganization of the cytoskeleton. For example, cytokeratin intermediate filaments are repressed while vimentin and fibronectin are

activated. All these changes promote cell migration and invasion, changing the interactions with the extracellular matrix and allowing its partial degradation via metalloproteases such as MMP2 and MMP9 (Mise et al., 2012; Zeisberg and Neilson, 2009). A summary of the markers for epithelial and mesenchymal cells is given in figure C2-4.

Although most studies rely on the induction of EMT through overexpression of EMT-TF, cellular stresses like UV or specific treatments such as TGFβ, meta-analysis have shown strong transcriptomic and epigenetic differences depending on the nature of the induction (Gröger et al., 2012a; Liang et al., 2016).

Although the main focus of EMT research has been on protein coding genes, an increasing number of studies support the role of long non-coding RNAs in its regulation, as well as its impact on tumor progression and metastasis.

2. LncRNAs associated with the EMT

LncRNA have been described to have a high specificity of expression compared to protein coding genes. Such specificity has been tightly linked to cell identity, pathological variations and particularly to cancer and EMT. This makes them good biomarkers for diagnosis and classification (Li et al, 2013). Many lncRNas have been shown to induce EMT and a few were actually shown to repress it, making them a new category of actors in the regulation of EMT.

2.1. Activators of EMT

Among the many lncRNAs upregulated in cancer, some of them have also been shown to be upregulated in various EMT models (Hu et al., 2014b; Liao et al., 2017). Here, I will discuss some examples of lncRNAs which were shown to activate EMT.

- **HOTAIR (HOX antisense intergenic RNA)**

First identified in 2007 (Rinn et al., 2007a), HOTAIR is a 2.2 kb lincRNA expressed antisense to the HOXC locus, between HOXC11 and HOXC12. Together with HOX A, B and D, the HOX C genes encode transcription factors involved in embryonic development. Among 231 HOX-originated ncRNAs, HOTAIR expression has been linked to cancer and metastasis. Indeed, HOTAIR is overexpressed in a wide variety of cancers such as breast, liver, lung, pancreas or colon cancers were it was associated with poor prognosis and/or invasion and metastasis (Geng et al., 2011;

Gupta et al., 2010a; Kim et al., 2013; Kogo et al., 2011a; Zhao et al., 2014). It is therefore a strong marker of tumor aggressivity and has been shown to be pro-oncogenic. A meta-analysis of 19 studies on tumors originating from various tissues showed HOTAIR to be a very efficient biomarker of poor prognostic and very low survival rate (Deng et al., 2014). Also, it has been suggested that HOTAIR is required for EMT to occur as well as for maintenance of the stemness of cancer cell lines (Pádua Alves et al., 2013).

Native HOTAIR adopts a single, well-defined conformation. It consists in 4 independent secondary structures that are highly stable, especially in the 5' half of the transcript (Somarowthu et al, 2015). Through its 5' and 3' extremities, HOTAIR has been shown to interact with epigenetic complexes PRC2 and Lsd1/CoREST/REST respectively, thus acting as a RNP repressor complex (figure C2-5). According to in vitro analysis, the truncation of nucleotides 1-300 and 1500-2146 abolish these interactions (Tsai et al., 2010a). It has been however suggested that PRC2 interacts with a longer portion (nt 1-530) of the transcript.

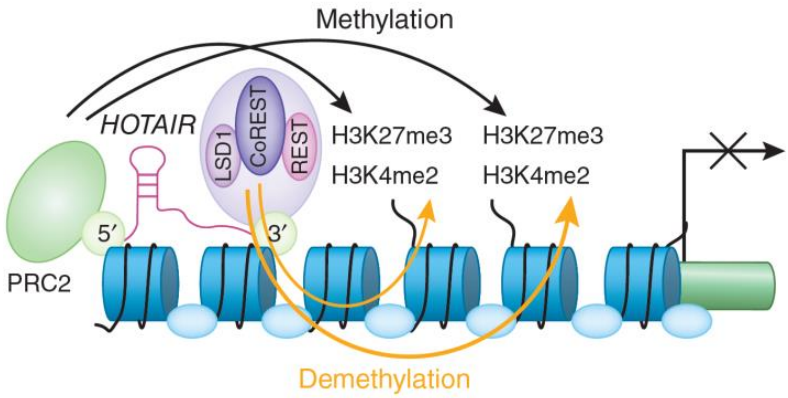


Figure C2-5. Mechanism of HOTAIR (pink)-mediated epigenetic modifications through its interaction with PRC2 (green) and Lsd1-CoREST-REST (purple) to induce transcriptional repression by chromatin (blue) modification (Croce, 2010).

PRC2 is involved in epigenetic regulation by methylation of H3K27 and although most of the initial studies regarding HOTAIR were done studying its interactions with PRC2, it has been recently suggested that HOTAIR-mediated regulation may be independent of PRC2, which could actually be recruited afterward (Portoso et al., 2017a; Qu et al., 2019). On the other hand, not much is known of HOTAIR's

interaction with the Lsd1/CoREST/REST complex despite their concomitant action to demethylate H3K4 (Li et al., 2013a).

Subject to many chromosomic rearrangements, cancer cells often undergo an epigenetic resetting and it has been shown that PRC2 subunits and Lsd1 are overexpressed in several types of cancer and have been linked to the promotion angiogenesis, metastasis and EMT (Berezovska et al., 2006; Ferrari-Amorotti et al., 2013). In addition, it was shown that HOTAIR may act in the cytoplasm to regulate the ubiquitination of certain proteins (Yoon et al., 2013b).

The HOTAIR-dependent pathways leading to metastasis as well as the genes it regulates are still unclear. Changes in expression, epigenetic modifications of hundreds of genes have been observed after HOTAIR knockdown or overexpression, in various systems.

- **MALAT1 (Metastasis-Associated Lung Adenocarcinoma Transcript 1)**

MALAT1 is the most prominent lncRNA associated to cancer metastasis. Its expression is remarkably increased in hepatocellular carcinoma, colorectal carcinoma, bladder cancer and lung cancer correlating with tumor metastasis potential and poor survival (Li and Chen, 2013). Originally described for its role in the formation of nuclear structures such as speckles and paraspeckles, MALAT1 depletion inhibits cell growth, cell cycle progression and invasion.

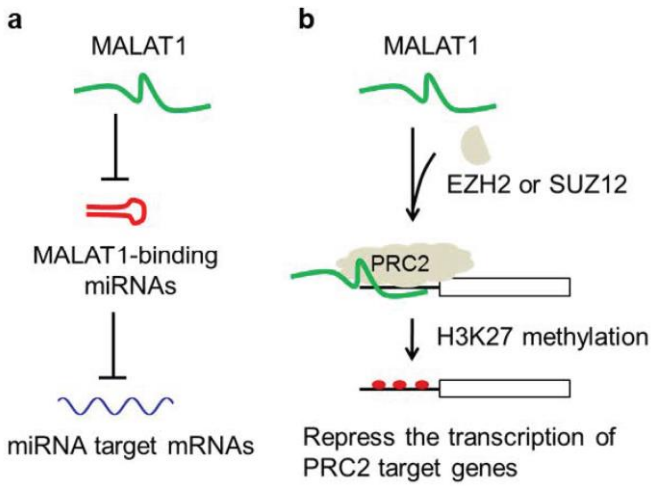


Figure C2-6. MALAT1 acts by interacting with (a) miRNAs (red) and epigenetic regulators such as (b) PRC2 (grey). From Sun and Ma, 2019.

At molecular level, MALAT1 has been shown to act as a ceRNA to block miR-205, miR-204 and miR-1, thus controling the levels of EMT-TF ZEB1, ZEB2 and SNAI2 (Ying et al., 2012) (figure C2-6a). It was also shown to modulate alternative splicing of some EMT-related genes through the repression of splicing factor RBFOX2. Although it is highly nuclear, it can also be processed into a smaller tRNA-like transcript mascRNA which is exported to the cytoplasm (Wilusz et al., 2008).
In addition, MALAT1 can also act as a scaffold for the recruitment of other transcription factors FOXN3 and SIN3N to also promote the expression of EMT genes in breast cancer, both *in vivo* and *in vitro* (Li et al., 2017b). It also interacts with Ezh2, the main component of the PRC2 epigenetic complex to induce epigenetic silencing of epithelial marker E-Cadherin (Hirata et al., 2015) (figure C2-6b).

- **PVT1 (Plasmacytoma variant translocation 1)**

The PVT1 lncRNA is transcribed from a frequently amplified region on chromosome 8 which contains the c-MYC oncogene where their expression is tightly correlated; indeed, PVT1 was shown to be essential for the maintenance of high levels of c-MYC (Tseng et al., 2014). In various cancer types, PVT1 expression has been correlated with metastasis and poor prognosis. Mechanistically, it was shown to promote EMT by regulating the SMAD2/3 phosphorylation and activation, part of the first steps of the TGFβ canonical induction of EMT (Zhang et al., 2018b). In pancreatic cancer cells, is associated with the repression of p21, a key player in the p53 pathway, thus promoting cell proliferation and EMT (Wu et al., 2017).

- **SNHG15 (Small Nucleolar RNA Host Gene 15)**

A more recent example of cancer- and EMT-associated lncRNA is SNHG15 which was first identified as a stress-responsive transcript (Tani 2013). Its expression has been linked to many types of tumors including gastric, breast, colorectal and renal clear cell carcinoma where it typically correlates with high proliferation, migration and invasion (Tong et al., 2019).
Interestingly, its first link was made in gastric cancer were it was shown to promote cell migration and invasion by regulating metalloproteases MMP2 and MMP9, two key markers of mesenchymal identity upon EMT (Chen et al., 2016). In breast cancer, its expression was also correlated with mesenchymal markers MMP2, MMP9, SNAI1

and VIM and it was actually shown to act as a ceRNA sponging miR-211-3p (Kong and Qiu, 2018). In colorectal cancer, it was shown to be upregulated by MYC and its knock-down inhibited proliferation, invasion, tumorigenicity and drug resistance (Saeinasab et al., 2019).
Since then, it has been shown to sponge a plethora of microRNAs, notably miR-141, to promote features associated with tumor progression and EMT pathways such as Wnt/β-Catenin (Liu et al., 2017a; Sun et al., 2019).

2.2. Repressors of EMT

- **GAS5 (Growth Arrest-Specific 5)**

Although most of the lncRNAs that have been associated with EMT are activators, there are a few examples of them which actually repress the transition. This is the case of GAS5 which was shown to suppress tumor proliferation, migration and the overall EMT phenotype in osteosarcoma by acting as a ceRNA for miR-221. In many types of human cancers, its expression is typically reduced and this lower expression correlates with increased tumor size and poor prognosis and it has been described a tumor suppressor lncRNA (Pickard and Williams, 2015).
Recently, GAS5 overexpression in pancreatic cancer was shown to actually reverse EMT, inducing a decrease in migration and invasion, as well as a repression of mesenchymal markers N-Cadherin, Vimentin and Snail, and an activation of epithelial marker E-Cadherin (Liu et al., 2018). This was also done through the repression of miR-221 which in turn cannot repress the expression of tumor suppressor SOCS3.
In addition to its function as a ceRNA, GAS5 acts as a decoy for the glucocorticoid receptor (GR). The GAS5-GR interaction prevents the receptor from binding to its response element, thus repressing GR-regulated genes and influencing many cellular functions such as cell survival and the response to apoptotic stimuli (Kino et al., 2010).

2.3. lncRNAs with controversial roles in EMT

The previous examples all typically tend to go toward one end of the EMT spectrum, either promoting (HOTAIR, MALAT1, PVT1, SNHG15) or inhibiting it (GAS5). However, some lncRNAs have been described to have very contradictory roles in the

regulation of EMT, both promoting and inhibiting it depending on the cellular context. Such lncRNAs display complex regulation and typically go toward tumor progression, regulating EMT-associated plasticity.

- **H19**

As mentioned in chapter 1, H19 was first identified as a maternally imprinted locus during development. Since then, many studies have linked it to cancer and tumor progression, making it a key oncofetal gene.

In normal conditions, p53 repressed the promoter of H19 which explains its overexpression in many types of cancer, where it correlates with metastasis and poor prognosis (Dugimont et al., 1998; Raveh et al., 2015). The regulation of H19 expression has been linked to fundamental hallmarks of cancer development such as genomic instability, hypoxic stress and high proliferative rates. H19 also suppresses apoptotic pathways and promotes the expression of genes involved in angiogenesis (Matouk et al., 2007). Thus H19 is a "super" oncogenic lncRNA promoting cancer progression at every stage.

H19 has been described to act in two ways: through its function as a host gene for miR675 and as a stand-alone RNA molecule which interacts with other microRNAs and proteins. miR-675 has been shown to downregulate many genes to regulate cancer and EMT pathways like Smad, Cadherins or TGFBI. On the other hand, H19 interact with proteins involved in transcription and epigenetic regulation such as MBD1 and PRC2 to guide them onto genomic targets (St Laurent et al., 2012). It can also sponge epithelial microRNAs such as miR-200 and miR-34.

In hepatocellular carcinoma, H19 transcription is activated by many EMT inducers such as TGFβ treatment and is actually essential for the activation of downstream EMT-TF such as Slug, setting a H19/Slug positive feedback loop. Upon overexpression, H19 induces a transcriptional shift from epithelial to mesenchymal markers as well as global cytoskeleton rearrangement, increase in migration and invasion. Similar findings were observed in bladder cancer (Luo et al., 2013).

In another hepatocellular carcinoma, contradictory data showed that mir-675 promoted MET by altering cell morphology, upregulating epithelial markers and downregulating mesenchymal markers such as the key EMT-TF Twist1 (Yuan et al., 2015b). In prostate cancer, mir-675 was reported to suppress the TGFβ-induced

transcript (TGFBI) which enhanced migration and invasion (The FANTOM Consortium, 2005b).
In their very comprehensive review, Raveh and colleagues suggest H19 modulates cell plasticity to either promote stemness (EMT) or differentiation (MET) as a mean to ultimately promote metastasis and tumor progression (Raveh et al., 2015).

- **MEG3 (Maternally Expressed Gene 3)**

Another example is MEG3 which has been reported to both promote and inhibit cancer progression. In lung cancer, MEG3 promotes a partial EMT by recruiting PRC2 to the promoter of the E-Cadherin and miR-200 genes locus to repress their transcription (Terashima et al., 2017). By contrast, it was shown to be repressed in gastric tumors compared to normal tissue and that it actually repressed migration *in vitro*, through the repression of mesenchymal markers such as metalloproteases and Snail (Xu et al., 2018).

Interestingly, these lncRNAs provide a cellular context for the regulation of typical EMT effectors, regulating both their expression and function in EMT. In this manuscript, I explore the relationship between lncRNAs in the subtler aspects of EMT regulation, through the mechanisms of HOTAIR interaction with Lsd1, and through the discovery of novel functionally relevant lncRNAs.

Objectives

Recently, the dogma that EMT is a strict switch from epithelial to mesenchymal is being questioned and hybrid EMT states have been shown to be relevant for pathological conditions as their phenotype vary in stemness, plasticity as well as migration and invasion properties. These traits are especially important in tumor progression, leading to drug resistance, metastasis and tumor recurrence.

Since lncRNA expression tends to be highly specific to cell identity, they were suggested to be good biomarkers for diseases and shown to regulate every stage of cancer development, in which they are often deregulated.

During my PhD, I decided to focus on defining the role of lncRNAs in the regulation of EMT by addressing first the identification of differentially-expressed lncRNAs upon EMT, define which ones are functionally relevant, what phenotypical changes they induce in the cells and finally through which mechanism do they act in this regulation.

To do so, we use an EMT system that was first developed in the group of Arturo Londoño (see chapter 3.1) and prior to my arrival, the former PhD of the lab Claire Bertrand found that the well-known lncRNA HOTAIR is upregulated upon EMT in this system. As a proof of concept for the study and discovery of lncRNAs in this system, I studied how the well-known lncRNA HOTAIR regulates EMT in close collaboration with Marina Pinskaya which lead to a first publication, see chapter 4 (Jarroux et al., 2019).

In this, we first aimed to define the role of HOTAIR's interacting domains with epigenetic modifiers PRC2 and Lsd1 through the overexpression of truncated variants of HOTAIR. We showed that the Lsd1-interacting domain is essential for the activation of cell migration, particularly through the transcriptional repression of genes involved in focal adhesion and interaction with the extracellular matrix. Although the PRC2-interacting domain seemed dispensable for the activation of migration, it was responsible for some transcriptomic changes.

Considering the importance of the Lsd1-interacting domain, we then asked if its recruitment onto the genome was modified upon HOTAIR expression. Interestingly, it seems HOTAIR actually promotes the relocation of Lsd1 from the promoter of its inherent genomic targets, resulting in a partial epithelial reprogramming. Depending on the cellular context, Lsd1 has been shown both promote and repress

EMT in the past, and HOTAIR seems to provide that context to promote a partial EMT.

In addition to studying the mechanism through which HOTAIR regulates EMT, my main project was to identify novel lncRNAs differentially expressed in EMT and do their functional characterization using a CRISPR-based transcriptional activation (CRISPRa) screening method. A manuscript is in preparation for this work and is presented here in chapter 5.

First, I aimed at doing a deep characterization of the non-coding transcriptome of EMT cells and used *de novo* lncRNA annotation coupled to a subcellular-fractionation approach to RNA-seq. This showed that chromatin-based RNA-seq allows for a better differential analysis of lncRNAs and a list of lncRNAs associated with the EMT in our system was established.

In order to define which of these novel lncRNAs were functionally relevant, I then performed a CRISPRa screen targeting the promoter of over 800 lncRNA genes to identify several ones which may regulated EMT, either by inducing a loss of epithelial identity (through the loss of the EpCAM surface marker by FACS), or a gain of mesenchymal indentity (through a gain in invasive and migratory properties by invasion assay).

On one hand, EpCAM-negative cells showed a strong enrichment for the CRISPRa guide-RNAs which targeted the most differentially-expressed lncRNAs in mesenchymal cells which I called MAL-1 for "Mesenchymal identity Associated LncRNA 1". On the other hand, the invasion-based screen did not seem to succeed and it will be discussed in the last part of chapter 5.

Although the validation experiments for MAL-1 in *cis* using CRISPRa are still ongoing, I characterized the expression and features of MAL-1 in our system. MAL-1 is a nuclear-enriched transcript associated with the mesenchymal identity. I then asked if it could act in *trans* as a stand-alone RNA molecule to regulated EMT and I generated epithelial cell lines which overexpressed it. In our system, MAL-1 expression correlates with the repression of epithelial markers such as EpCAM or other junction proteins, and increase the migratory properties of the cells.

Altogether, the validation of HOTAIR function as well as the identification of the novel MAL-1 shows lncRNAs represent an additional layer of regulation in the EMT, potentially promoting hybrid states.

MATERIAL AND METHODS

CHAPTER 3

Chapter 3. Methods

This chapter does not only encompass the protocols used during my PhD, it also contains a description of the in vitro system we use to study EMT (chapter 3.1) as well as the logic behind the use of a CRISPR-based transcriptional activation screen for the functional identification of lncRNAs (chapter 3.2). Unless mentioned otherwise, all NGS data processing were performed by Marc Gabriel, the bioinformatician of our team.

1. *In vitro* cell model to study EMT

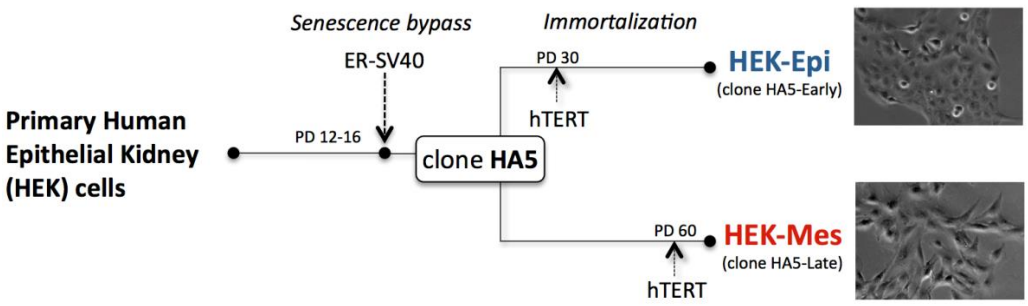


Figure C3-1. Diagram of the *in vitro* system used to study EMT (Castro-Vega et al., 2013a).

Most of the studies concerning EMT have been done on immortal epithelial cancer cell lines in which the transition is induced by stress conditions, specific treatments such as TGFβ, or overexpression of EMT-TF from the SNAI, ZEB or TWIST families. However, meta-analysis have shown strong transcriptomic differences depending on the nature of the induction (Gröger et al., 2012a; Liang et al., 2016).

In this study, we decided to take advantage of an *in vitro* system based on primary Human Epithelial Kidney (HEK) cells which was developed in the group of Arturo Londoño. In the life span of cultured cells, as telomere shortening and chromosome instability are initiated, cells start to naturally undergo EMT (Castro-Vega et al, 2013) (figure C3-1). A population of primary cells were maintained in culture after the bypass of senescence and were immortalized (hTERT) early in their life span, still at the epithelial state (Epi) or maintained in culture for 30 more population doublings as cells went through EMT, and were then immortalized at the

mesenchymal state (Mes). Therefore, molecular characterization and comparison between Epi and Mes cell lines allow the investigation of the EMT program in a highly stable in vitro system, without any specific treatments ensuring better insights into the EMT program naturally occurring during malignant transformation.

Upon EMT, Epi cells lose the expression of epithelial markers such as junction proteins EpCAM, ZO-1/TJP1, b-Catenin or Claudin and gain mesenchymal markers such as Vimentin and Fibronectin 1 and EMT-TF Slug, Snail or Zeb1. Phenotypically, Mes cells have increased migratory and invasive properties (figure C3-2).

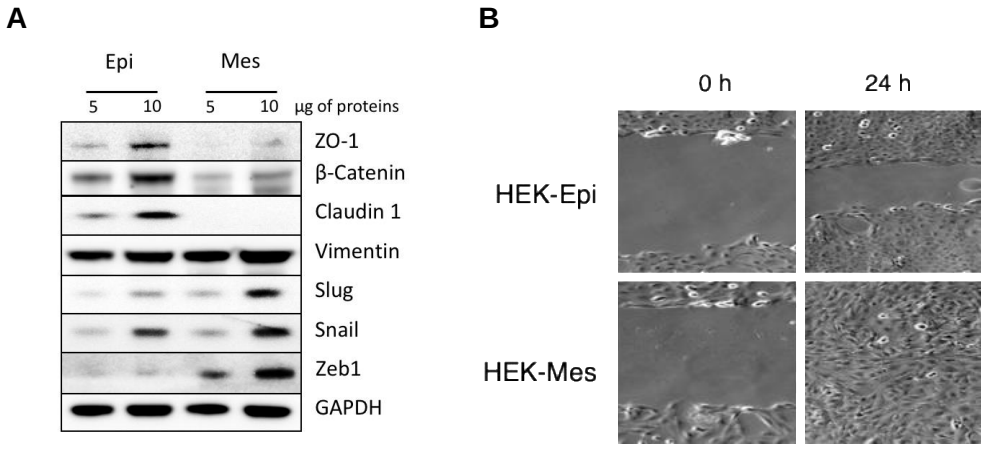


Figure C3-2. Phenotypic properties of Epi and Mes cells. (A) Western blot of EMT markers. (B) Wound healing assay over 24 hours to measure migratory properties.

2. CRISPR-based transcriptional activation screening

2.1. Generalities on CRISPR-based screens

In the last few years, Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) genome editing tools have developed tremendously from genome editing with the Cas9 nuclease to many other applications. Indeed, with the inactive dead (d)Cas9 fused to other effector proteins, it is now possible to use these technologies to target genomic loci to induce histone modification, DNA methylation, as well as transcriptional activation or repression (Montalbano et al., 2017) (figure C3-3).

Recently, some other CRISPR technologies have emerged to directly target RNA molecules and induce their degradations with Cas13 (Cox et al., 2017). All these technologies rely on the use of a RNP complex made up of a single guide (sg)RNA molecule and a protein effector (Cas9, the various dCas9, Cas13, etc.).

Through sequence complementarity of the sgRNA to its target, the complex can bind to specific genomic loci where the protein will then either cut the target DNA (Cas9) or recruit other effectors for transcriptional activation for example, as is the case of the fusion-protein dCas9-VP64 which recruits the p65-HSF1 proteins to activate transcription.

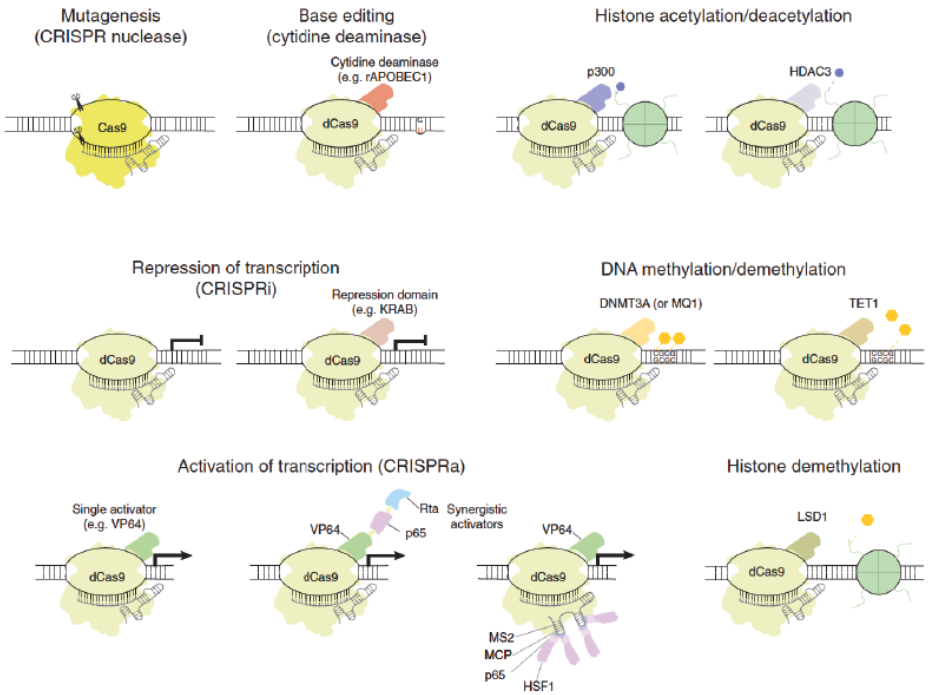


Figure C3-3. Different pooled CRISPR approaches using Cas9 and dCas9 to cut the target DNA loci, induce DNA/histone modification or activate/repress transcription. Figure from Montalbano et al., 2017.

This latter example is what is used here for our CRISPR-based transcriptional activation (CRISPRa) screen. These recent technologies are particularly interesting to investigate lncRNAs since mutation in their promoters or in the final transcript sequence may not yield any consequences for their function, unlike protein-coding genes. So far, CRISPR interference (CRISPRi) and CRISPRa screens have been successfully used to functionally characterize the non-coding transcriptome in a high throughput manner (Joung et al., 2017a; Liu et al., 2017b).

2.2. CRISPRa library cloning and phenotypic screening

The protocol used for CRISPRa screening goes on for several months and has been published in great details by Joung and colleagues. Here, we'll discuss a shorter version of it, as summarized in figure C3-4.

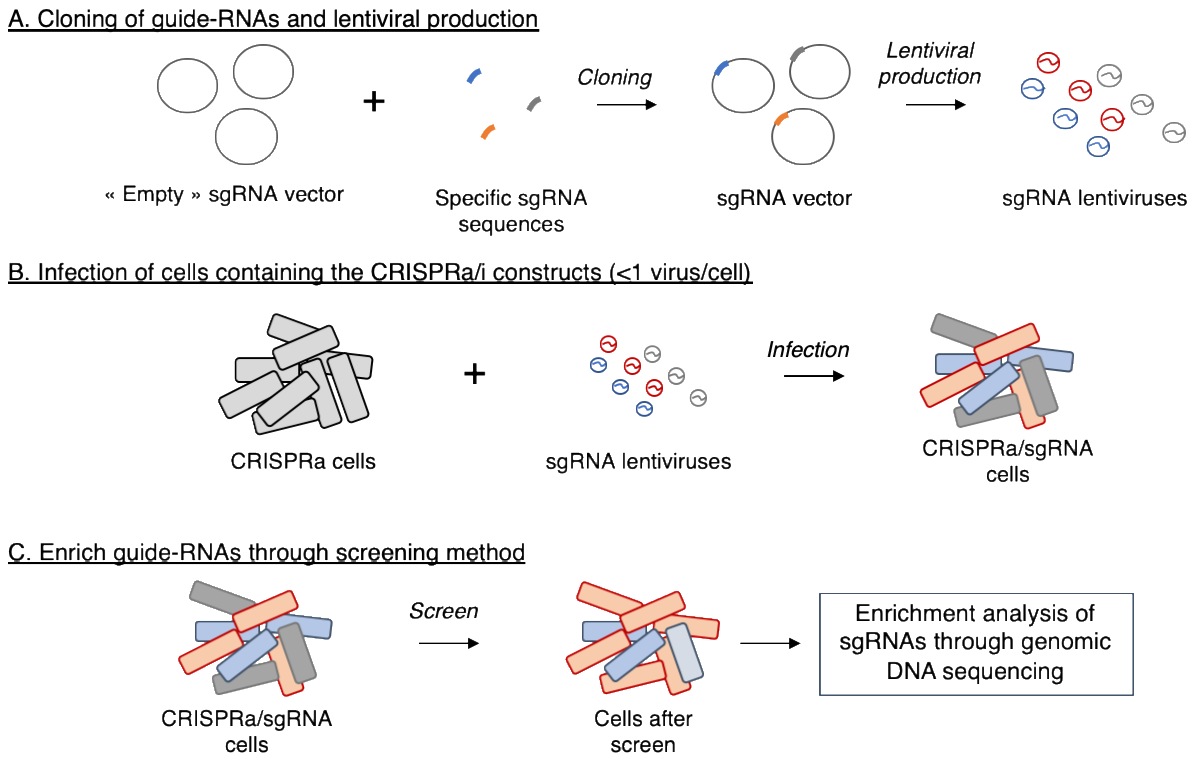


Figure C3-4. Overview of the CRISPRa screening strategy. (A) Cloning of sgRNAs in lentiviral vector by Gibson cloning followed by viral production. **(B)** Transduction of CRISPRa cells by spinfection at 0.3 MOI. **(C)** Cell screening by phenotypic selection for the functional enrichment of sgRNAs, followed by genomic DNA extraction and enrichment analysis.

These screens consist in cloning sgRNAs in lentiviral vectors as a pool. Once transduced in the cells constitutively expressing the CRISPRa machinery (dCas9-VP64 and MS2-P65-HSF1), a population of cells with a theoretically homogenous distribution of sgRNAs will be generated. Through various phenotypic methods, sgRNAs inducing changes in phenotypes would be enriched or depleted from the population. For example, if the CRISPRa screen is based on cell-proliferation, a sgRNA targeting a gene which represses cell proliferation or induces apoptosis would be depleted from the population, whereas a sgRNA targeting a gene which activates proliferation or represses apoptosis would be enriched in the population. The genomic DNA of the cells can then be extracted and after a targeted sequencing library preparation the sgRNA distribution can be measured. The methods described has been published by the Zhang laboratory and I performed it in collaboration with the group of Neville Sanjana (Joung et al., 2017b).

- **sgRNA library design**

The first step is the design of sgRNAs to target genes of interest. This was done by Meer Mustafa from Neville Sanjana lab using a confidential in-house pipeline from the TSS annotation defined for *de novo* lncRNAs validated through ChIP- and ATAC-seq as well as already annotated genes (see chapter 5). We aimed for 5 sgRNAs in the 200 basepairs upstream of the TSS. Our designed EMT-lncRNA pool consists of 3240 sgRNAs targeting *de novo* lncRNAs, 799 targeting GENCODE-annotated lncRNAs, 75 targeting protein-coding control genes and 100 non-targeting negative controls, for a total of 4214 sgRNAs. Flanking regions were added for Gibson cloning and an ssDNA oligo pool was ordered from Twist Bioscience.

- **sgRNA library cloning, transformation and verification**

To clone the sgRNA library, the oligo pool was amplified by PCR using the NEBNext High Fidelity PCR Master Mix (NEB) and the product was purified by electrophoresis on a 2% EX E-Gel (Invitrogen). The CRISPRa vector "lenti sgRNA(MS2)_zeo backbone" (Addgene 61427) was digested, dephosphorylated and gel-purified. The PCR mix and the digested vector were then clones using the NEB Gibson mix and purified using isopropanol precipitation.

Then, the cloned vectors were transformed in DUO Endura competent bacteria by electroporation and plated onto large LB-Carbenicilin plates. The estimation of expected clones was done using a titration method prior to the final transformation as it is extremely important to have a sufficient number of colonies before proceeding. The number to aim for is typically a 100 x representation of the library during cloning, which means that for a library of 4214 sgRNAs, 421.400 colonies are needed after transformation. Below this, the risk of losing some sgRNAs in the pool due to their low representation becomes greater at each step. To avoid this, a 500 x representation was aimed, with a minimum of $2.1 .10^6$ colonies.

After 12 hours at 37°C, plates were washed thoroughly with LB by pipetting and scraping the agar to retrieve as many bacteria as possible and the final vector pool was extracted using the Nucleobond Xtra Maxi Plus kit (Macherey-Nagel) according to manufacture instructions.

The integrity of the library was then checked by next-generation sequencing. For this, a specific sequencing library preparation was done as described in Joung et al.,

2017. Using R, we checked there were no skew in representation and the library seemed homogenous although some sgRNAs were totally absent. This is probably due to their absence in the pool of synthesized ssDNA oligos from the very beginning.

- **CRISPRa cell transduction**

To generate the CRISPRa-sgRNA cell line, HEK293T cells were transfected using the Jetprime Polyplus transfection reagent with the lentiviral packaging vectors pCMV-VSV and psPAX2, as well as the sgRNA-library vector. After viral titration, Epi_CRISPRa cells were spinfected at 0.3 MOI to ensure only one sgRNA gene could integrate per cell; this was done in duplicates and they were treated separately for the rest of the experiments. Cells were then selected using Zeocin over 5 days and passed before reaching 80% of confluence. Once again, making sure to infect enough cells to maintain a sufficient representation of the library (500X) is crucial, as well as a close monitoring of cell culture afterward to make sure that cells do not go beyond 80% of confluence, which could impact library representation.

- **Phenotypic screen**

After two days of antibiotic-free culture, cells were then screened using two distinct phenotypes. The first method relied on flow cytometry analysis of EpCAM for which cells were stained (EpCAM-APC, Miltenyi Biotech 130-098-118) a gate was defined for EpCAM-negative cells, thus targeting Epi cells losing epithelial identity. They were isolated using the BD Aria II flow cytometer, pelleted and frozen. The second method relied on an invasion assay through Boyden chamber. For this, commercial transwell inserts (Corning) were rehydrated according to manufacture instructions and an appropriate number of cells was seeded for screening. After 72 hours of invasion, the top of the membrane was cleaned using cotton swabs and the bottom of the membrane as well as the bottom of the culture dish were trypsinized to recover cells and put them back in culture for 2 extra days to ensure sufficient number of cells. In parallel to the start of the invasion assay, a sufficient number of cells was also kept in culture over the course of the experiment to act as a control to correct proliferation bias. Again, cells were then detached, pelleted and frozen.

Once all screening experiments had been performed, the genomic DNA was extracted using the classical isopropanol precipitation method, and sgRNA sequences were

amplified through a specific sequencing library preparation to measure their distribution in the population of cells.

- **sgRNA distribution analysis**

For analysis, each sgRNA was assigned a number of reads after sequencing and normalized on the total number of reads. Then, enrichment score was calculated as the ratios of normalized reads for a given condition to the initial control condition. Finally, the RIGER software was used to correlate sgRNA enrichment to the overall enrichment of targets genes in the screen (Luo et al., 2008).

3. General methods

- **Plasmids and oligonucleotides**

The plasmids used for the generation of the stable Epi-CRISPRa cells were plenti dCAS-VP64_Blast (Addgene, #61425) and MS2-P65-HSF1_GFP (Addgene, #61423). The CTR construct corresponds to pLenti_CMV_GFP (659-1) (Addgene, #17445). The MAL-1 constructs corresponds to the PCR-amplicon for MAL-1 amplified from the DNase-treated RNA from Mes cells, sub-cloned into pLenti CMV Blast DEST (706-1) (Addgene, #17451) (Campeau et al., 2009) using the Gateway system (Thermo Fisher Scientific). Oligonucleotides sequences for PCR and RTqPCR are available in the table below:

Target	Forward	Reverse
POLR2F	ATGTCGAGATCCTCCCCTCTG	GGCCTTGAGTTCCTTCATGG
RPL11	AGCAGCCAAGGTGTTGGAG	TACTCCCGCACCTTTAGACC
HOTAIR	GGTAGAAAAAGCAACCACGAAGC	GTGAGTGCCCGTCTTGCCCT
EAL-1	TTATTCCCGCGATGAGTTTC	CACCGAGACCACCCTTAAAC
EAL-2	TACCTTGTTGGCTCAGAACT	TGTGGATCCTTCTAGGTGTC
EAL-3	AGATGTCAAAGCACAAGCTC	AGAGATCAATGGTGTCCACT
MAL-1	TTCTTATCAGCCAGCCCAG	TTTGTCACAGCCCACAATG
MAL-2	TGGTGTAACACCTGGCAG	CATCTTCACTTGGGCAACAG
MAL-3	TCTCCTGAATTTTGTTTGCT	TGCAGTTTCATATGCGTCTA
GAPDH exon 3-4	AAAGCCTGCCGGTGACTAAC	ATGAAGGGGTCATTGATGGCA
GAPDH intron 3	GCTGCATTCGCCCTCTTAATG	GACAAGAGGCAAGAAGGCATGA
MALAT1	GAATTGCGTCATTTAAAGCCTAGTT	GTTTCATCCTACCACTCCCAATTAAT
XUT1150 (S. cerevisiae)	CTCAACGAGATGAGCCAACA	GCTTTTGCGGTTGTTATTCA
OCLN	CAGGACGTGCCTTCACCCCC	CCACCGCTGCTGTAACGAGGC
TJP3	GCCAGTTTCAAGCGCCCGGT	TCTGCAATCACCCGCACGGTG

FN1	GGTTTCCCATTATGCCATTG	TTCCAAGACATGTGCAGCTC
SNAI1	TGACCTGTCTGCAAATGCTC	CAGACCCTGGTTGCTTCAA
CLU	TGCGGATGAAGGACCAGTGTGA	TTTCCTGGTCAACCTCTCAGCG
CD44	TGCCGCTTTGCAGGTGTAT	GGCCTCCGTCCGAGAGA
PRICKLE1	GACAGTCTCTCCTCTTATCG	CTCTGCCTTTCCAAAATTCTTCAC
MTUS	AGCTTCGGGACACTTACATT	ATAGGCCTTCTTTAGCAATTC
EGR1	CTTCAACCCTCAGGCGGACA	GGAAAAGCGGCCAGTATAGGT
TGFB2	CCAAAGGGTACAATGCCAAC	CAGATGCTTCTGGATTTATGGTATT
CCND1	GTGTGCAGAAGGAGGTCCTGC	CCTCCTCGCACTTCTGTTCC
DNER	ATGCCAGTTCTAACAGCTCTGC	GGAGCACTGTTGGAATCCTGTGG

- **Cell lines and culture**

All cell-lines were cultivated in a humidified 5% CO_2 atmosphere at 37°C.

- HEK293T, A549, MDA-MB-231, MDA-MB-468 were cultivated in high-glucose DMEM with GlutaMAX, 10% Fetal Bovine Serum (FBS).
- Epi, Mes, Epi-CRISPR, MCF7, Epi_CTR and Epi_MAL-1 cells were cultivated in MEM alpha without nucleosides for HEK cells (Epi, Mes, Epi-CRISPRa, Epi_CTR and Epi_MAL-1) supplemented with 10% FBS, non-essential amino acids (NEAA) and sodium pyruvate.
- HCT116 were cultivated in McCoy medium supplemented with NEAA and 10% FBS.
- All cell lines were systematically tested and found negative for mycoplasma.

- **Cell-line generation**

For the generation of HOTAIR cell-lines, HEK293T cells were cultivated at 50-70% of confluence at T25 flasks were co-transfected with 1.3 µg of psPAX2 (Addgene, #12260), 0.8 µg of pVSVG (Addgene, #36399) and 0.8 µg of the lentiviral plasmid bearing cDNA of GFP (CTR), full-length (HOT) or truncated HOTAIR (HOTΔP and HOTΔL) and 5 µL of Lipofectamine® 2000 Transfection Reagent according to manufacturer's protocol (Thermo Fisher Scientific). Virus supernatant was recovered, filtered 48 h post-transfection and added to Epi cells at 50-70% of confluence. After 24 h post-transduction cells were sub-cultured every two days at a 1:4 ratio in 10 µg/ml of blasticidin supplemented MEM alpha medium for one week and then in MEMalpha medium for additional two weeks prior to any experiment.

For the generation of the Epi-CRISPRa cell-line, HEK293T cells were amplified to 50-70% confluence in T25 flasks and co-transfected with 1.3 µg of psPAX2

(Addgene, #12260), 0.8 µg of pVSVG (Addgene, #36399) and 0.8 µg of the dCas9-VP64, or the MS2-P65-HSF1_GFP lentiviral plasmids. Then, 200µL of Jetprime® buffer (Polyplus Transfection), mixed and supplemented with 9 ul of Jetprime® reagent. The mix was incubated at RT for 10 minutes and then added to the HEK293T cells. 48 hours after, viral supernatants were retrieved and 500 ul of dCas9-VP64 and 500 ul of MS2-P65-HSF1_GFP virus were added to a 50% P10 dish of Epi cells. After 10ug/ml Blasticidin selection over 5 days, cells were FACS-sorted based on GFP expression into a 96-well plate. Clones were amplified and one clone was selected as the Epi-CRISPRa cell line.
For the generation of the Epi_CTR and Epi_MAL-1 cell-lines, the same method was used with lentiviral plasmid bearing cDNA of the CTR or MAL-1 constructs. However, these two cell-lines were not obtained through clonal selection but maintained as a bulk-population after Blasticidin selection.

- **siRNA transfection**

siRNAs were transfected using the Jetprime® kit as well. In 6-well plate, 50% cells were washed and added fresh medium. A mix was prepared containing 2 ul of siRNA targeting MAL-1 or scrambled siRNA, 200ul Jetprime® buffer and 8ul Jetprime® reagent, vortexed and added to the cells. Cells were then incubated for at least 24 hours before RNA extraction or wound healing assay.

- **Wound healing assay**

Cells were cultured in triplicates in 6-well-plates to 95% confluence. Four scratches per well were made on each cell monolayer using a 10 µl pipette tip. Cells were then washed in PBS and cultured in fresh complete media. Wound images were taken at time 0 and 24 hours post-scratch with a Zeiss Axiovert 135 Microscope. The cell-free area was quantified using the TScratch software (Gebäck et al., 2009) for each photo and calculated as follows: Invaded area (%) = $(A_0 - A_{24})/A_0$
Results are presented as a mean ± SEM and p-values were calculated using the Student's t-test.

- **Colorimetric cell proliferation assay**

1-2*103 cells were seeded on 96-well-plates in 100 µl of culture medium (see 1.a) in triplicates for each cell line and each time point, and incubated at 37°C. Every day for 4 days and for each cell line, the quantity of cells was measured using the CellTiter 96 Aqueous One Solution Cell Proliferation Assay (G3582, Promega) according to manufacturer's instructions. Population doubling was calculated with the following formula: PD = (t2-t1)*[log10(2)/log10(A2/A1)]

- **Cell cycle analysis by Propidium Iodine staining.**

80% confluent cells were trypsinated, centrifugated and the pellet was well resuspended in 0.5 ml 1X PBS. To fix cells, 1.5 ml of ice-cold 100% ethanol were added drop by drop and incubated on ice for 15 minutes. Then, cells were centrifugated again for 10 min at 1400 rpm and resuspended in a mix of an RNAse cocktail and PI to a finale concentration of 40 ug/ml in 1X PBS. Cells were then incubated for 15 minutes in the dark, at room temperature and then sorted using a LSRII flow cytometer.

- **RNA extraction and reverse transcription**

Total RNA extraction was performed directly from cell cultures with miRNeasy kit according to the manufacture instructions (Qiagen). Only RNAs with the RNA Integrity Number (RIN) above 6 were used for further experiments. Reverse Transcription (RT) was performed on 500 ng of RNA with either random and oligo(dT) primers mix (iScriptTM cDNA Synthesis Kit) or specific oligonucleotides (SuperScript II Reverse Transcriptase, Thermo Fisher Scientific) (Table S5). Reactions without reverse transcriptase were included as a negative control for DNA contamination.

- **Quantitative PCR analysis**

For quantification of cDNA in RT experiments, RT reactions were diluted 10-40 times in water. 5 µl of each undiluted RT were pooled together and used to make 8 samples of reference standards corresponding to two fold serial dilutions.

Each qPCR reaction was performed in duplicates on the Roche Light cycler 480III. Relative values for cDNA/RNA amount in each sample was extrapolated from the standard curve generated from the reference standards using LightCycler480

Software, reported to the POLR2F or RPL11 mRNA and correspond to the mean ± standard deviation of three independent biological experiments.

- **PolyA pull-down**

Isolation of poly(A) RNA was performed using PolyATract® mRNA Isolation Systems (Promega) according to manufacture instructions. 5 µg of total RNA were incubated for 10 minutes in 65°C prior the Biotinylated-Oligo (dT) Probe annealing. The mix was incubated again for 10 minutes at room temperature. Streptavidin MagneSphere® Paramagnetic Particles (SA-PMPs) were then added and incubate 10 minutes at room temperature, washed 4 times in 0.1X SSC buffer. PolyA-RNAs were eluted in RNase-free water and purified once more by isopropanol/sodium acetate precipitation.

- **Terminator assay**

The exonuclease treatment was accomplished using Terminator™ 5' –Phosphate-Dependent Exonuclease (Epicentre) according to manufacture instructions. As input, 4 µg of total RNA were mixed to 1 µg RNA of S. Cerevisiae XRN1Δ as a spike-in control (using the yeast lncRNA XUT1150 as a RTqPCR control of non-capped transcript). The 5 µg of RNA were mixed with 2 µl Terminator 10X Reaction Buffer A, 1 µl of Terminator Exonuclease (or without for "non treated"), 0.5 µl RiboGuard RNase Inhibitor and RNase-free water to a final volume of 20 µl. The reaction mix was incubated at 30°C for 60 minutes and purified by phenol extraction and ethanol precipitation.

- **Subcellular fractionation for Cyto- and Chro-seq**

This protocol was adapted from Gagnon et al., 2014. All steps were performed on ice or at 4°C with ice-cold buffers supplemented with 25µM α-amanitin and 20 U/µl SUPERase-IN (AM2696, Ambion) for RNA extraction or 0.1mM AEBSF (A8456, SIGMA) for protein extraction. For each recovered fraction, 1 ml of RNA Precipitation Solution RPS (150 mM sodium acetate in 100% ethanol) was added to the RNA samples and kept at -20°C; and NaCl to a final concentration of 150 mM was added to the protein samples and kept on ice.

Cells were cultured until 90% confluence on T150 flasks. After media removal, cells were washed twice in PBS, scraped and recollected in Falcon tubes. They were divided to have approximately 10^7 cells per tube, pelleted by centrifugation at 1500 rpm for 5 min at 4°C and resuspended in 380 µl of Hypotonic Lysis Buffer HLB (10mM Tris HCl pH 7.5, 10 mM NaCl, 3 mM MgCl2, 0.3% NP-40, 10% glycerol). Cells were then incubated on ice for 10 min. After brief vortexing, lysates were centrifuged at 1000g for 3 min at 4°C and the cytoplasm-containing supernatant was recovered.

Nuclei pellets were washed 3 times in 1 ml HLB by gently pipetting once and centrifugated at 300g for 2 min at 4°C. Then, nuclei were resuspended in 380 µl Modified Wuarin-Schibler buffer MWS (10 mM Tris HCl pH 7, 300 mM NaCl, 4 mM EDTA, 1% NP-40, 1M urea), vortexed for 30 seconds and incubated on ice for 5 min. They were vortexed again for 30 seconds and put back on ice for 10 min. After centrifugation at 1000g for 5 min at 4°C, nucleoplasm-containing supernatant was recovered.

Chromatin pellets were washed 3 times in 1 ml MWS by quick vortexing and centrifugation at 500g for 3 min at 4°C. For RNA extraction, 700 ml Qiazol was added to the pellet and briefly vortexed. EDTA was then added to a finale concentration of 5 mM and incubated at 65°C for 10 min with regular vortexing in order to resuspend the chromatin pellet as much as possible. For protein extraction, 100 µl Nuclear Lysis Buffer NLB (20 mM Tris HCl pH 7.5, 3 mM MgCl2, 150 mM KCl, 0.3% NP-40, 10% glycerol) was added to the pellet and samples were sonicated for 15 min (30s on, 30s off, "high power" mode) on the Bioruptor Plus (Diagenode).

For RNA samples, cytoplasmic and nucleoplasmic fractions in RPS solution at -20°C were centrifuged at 16000g for 15 min at 4°C. RNA pellets were resuspended in 700 ul Qiazol followed by RNA extraction according to the manufacturer's instructions.

For protein samples, all 3 fractions supplemented with NaCl were centrifuged at 16000g for 20 min at 4°C. Protein-containing supernatants were recovered and 5 µl were used to measure protein concentration with the Pierce™ BCA Protein Assay Kit (23225, Thermo Scientific) according to the manufacturer's instructions. The remaining samples were stored at -20°C.

- **RNA-seq library preparation**

1 µg of RNA was depleted for ribosomal RNA with the RiboMinusTM Eukaryote Kit for RNA-seq (Thermo Fisher Scientific) and converted into cDNA library using a TruSeq Stranded Total Library Preparation kit (Illumina). cDNA libraries were normalized using an Illumina Duplex-specific Nuclease (DSN) protocol prior to a paired-end sequencing on HiSeq™ 2500 (Illumina). At least 20x coverage per sample was considered as minimum of unique sequences for further data analysis. Raw RNA-seq data from the HOTAIR article is available at Gene Expression Omnibus (GEO) under accession number GSE106517.

- **RNA-seq data analysis**

Reads were mapped allowing 3 mismatches using TopHat 2.0.4 (Trapnell et al., 2009) and the human genome version hg19. Uniquely mapped reads were assembled using the BedTools suite (Quinlan and Hall, 2010) and merged in segments if mapped in the same strand to the Gencode V15 (chapter 4) or v27 (chapter 5) annotation to extract protein-coding genes and annotated noncoding genes including lncRNA, antisense, sense_intronic, sense-overlapped and pseudogenes. Finally, differential expression analysis was performed using DESeq (Anders and Huber, 2010) and gene ontology analysis were done using DAVID and GSEA webservers.

RESULTS

CHAPTER 4
A role for HOTAIR in the EMT

"HOTAIR promotes an epithelial-to-mesenchymal transition through relocation of the histone demethylase Lsd1."

Publication n°1

CHAPTER 5
Functional discovery of novel lncRNAs in the EMT

"CRISPRa screen of chromatin-enriched lncRNAs reveals a new regulator of epithelial identity."

Publication n°2

Chapter 4. A role for HOTAIR in the EMT

1. Introduction

In collaboration with the laboratory of Arturo Londoño-Vallejo who first initiated the system to study EMT (Castro-Vega et al., 2013b), our laboratory initiated the study of lncRNAs in EMT. Focusing on already annotated and known lncRNAs, we performed Total RNA-seq and differential expression analysis in epithelial and mesenchymal cells. Among the lncRNAs upregulated in Mes cells is the well-known HOTAIR which has been associated to cancer progression and metastasis. It has been described to act as a scaffold through its 5' and 3' structural domains for the recruitment of chromatin-modifying complexes PRC2 and Lsd1-CoREST-REST respectively, in order to repress gene expression.

From there started the PhD work of Claire Bertrand prior to my arrival. She notably cloned the full-length HOTAIR and versions truncated for its 5'- and 3'-end domains and first showed that the full-length HOTAIR induces an increase in migration in our system. Under the direct supervision of Marina Pinskaya, I continued this work during my master and then my own PhD as a side project, studying the truncated versions of HOTAIR and their role in EMT.

We showed that the 3'-end domain of HOTAIR which interacts with Lsd1 is essential for HOTAIR-mediated activation of cell migration, particularly through the repression of genes involved in focal adhesion and interaction with the extracellular matrix. ChIPseq of Lsd1 showed that HOTAIR overexpression induces a relocation of Lsd1 from its inherent genomic loci, resulting in partial epithelial reprogramming. Our results thus show how HOTAIR modulates the role of Lsd1 as a guardian of the epithelial identity.

For the sake of manuscript clarity, the method section was merged to chapter 3. The full article is available online on bioRxiv; DOI: https://doi.org/10.1101/724948.

2. Publication n°1

HOTAIR promotes an epithelial-to-mesenchymal transition through relocation of the histone demethylase Lsd1

Julien Jarroux[1], Claire Bertrand[1], Marc Gabriel[1], Dominika Foretek[1], Zohra Saci[1], Arturo Londoño-Valejo[2], Marina Pinskaya[1€] and Antonin Morillon[1€]

[1] ncRNA, epigenetic and genome fluidity, CNRS UMR3244, Sorbonne Université, PSL University, Institut Curie, Centre de Recherche, 26 rue d'Ulm, 75248 Paris, France

[2] Telomeres and cancer, CNRS UMR3244, Sorbonne Université, PSL Université, Institut Curie, Centre de Recherche, 26 rue d'Ulm, 75248 Paris, France

Contact:
marina.pinskaya@curie.fr
antonin.morillon@curie.fr
[€] co-corresponding

Summary
Epithelial-to-mesenchymal transition (EMT) drives a loss of epithelial traits by neoplastic cells enabling metastasis and recurrence in cancer. HOTAIR emerged as one of the most renowned long noncoding RNAs promoting EMT mostly as a scaffold for PRC2 and repressive histone H3 Lys27 methylation at gene promoters. In addition to PRC2, HOTAIR interacts with the Lsd1 lysine demethylase, a known epigenetic regulator of cell fate during development and differentiation. However, Lsd1 role in HOTAIR function is still poorly understood. Here, through expression of truncated variants of HOTAIR, we revealed that, in contrast to PRC2, its Lsd1-interacting domain is essential for acquisition of migratory properties by epithelial cells. HOTAIR induces Lsd1 relocation from its inherent genomic loci hence reprogramming the epithelial transcriptome. Our results uncovered an unexpected role of HOTAIR in EMT as an Lsd1 effector and pointed to the importance of Lsd1 as a guardian of the epithelial identity.

Highlights

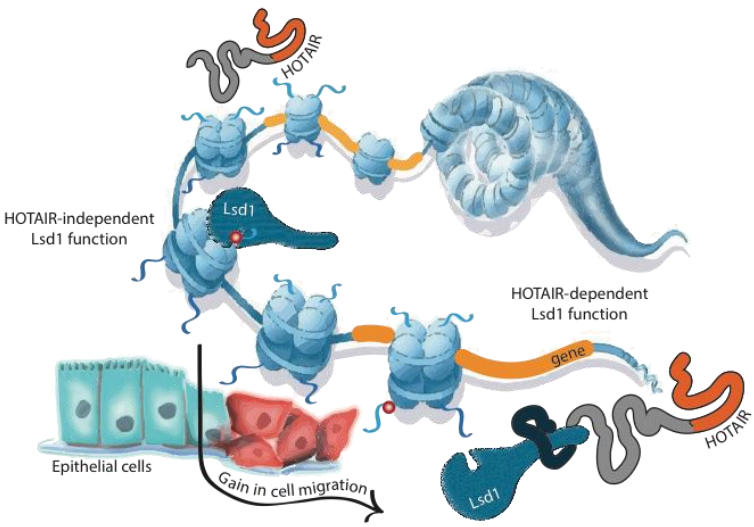
- HOTAIR promotes migration of immortalized normal epithelial cells.
- Lsd1-interacting domain, but not PRC2, is essential for HOTAIR function.
- When expressed, HOTAIR reshuffles Lsd1 from its inherent genomic locations.
- Lsd1 dislocation switches gene expression pattern in favor of mesenchymal identity.

eTOC Blurb

HOTAIR is a long noncoding RNA scaffolding PRC2 and Lsd1 chromatin modifiers to repress transcription, promote cell migration and tumor metastasis. Jarroux et al. reveal that HOTAIR acts independently of PRC2 by genome-wide reshuffling of Lsd1 chromatin occupancy and disrupting its function in maintenance of epithelial identity.

Running title: HOTAIR as Lsd1 molecular switch

Graphical abstract



INTRODUCTION

The epithelial-to-mesenchymal transition (EMT) allows normal or neoplastic cells to gradually lose their differentiated epithelial characteristics including cell adhesion and polarity, and to acquire mesenchymal traits enabling cytoskeleton reorganization and motility (Lamouille et al., 2014a). EMT is closely linked to carcinogenesis since it progressively endows epithelial cells with multiple properties

required for invasion and metastasis, but also for acquisition of stem-like properties contributing to tumor recurrence and drug resistance (Ye and Weinberg, 2015). This dynamic and reversible process is driven by complex changes in signaling circuits and reprogramming of gene expression. Transcriptional factors such as zinc-finger E-box-binding (ZEB), Slug, Snail and Twist have been identified as master regulators of EMT, coordinating repression of epithelial genes and activation of mesenchymal genes (Lamouille et al., 2014a). The expression and action of these transcription factors can, in turn, be regulated at post-transcriptional level by RNAi pathway (Lamouille et al., 2013), but also epigenetically. In the latter case, chromatin-modifying complexes, such as histone lysine methyltransferases (Polycomb), histone deacetylases (NuRD) and demethylases (Lsd1, PHF2) may determine the transcriptional activity of a genomic locus through covalent chromatin modifications and, as a consequence, may govern the epithelial-mesenchymal plasticity (Tam and Weinberg, 2013). In particular, epigenetic landscape and balance in expression of EMT genes may contribute to the residency of cells in an epithelial state, preserving or maintaining epithelial identity or, in the contrast, allowing transition to a mesenchymal state.

An increasing number of examples supports the involvement of long noncoding (lnc)RNAs in the EMT and metastasis (Huarte and Marín-Béjar, 2015), (Liang et al., 2018), (Shi et al., 2015), (Richards et al., 2015). These RNA polymerase II transcripts of at least 200 nucleotides long and of any or low coding potential can intervene in regulation of gene expression in the nucleus through RNA-protein or RNA-DNA pairing mechanisms, scaffolding and guiding chromatin modifying complexes to specific genomic locations (Quinn and Chang, 2015), (Hendrickson et al., 2016), (Morlando et al., 2014). LncRNAs often show cell- and tissue-specific expression and are highly deregulated in cancers. However, molecular mechanisms underlying lncRNAs dysregulation and action remain largely unknown. Among the most prevalent cancer-associated lncRNAs is HOTAIR (for HOX transcript antisense intergenic RNA). Clinical studies have clearly shown its overexpression in most human cancers and its association with poor prognosis, metastasis and acquisition of stemness (Tsai et al., 2011), (Kogo et al., 2011b), (Li et al., 2017a), (Gupta et al., 2010b), (Zhang et al., 2015). HOTAIR has been firstly identified in human fibroblasts as a molecular scaffold RNA, responsible for epigenetic regulation of cell fate during

differentiation (Rinn et al., 2007b). Indeed, the majority of nuclear HOTAIR functions have been attributed to its interaction with the Polycomb repressive complex 2 (PRC2) and Histone H3 Lys27 (H3K27) methylation of EMT genes promoters in *trans* (Kogo et al., 2011c), (Gupta et al., 2010c). If the exact mode of the lncRNA targeting to genomic loci remains unclear, the molecular outputs are highly cell-type specific. In hepatocytes, HOTAIR has been reported to mediate a physical interaction between the Snail1 transcription repressor and the Enhancer of Zeste Homolog 2 (EZH2) subunit of PRC2, guiding both to specific loci for regulation of hepatocyte *trans*-differentiation program (Battistelli et al., 2016). However, some publications have also demonstrated that PRC2 promiscuously interacts with many structured coding and noncoding RNAs and have claimed PRC2 dispensability for HOTAIR-mediated transcriptional repression (Kaneko et al., 2013), (Kaneko et al., 2014), (Portoso et al., 2017b). Instead, HOTAIR have been proposed to play a role in anchoring PRC2 at specific repressed loci, though the ultimate action of HOTAIR and its protein cofactors are still not fully depicted.

The full length HOTAIR is 2.1 kilonucleotides long and has a modular secondary structure (Somarowthu et al., 2015). In addition to PRC2 binding to first 300 nucleotides of the Domain 1, HOTAIR within its last 500 nucleotides contains another independent domain associated with the Lsd1/REST/CoREST complex (Wu et al., 2013), (Tsai et al., 2010b), (Somarowthu et al., 2015). Lsd1/KDM1A, the lysine specific demethylase-1, has been proposed to demethylate H3K4me2 and to reinforce HOTAIR/PRC2-mediated repression of transcription. Chromatin immunoprecipitation (ChIP) and isolation by RNA purification (ChIRP) allowed identification of GC-rich regions of Lsd1 binding sites and a GA-rich consensus sequence for HOTAIR targeting in epithelial cancer cells (Tsai et al., 2010b), (Chu et al., 2011). However, little is known of whether and how Lsd1 contributes to HOTAIR action. Lsd1 is a well-known epigenetic regulator of EMT and cancer with, in few cases, a tumor suppression function (Wang et al., 2009), but mostly playing an oncogenic role (Hino et al., 2016), (Harris et al., 2012), (Sun et al., 2016), (Lim et al., 2010), (Schenk et al., 2012), (Feng et al., 2016). The functional duality of Lsd1 can be attributed to the versatility of its substrates and of Lsd1 interacting partners in different biological contexts (Shi et al., 2004a), (Metzger et al., 2005a), (McDonald et al., 2011a). Indeed, in mouse hepatocytes Lsd1 was reported to control the

establishment of large organized heterochromatin H3K9 and H3K4 domains (LOCKs) across the genome during EMT. Large scale immunoprecipitation has revealed that Lsd1 interacts with REST/coREST co-repressors in differentiated epithelial cells, as though in TGFβ treated cells undergoing EMT Lsd1 is mostly associated with transcriptional co-activators including several catenins (McDonald et al., 2011a). In addition, non-histone targets, such as p53 and DNMT1, and non-enzymatic, scaffold roles have been proposed for Lsd1, particularly, in regulation of enhancer activity in mammals (Lan et al., 2007a), (Lan et al., 2007b), (Zeng et al., 2016), (Wissmann et al., 2007), (Wang et al., 2001), (Huang et al., 2007), (Scoumanne and Chen, 2007), (Roth et al., 2016). Pharmacological inhibitors of Lsd1 impairing its catalytic activity, unexpectedly, have been shown to act through disruption of its scaffold function, particularly with SNAG domain containing proteins, such as the transcriptional repressor GFI (Maiques-Diaz et al., 2018). Whatever the molecular basis of Lsd1 action may be, the biological outcome depends on a balance between activated and repressed genes underlying the pivotal role of Lsd1 in the phenotypic plasticity of a cell.

In the present study, we aimed to understand a role for HOTAIR interaction with Lsd1 in the EMT reprogramming. For this purpose, we used gain- and loss of function approaches overexpressing HOTAIR in immortalized primary epithelial cells and disrupting HOTAIR interactions with chromatin modifying complexes by deletion of either the 5'-PRC2 or 3'-Lsd1-interacting domains within the lncRNA. As expected, HOTAIR promoted migration of epithelial cells; however, this required the presence of the Lsd1-interacting domain while the PRC2 one was dispensable. At molecular levels, epithelial cells expressing the HOTAIR variant truncated for the Lsd1-interacting domain expressed more and showed less diffused outer membrane distribution of the tight junction protein ZO-1/TJP1, compared to the full-length and the truncated for PRC2 HOTAIR variant. Genome-wide Lsd1 profiling confirmed that the expression of HOTAIR with the intact 3'-extremity induces dramatic changes in chromatin distribution of Lsd1. We propose that HOTAIR, when expressed in epithelial cells, promotes a displacement of Lsd1 from its inherent targets resulting in transcriptomic changes in favor of mesenchymal traits. Our findings pinpoint Lsd1 as a guardian of epithelial identity and support a PRC2-independent function of

HOTAIR in acquisition of migratory properties by epithelial cells at very early steps of carcinogenesis.

RESULTS

Generation of Epi cell lines expressing full-length and truncated variants of HOTAIR

To decipher a role of the Lsd1 interacting domain in HOTAIR function, our rational was to generate expression vectors containing the lncRNA as a full-length transcript (HOT), but also truncated for the first 300 or the last 500 nucleotides sequences, previously reported to be involved in PRC2 and Lsd1 interactions (HOTΔP and HOTΔL), respectively (Figure 1A). These constructs were transduced into immortalized human epithelial kidney cells, HA5-Early. This cell line, originally obtained from a primary kidney epithelium by ER-SV40 and hTERT transformation very early in their lifespan, is characterized by normal karyotype and epithelial traits, such as rounded cobblestone morphology, low migration and expression of epithelial markers (zonula occludens-1/ZO-1, β-catenin, claudin-1) (Figure S1A-S1C) (Castro-Vega et al., 2013b). To facilitate further reading, the HA5-Early cell line is referred to as Epi, and its derivatives as Epi-CTR, Epi-HOT, Epi-HOTΔP and Epi-HOTΔL.

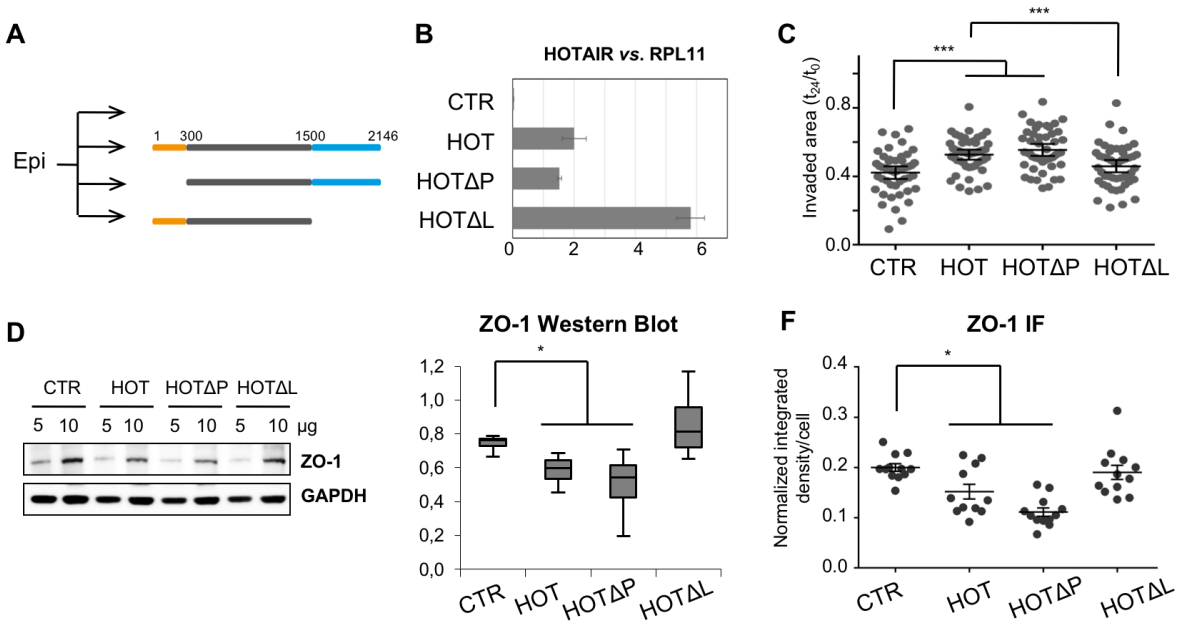


Figure 1. HOTAIR expression in epithelial Epi cells promotes cell migration in Lsd1-dependent manner:

(A) Stable Epi cell lines overexpressing CTR, full-length and truncated variants of HOTAIR lacking PRC2 or Lsd1-interacting domains, HOTΔP and HOTΔL,

respectively; **(B)** Random-primed RT-qPCR measurement of HOTAIR expression in Epi cell lines. cDNA levels are presented as a mean ± standard deviation (SD) for at least three biological replicates; **(C)** Quantification of the wound area invaded in 24 hours by Epi cells as a mean ± Confidence Interval (CI) of 95%; **(D)** Abundance of ZO-1 in Epi cells assessed by Western blot of whole protein extracts; **(E)** ImageJ quantification of ZO-1 in four independent Western blot experiments; **(F)** ZO-1 quantification of IF images performed using the Fiji software and bar-plotted as normalized integrated densities per cell with as a mean ± standard error of the mean (SEM) for at least 11 high-field units representing at least 100 cells; * p-value < 0.05, Student's t-Test.

Expression levels of all HOTAIR variants were measured relative to the housekeeping protein-coding gene *RPL11* showing stable expression in all experimental conditions. For comparison, we also used the mesenchymal cell line HA5-Late, below referred to as Mes, which derives from the same primary kidney tissue as Epi, but through immortalization at the late steps of the life span after natural accomplishment of EMT (Castro-Vega et al., 2013b). In addition to expression of key mesenchymal markers and increased migration properties, this cell line is characterized by high levels of HOTAIR comparing to Epi (Figure S1A-S1C). We found that the ectopic expression of HOTAIR from the CMV promoter was at least 36 times higher comparing to its inherent levels in Epi cells, and 3.5 times higher than in Mes cells expressing it endogenously. While comparing expression levels of the full-length and truncated variants, the HOTΔL transcript was at least twice more abundant than HOT or HOTΔP transcripts in Epi cells (Figure 1B). Subcellular fractionation into cytosolic, nucleoplasm and chromatin fractions confirmed that the overexpression, as well as sequence truncations did not change HOTAIR subcellular residence and, in particular, its association with chromatin in Epi cells in comparison to Mes (Figure S1D, S1E). Moreover, subcellular localization and expression levels of HOTAIR protein partners EZH2 and Lsd1 were the same in all cell lines (Figure S1F and S1G). Generated Epi cell lines expressing the full-length HOT and truncated HOTΔP and HOTΔL were further used as a model system to assess a role of PRC2- and Lsd1-interacting domains in HOTAIR function at cellular and molecular levels.

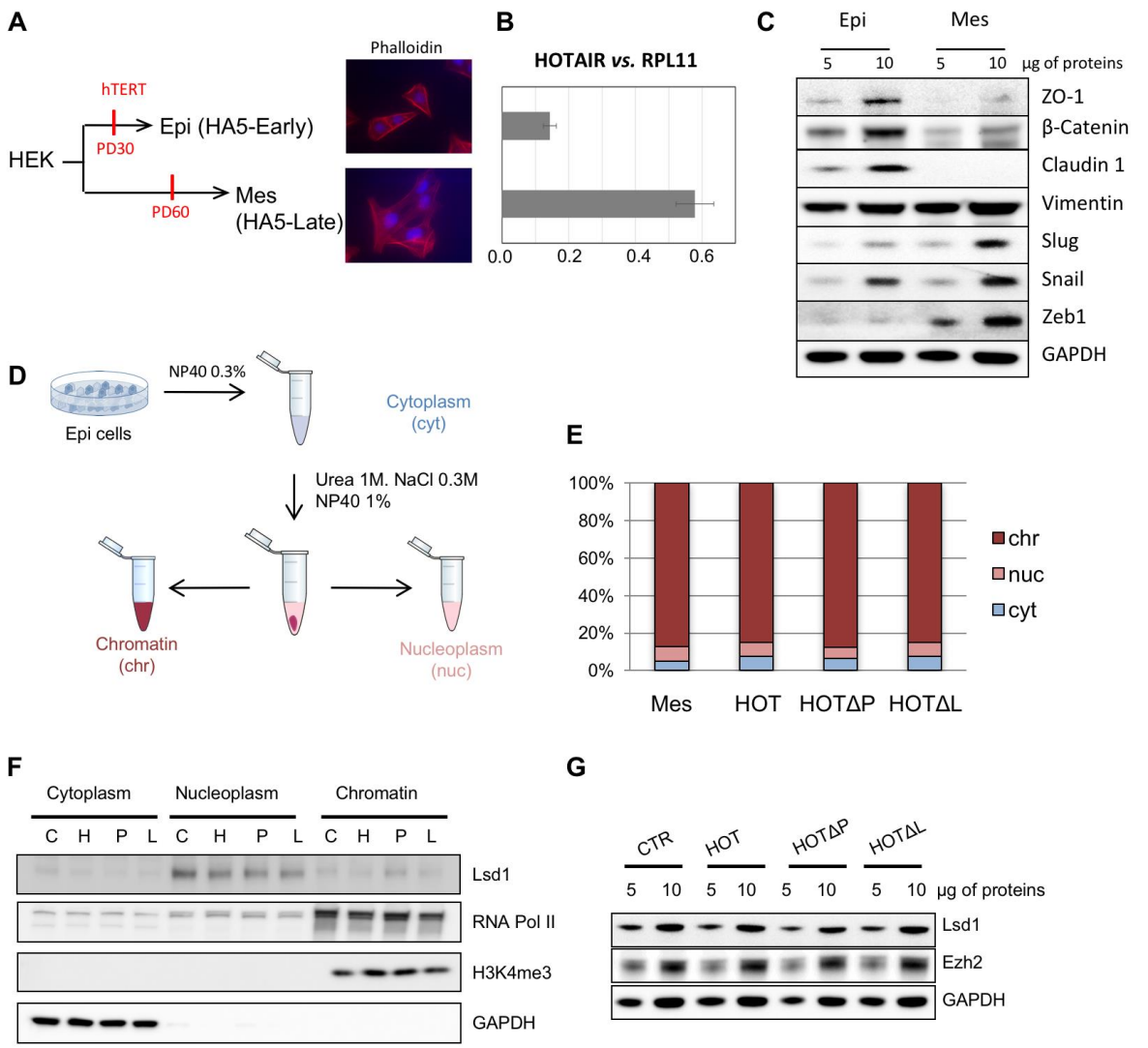


Figure S1. In *vitro* EMT system used to study HOTAIR function:
(A) Epi and Mes cell lines corresponding to HA-Early5 and HA5-Late, respectively (Castro-Vega et al., 2013), stained for F-actin fibers by Phalloidin-TRITC (x40); **(B)** Quantification of HOTAIR expression in Epi and Mes cells by random-primed RT-qPCR; **(C)** Expression levels of EMT markers in whole protein extracts of Epi and Mes cells assessed by Western blot; **(D)** Protocol of subcellular fractionation into cytoplasm, nucleoplasm and chromatin fractions; **(E)** Distribution of full-length and truncated variants of HOTAIR between cytoplasm, nucleoplasm and chromatin fractions in Mes and Epi cells assessed by RT-qPCR relative to GAPDH mature mRNA; **(F)** Subcellular distribution and levels of Lsd1, RNA Pol II, H3K4me3 and GAPDH in cytoplasm, nucleoplasm and chromatin fractions and **(G)** levels of Lsd1, EZH2 and GAPDH in whole protein extracts assessed by Western blot in Epi cells expressing CTR (C), HOT (H), HOTΔP (P) and HOTΔL (L).

Lsd1-interacting domain is essential for HOTAIR function in promoting cell migration

One of the most robust phenotypes associated with HOTAIR expression is the increase in ability of epithelial cells to migrate (Ding et al., 2014), (Dong and Hui, 2016). Therefore, we assessed whether HOTAIR affects migration of Epi cells using the wound healing assay (WHA). As expected, HOTAIR promoted migration of epithelial cells, though the wound healing was much slower than in fully reprogramed mesenchymal Mes cells (Figure 1C, Figure S2A). Surprisingly, deletion of the PRC2-interacting domain did not have an effect as the Epi-HOTΔP cell line migrated as fast as Epi-HOT. On the contrary, HOTAIR missing the Lsd1-interacting domain healed the wound as slowly as the control epithelial cells Epi-CTR (Figure 1C, Figure S2A). We also assessed the proliferation by measuring population doubling (PD) rates of each cell line and did not find significant differences that could explain observed gain or loss of the wound healing efficiency (Table S1).

Table S1. Population doubling (PD) time. Expressed as a mean with a standard deviation (SD) calculated for exponentially growing cells according to a formula $PD=(t_F-t_I)*\ln 2/\ln(N_F/N_I)$; t stands for time; F and I stand for Final and Initial, and N stands for the number of cells.

	Epi	Mes	Epi-CTR	Epi-HOT	Epi-HOTΔP	Epi-HOTΔL
Mean (hours)	16.13	17.62	16.90	16.68	16.44	16.40
SD	2.02	2.47	2.32	1.71	2.57	3.30

Cellular migration is a highly complex phenomenon associated with changes in cell-cell junctions, cytoskeletal organization and apico-basal polarity of epithelial cells (Lamouille et al., 2014a). We assessed the general morphology of cells expressing HOTAIR by the phalloidin staining of F-actin fibers, but no change in cell shape or in formation of cell sheets was detectable (Figure S2B). During EMT, the acquisition of migratory properties is known to result from the decrease in the formation of tight junctions involved in cell-to-cell contacts (Tornavaca et al., 2015). Therefore, we measured protein levels of the tight junction protein ZO-1/TJP1 by Western blot as well as its subcellular localization by Immunofluorescence (IF). Strikingly, the ZO-1 abundance in the whole protein extracts decreased in HOT and HOTΔP expressing Epi cells showing higher migration, but not in low-migrating Epi-HOTΔL and Epi-

CTR cells (Figure 1D and 1E). Concordantly, we observed a more diffused localization of ZO-1 by IF, especially, at cell-cell junctions in HOT and HOTΔP comparing to HOTΔL and CTR (Figure 1F and Figure S2C). The expression of other epithelial markers, β-Catenin and Claudin-1, and mesenchymal markers, Slug, Snail, Zeb1 and Vimentin, was unchanged at protein levels as assessed by Western blot (Figure S3).

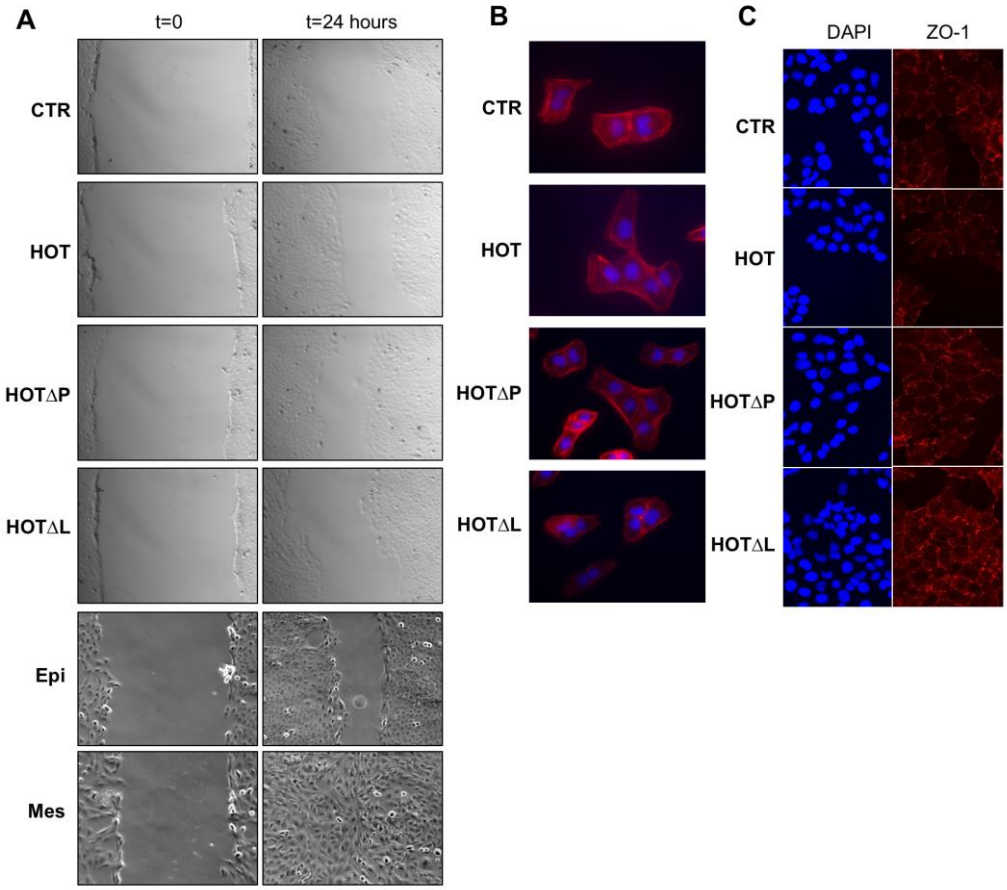


Figure S2. EMT characteristics of Mes and Epi cells used in this study: (A) Assessment of migration capacities by WHA: **r**epresentative images at zero and 24 hours post-scratch; **(B)** Phalloidin-TRITC staining of F-actin fibers (x40); **(C)** ZO-1 subcellular localization assessed by Immunostaining and fluorescence microscopy: representative images of ZO-1 (red) and DNA/nucleus (DAPI, bleu) generated from three ApoTome stacks in Epi cells expressing none (CTR), full-length (HOT) and truncated variants (HOTΔP and HOTΔL) of HOTAIR.

Together, these findings suggested that acquisition of migratory properties by epithelial cells is promoted by high levels of HOTAIR and relies on its interaction with Lsd1 rather than with PRC2. The gain in migration is associated amongst other factors with the weakening of cell-cell junctions. The expression of EMT drivers, the

key transcription factors known to induce cell reprogramming towards mesenchymal identity, remained unchanged in Epi cells expressing HOTAIR.

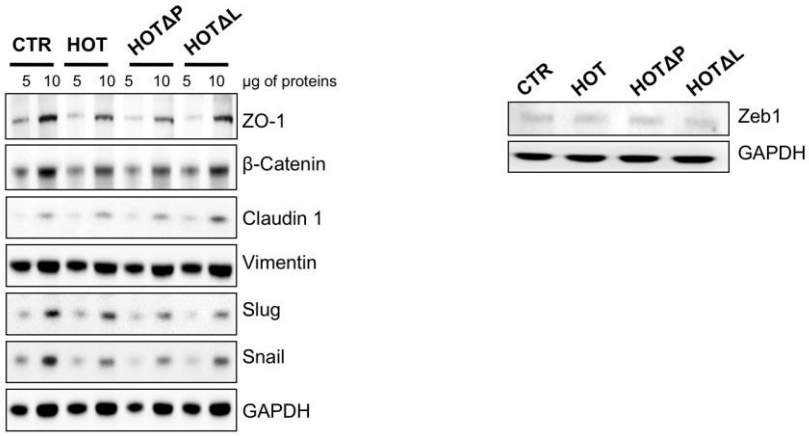


Figure S3. Expression levels of EMT markers in whole protein extracts of Epi-CTR, HOT, HOTΔP or HOTΔL cell lines assessed by Western blot.

HOTAIR expression in epithelial cells induces global transcriptomic changes, majorly dependent on both PRC2- and Lsd1-interacting domains

To get insights into the molecular mechanisms driving the changes of migratory properties and cell identity upon HOTAIR expression, we performed RNA-sequencing (RNA-seq) and differential expression analysis of CTR, HOT, HOTΔP or HOTΔL expressing Epi cells. First, the transcriptome of Epi-CTR was compared to Epi cells expressing each of HOTAIR variants to define HOTAIR induced transcriptomic changes associated with each condition. Then, differentially expressed (DE) genes of each set were intersected to query the ones common to all, at least to two or exclusive to one specific condition.

First and as expected, HOTAIR expression in Epi cells induced global changes in expression of protein-coding genes (PCGs) with a prevalence of a repressive effect (Figure 2). A total of 743 PCGs were retained as significantly dysregulated in Epi-HOT with a fold-change (FC) above 2 and the adjusted p-value below 0.05 (Figure 2A and 2B, Table S2). Deletion of either PRC2 or Lsd1-interacting domains within HOTAIR resulted in more moderate transcriptomic perturbations with 191 and 347 DE-PCGs, respectively, again with a prevalence of down-regulation. Further intersection of up- and down-regulated genes associated with each variant identified 495 DE-PCGs genes strictly requiring the presence of both domains (Figure 2B). These genes were grouped into HPL-neg and HPL-pos sets for down-

(n=379) and up- (n=116) regulated genes, respectively, representing putative HOTAIR/PRC2/Lsd1-dependent targets (Figure 2B, Table S3). Some genes of the HPL set have already been reported among EMT markers (Gröger et al., 2012b) and some identified as repressed by HOTAIR and PRC2 mediated histone H3 Lysine 27 methylation in previous studies (Gupta et al., 2010c) (Table S3, Gröger and Gupta sets). Among them were genes involved in proteolysis of extracellular matrix, *SERPIN2* and *MMP3*, and the protocadherin gene family member *PCDH18* (Gupta et al., 2010c), (Xu et al., 2013), (Qiu et al., 2014). Gene ontology (GO) analysis revealed enrichment of the HPL-set for genes involved in several KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways tightly linked to EMT, cancer and metastasis (Figure 2D). Importantly, the most significantly depleted pathways were enriched in genes of extracellular matrix (ECM) receptor interactions, focal adhesion and Hedgehog signaling, whereas the up-regulated genes represented Jak-STAT and bladder cancer pathways (Figure 2D). Notably, the DE-gene sets associated with HOTAIR expression were devoid of the key transcription factors inducing EMT and described as EMT drivers.

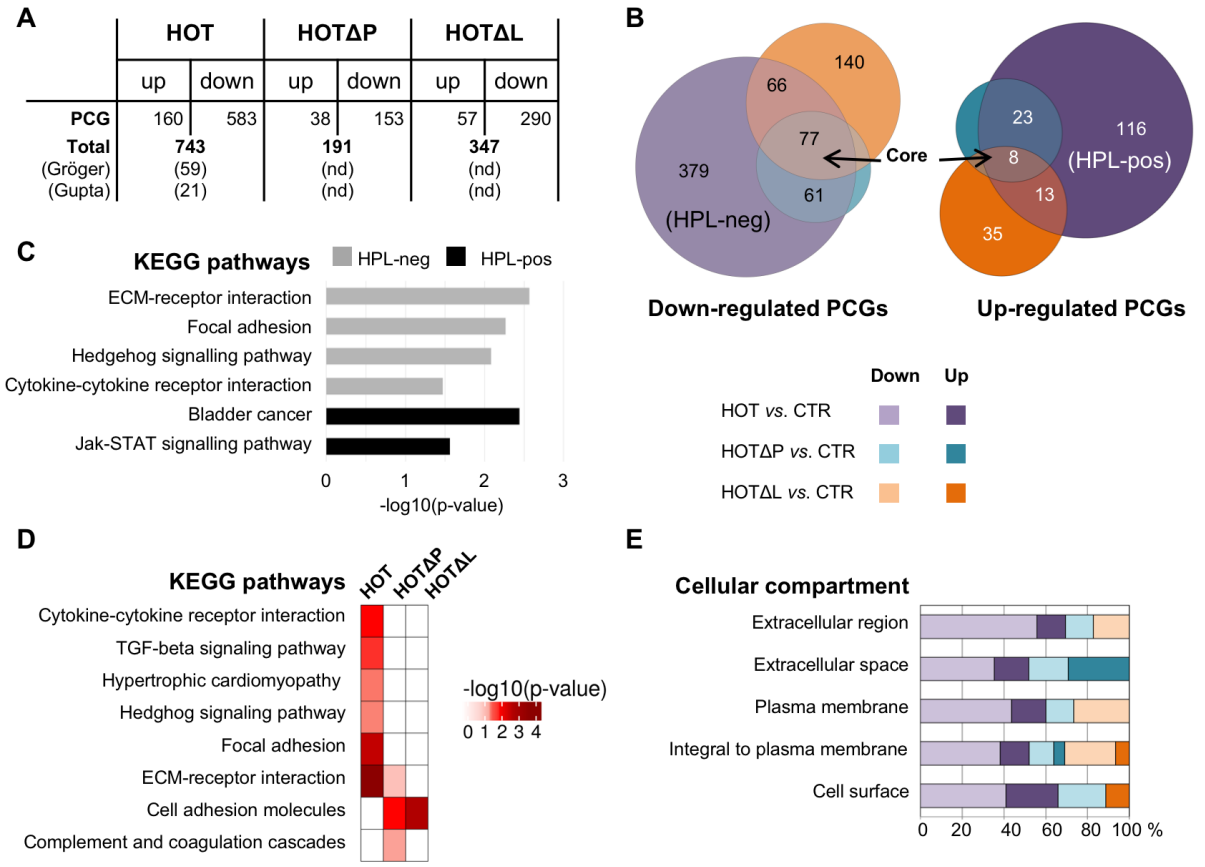


Figure 2. HOTAIR expression in Epi cells induces drastic changes in expression of PCGs, majorly dependent on the presence of both, PRC2- and Lsd1-interacting domains:
(A) Number of up- and down regulated genes defined as differentially expressed (DE) in HOT, HOTΔP and HOTΔL expressing Epi cells comparing to Epi-CTR by DESeq (FC above 2 and adjusted p-value below 0.05), including those associated with EMT and already identified as HOTAIR/PRC2 targets; nd stands for non-determined; **(B)** Venn diagram of intersection of down- and up-regulated PCGs in Epi-HOT, HOTΔP and HOTΔL cells comparing to Epi-CTR; **(C)** KEGG pathways identified by DAVID as significantly enriched (adj. p-value < 0.05) for HPL-set of PCGs; **(D)** KEGG pathways identified by DAVID as significantly enriched (p-value < 0.05) for DE-PCGs of HOT, HOTΔP and HOTΔL expressing Epi cells comparing to Epi-CTR; **(E)** Cellular compartments of differentially expressed protein counterparts shared by up- and down-regulated PCGs in Epi-HOT, HOTΔP and HOTΔL cells comparing to Epi-CTR.

By contrast, expression of 85 genes was affected whatever the HOTAIR truncation (Figure 2B). These genes were grouped into a Core-set representing potential HOTAIR targets, most likely regulated independently of PRC2 and Lsd1 but through alternative mechanisms (Table S3). The majority of the Core-set genes was down-regulated (n=77/85) and was involved in protein maturation and processing pathways, proliferation and extracellular matrix organization processes. Since this particular group of genes was not the focus of this study it was excluded from further analysis.
Together, our protein and global transcriptomic results strongly suggest that HOTAIR has no function in the EMT reprogramming in immortalized primary epithelial cells, instead, they further reinforce its role as modulator of the epithelial-mesenchymal plasticity towards acquisition of some mesenchymal traits particularly affecting signal transduction and migration pathways.

Distinct roles of PRC2- and Lsd1-interacting domains in HOTAIR-mediated regulation of gene expression

To further discriminate HOTAIR targets dependent on its interaction with either PRC2 or Lsd1, we analyzed more in detail DE-PCGs in Epi cells expressing truncated variants of HOTAIR. As aforementioned, deletion of either PRC2- or Lsd1-interacting domain within HOTAIR resulted in decreased number of DE genes (R^2 of 0.986 and 0.982, respectively) comparing to changes induced by the expression of the full-length transcript (R^2 of 0.972) (Figure 2A, Figure S4).

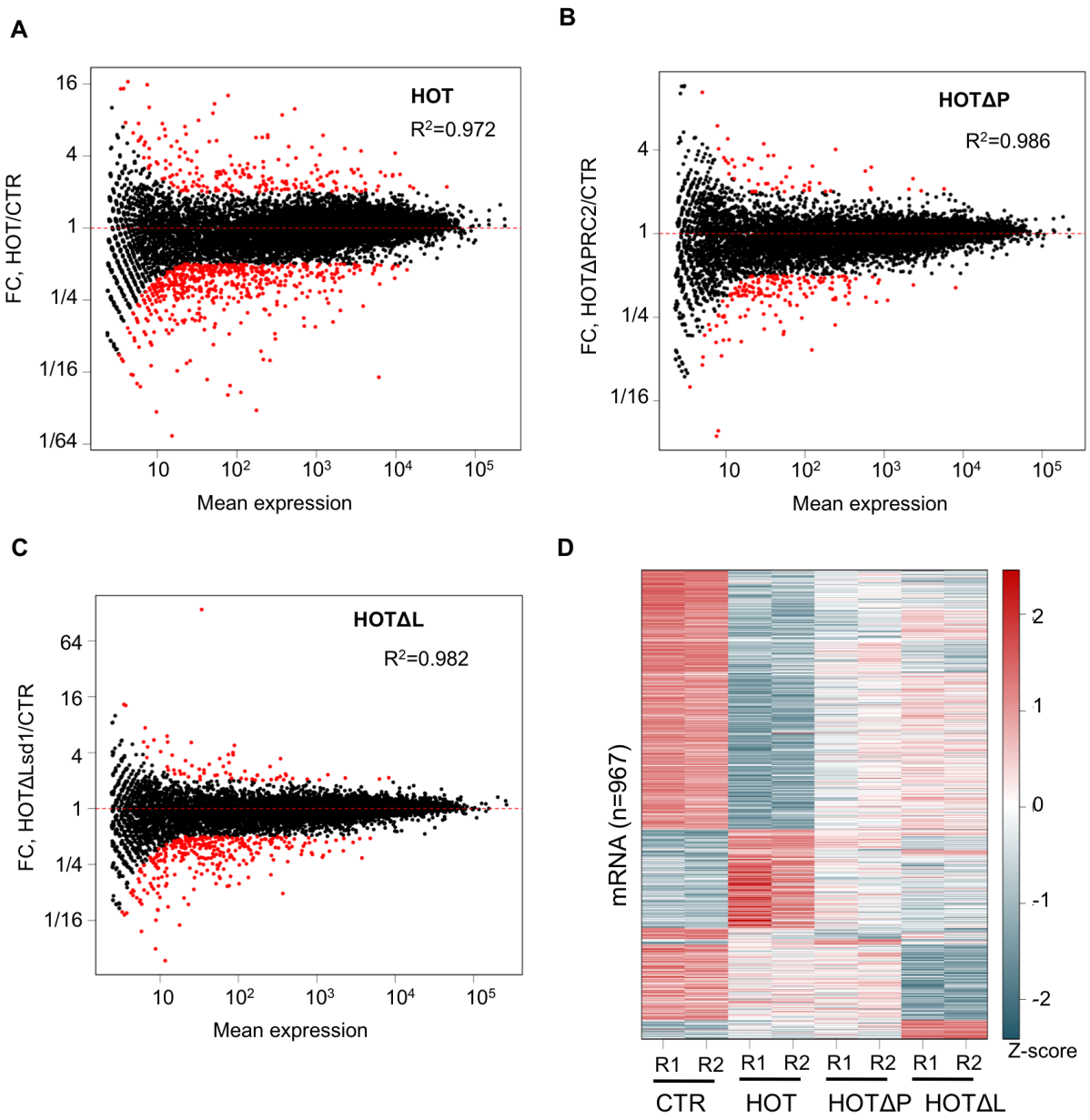


Figure S4. Expression of full-length and Lsd1-interacting domain deleted variant of HOTAIR induces the most drastic changes in PCG transcriptome:
(A-C) MA-plot of protein-coding genes expression in Epi-HOT, HOTΔP and HOTΔL cells comparing to Epi-CTR; Black dots represent all counted PCG, red dots only those with the fold-change (FC) above 2 and adjacent p-value below 0.05; **(D)** Heatmap of DE-PCGs defined by RNA-seq and DESeq analysis.

Paradoxically, even if there were more drastic transcriptome perturbations in Epi-HOTΔL (347 DE-PCGs), they were not sufficient for the Lsd1-domain truncated HOTAIR variant to promote migration, whereas the HOTΔP variant induced less changes (191 DE-PCGs) but still showed increased migration as much as the full-length HOTAIR (Figure 1C). Moreover, while querying GO terms for DE-PCGs in Epi

cells expressing truncated variants of HOTAIR, we retrieved cell adhesion molecules for both Epi-HOTΔL and HOTΔP, but for the rest the transcriptomic landscape of these two cell lines was quite distinct (Figure 2D). In addition to KEGG, we searched for cellular compartments of differentially expressed protein counterparts. Strikingly, DE-PCGs of HOT and HOTΔP were particularly enriched in genes localized to cell surface, extracellular region and matrix, as though DE-PCG of HOTΔL were much more represented by plasma membrane and intracellular locations (Figure 2E). Even if GO terms comparisons are difficult to interpret because of the low number of misregulated genes, fast-migrating Epi-HOT and Epi-HOTΔP cells were characterized by overexpression of genes featuring extracellular space and involved in cell adhesion, whereas the low-migrating Epi-CTR and Epi-HOTΔL cells showed more alterations in expression of genes with intracellular functions as cell signaling.

The LSD1 interacting domain of HOTAIR shapes the fast migrating transcriptome landscape

Epithelial plasticity is defined by a balance in expression of epithelial and mesenchymal genes. Juxtaposition of migration and transcriptome changes in Epi cells overexpressing full-length or truncated variants of HOTAIR strongly suggested that HOTΔL cells maintain their epithelial balance to a larger extent than HOT or HOTΔP cells, both able to interact with Lsd1. This observation nourished a hypothesis that Lsd1/HOTAIR crosstalk may affect epithelial-mesenchymal balance. To define other genes involved in this regulation, we performed a differential expression analysis of Epi-CTR and Epi-HOTΔL transcriptomes against Epi-HOT and Epi-HOTΔP. Knowing that Lsd1 is a component of multiple complexes with repressor or activator activities, we assigned all PCGs that are significantly up-regulated in Epi-CTR and Epi-HOTΔL datasets relatively to Epi-HOT and Epi-HOTΔP (n=148) (DESeq, FC > 1.5 and p-value < 0.05) but absent in the Core-set, as a *Low Migration Signature* (LMS, n=131) (Figure 3A, Tables S3 and S4). Similarly, all up-regulated PCGs in Epi-HOT and Epi-HOTΔP (n=77), but absent in the Core-set were grouped into a *High Migration Signature* (HMS, n=75) (Figure 3A, Tables S3 and S4). Remarkably, the LMS-set was composed of genes mostly involved in pathways linked to cardiac development, steroid biosynthesis and cytokine-cytokine receptor interactions, whereas the HMS-set was clearly enriched in cancer and metastasis

related pathways including the ECM receptor interaction and focal adhesion (Figure 3B). Notably, transcriptomic changes induced by HOTAIR did not result in a complete switch of EMT program but rather in modulation (attenuation or increase) of gene expression (Figure 3C). Among LMS genes highly expressed in Epi-CTR and Epi-HOTΔL were the tumor suppressor *MTUS1* (Di Benedetto et al., 2006), the nuclear receptor *PRICKLE1/RILP* implicated in the nuclear trafficking of REST/NRSF and REST4 transcription repressors (Shimojo and Hersh, 2006). HMS genes highly expressed in Epi-HOT and Epi-HOTΔP included the cell cycle regulator *CCND1*, the *DNER* activator of the NOTCH pathway, but also the *CD44* EMT marker (Figure 3D). In conclusion, disruption of HOTAIR interaction with PRC2 or Lsd1 does not abolish completely its function as a regulator of gene expression; however, HOTAIR association with Lsd1 is essential for the modulation of the transcriptomic pattern of epithelial cells in favor of mesenchymal identity.

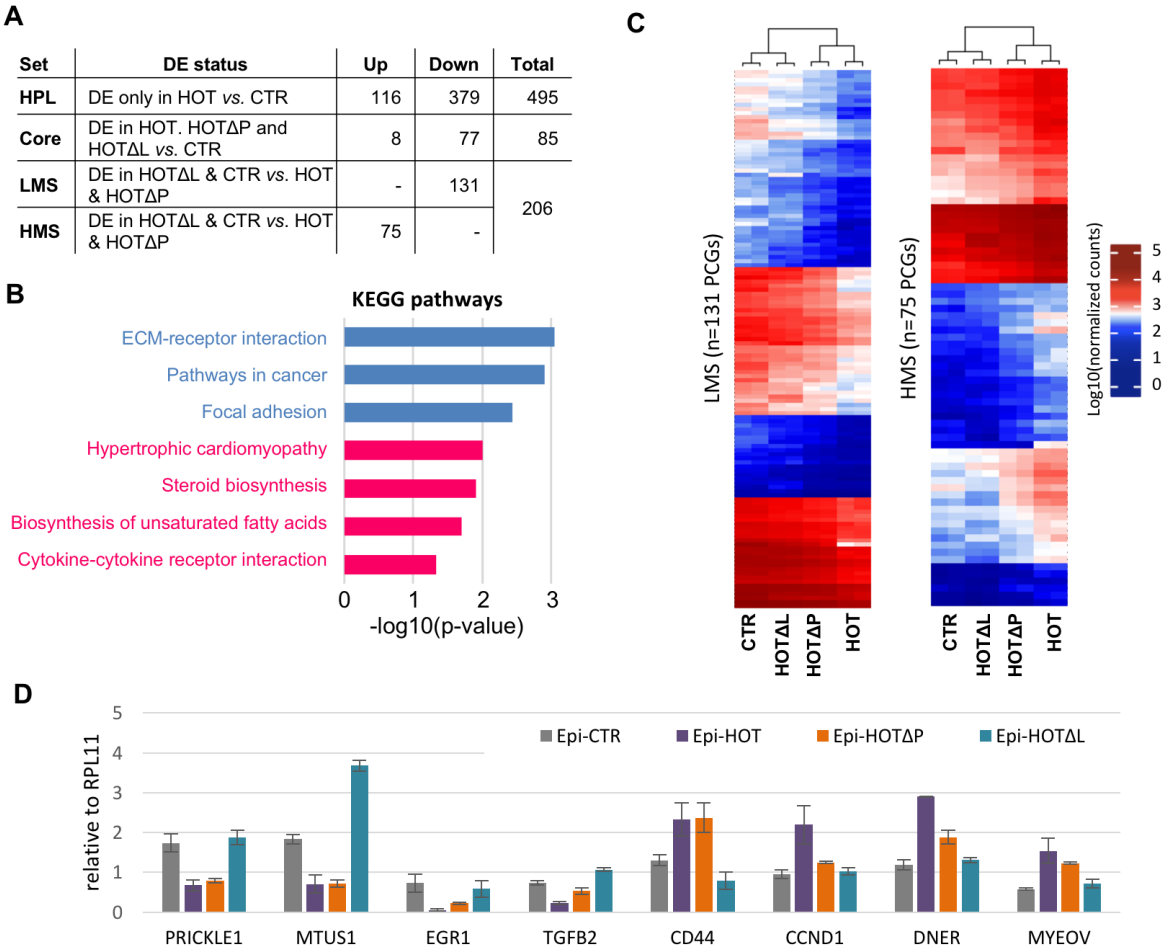


Figure 3. Transcriptome signature of low and fast migrating epithelial cells:

(A) PCGs assignment to different gene sets according to DE features; **(B)** KEGG pathways enriched by PCGs from LMS (blue) and HMS (pink) sets (DAVID, p-value below 0.05); **(C)** Unsupervised hierarchical clustering heatmap of LMS and HMS gene sets; **(D)** Random-primed RT-qPCR quantification of gene expression levels relative to RPL11 in Epi-CTR and Epi cells expressing HOT, HOTΔP or HOTΔL variants of HOTAIR.

HOTAIR and its Lsd1-interacting domain are essential for Lsd1 chromatin redistribution

With the support of previous studies, we hypothesized that the epithelial-mesenchymal plasticity is controlled by the function of Lsd1 in gene expression regulation, which may be modulated in cells expressing HOTAIR. Since Lsd1 protein levels in total and nuclear cell extracts did not show any changes in response to HOTAIR overexpression in epithelial cells (Figure S1F, S1G), we aimed to test whether HOTAIR would affect Lsd1 chromatin occupancy and distribution in two phenotypically distinct groups: Epi-CTR and Epi-HOTΔL cells would represent a biological context, in which Lsd1 exhibits its function independently of HOTAIR maintaining epithelial identity and low migration, whereas Epi-HOT and Epi-HOTΔP would designate a context with both free and HOTAIR associated Lsd1 function. We performed a Chromatin ImmunoPrecipitation sequencing (ChIP-seq) of Lsd1 in Epi cells expressing full-length or truncated variants of HOTAIR in comparison to the control Epi-CTR condition to define Lsd1 chromatin occupancy. The uniquely mapped reads of two replicates per condition were subjected independently to a peak calling procedure of SICER, an algorithm specifically designed for identification of dispersed IP-DNA enriched islands relative to a corresponding Input-DNA signal (Zang et al., 2009). The blacklisted by ENCODE genomic regions were excluded from further consideration (ENCODE Project Consortium, 2012b) and only peaks showing at least 1 nucleotide overlap in two replicates were merged and retained for further analysis (Figure 4A, Table S3). The number of detected Lsd1 peaks was strikingly heterogeneous between conditions, in particularly, CTR and HOTΔL datasets showed as much as 20 times more peaks than HOT and HOTΔP regardless the identical ChIP-seq metrics (Figure S5A, S5B). This result correlated with the differences in migration capacities of the cell lines; CTR and HOTΔL being of lower and HOT and HOTΔP of higher migration capacities. We interrogated genomic locations occupied by Lsd1 in both replicates and revealed that

only few peaks were located to promoter regions, transcriptional start sites (TSS) as though the majority was detected within noncoding 5'UTR, intergenic and intronic regions (Figure 4A, Figure S5B). Herein, the Epi-HOTΔP cell line was particularly depleted for Lsd1 in promoter and 5'UTR regions. However, no specific, discriminating feature was found when comparing Lsd1 peak locations between the two phenotypically distinct groups, CTR and HOTΔL *versus* HOT and HOTΔP. Distribution of Lsd1 peaks as a distance from genes TSS did not show any significant difference between CTR and three other conditions (p-value > 0.3, Wilcoxon test) (Figure S5C).

Finally, we determined the number of covered bases per peak for every condition and revealed that all three cell lines expressing HOTAIR showed broader Lsd1 peaks than Epi-CTR (p-value < 10^{-11}, Wilcoxon test) (Figure S5D), nevertheless according to the density plot, the two low migrating cell lines, Epi-CTR and Epi-HOTΔL, presented more sharp peaks with the mode values of 3.6 and 6 kb, respectively, than Epi-HOT and Epi-HOTΔP with the mode values of 8.8 and 13.4 kb, respectively (Figure 4B).

Although Lsd1 is not a strictly a promoter associated factor and can be found in distal, enhancer regions, gene bodies, but also cover large chromosomal regions, we decided to explore the Lsd1 landscape on a gene-based approach. For this, we assigned peaks, unique to each cell line and common for two replicates, within the 5kb window around the TSS and within the TSS-TTS window to a corresponding gene and searched for specific Lsd1 patterns associated with low and high migration phenotypes but also with transcriptomic high- and low-migration signatures (HMS and LMS) retrieved from the RNA-seq differential expression analysis (Figure 4C and 4D, Figure S5). Firstly, intersection of the Lsd1-associated genes between conditions revealed high cell-line specificity of Lsd1 loci with only one gene (RNU2-38P, snRNA gene) common to Epi-HOT and Epi-HOTΔP and 555 genes shared by Epi-CTR and Epi-HOTΔL (Figure 4C). We considered the common Lsd1 associated genes of Epi-CTR and Epi-HOTΔL datasets as genomic locations independent of Lsd1/HOTAIR interactions and specific to the low migration phenotype. Among them, 312 represented PCGs and 243 were noncoding genes. Gene enrichment analysis revealed Jak-STAT signaling pathway as the significantly enriched KEGG pathway and several biological processes tightly linked to EMT, such as positive regulation of cell motion and TGF-beta signaling, cytokine-mediated signaling

pathways among PCGs presenting Lsd1 peaks 5kb upstream of their TSS and within the gene body (Figure S5E).
Secondly, intersection of Lsd1-associated protein-coding genes (n=353) with LMS (n=131) and HMS (n=75) transcriptomic sets revealed the presence of Lsd1 for 7 up-regulated genes, including the *PRICKLE1*, and for only 1 down-regulated genes, the antagonist of fibroblast growth factor pathways SPRY2, in Epi-CTR and Epi-HOTΔL cells (Figure 4D).

In sum, HOTAIR expression in epithelial cells dramatically affects Lsd1 genomic localization and, in particular, results in its dislocation from specific genomic locations. As a consequence, this imbalances transcription and promotes expression of mesenchymal genes endowing partial transition of epithelial cells to a more mesenchymal phenotype (Figure 4E).

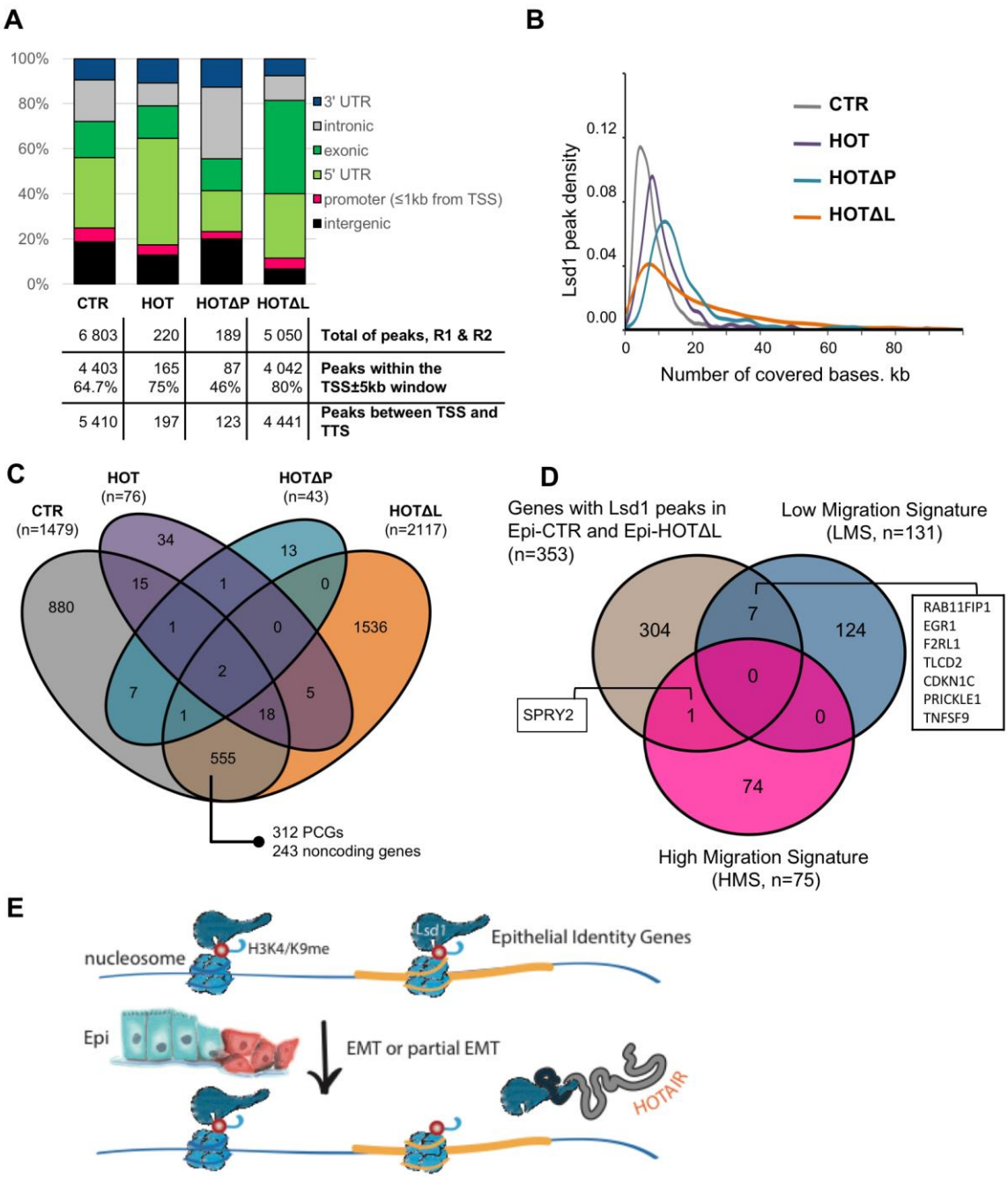


Figure 4. HOTAIR expression promotes Lsd1 dislocation from inherent genomic locations through its 3'-Lsd1-interacting domain:
(A) Lsd1 peaks identified by ChIP-seq of Lsd1 and SICER peak calling protocol and their distribution across distinct genomic features; **(B)** Density plot of the number of bases covered by Lsd1 peaks; **(C)** Venn diagram representing intersections of genes possessing Lsd1 peaks within the 5 kb window upstream their TSS and within the gene body in Epi-CTR and cells expressing HOT, HOTΔP and HOTΔL; **(D)** Venn diagram presenting the intersection of genes with TSS-associated Lsd1 peaks found in low migrating Epi-CTR and Epi-HOTΔL cells with LMS and HMS sets; **(E)** Model illustrating HOTAIR-mediated disruption of Lsd1 function as a guardian of epithelial identity.

A

ChIP_ID	Name	Number of total reads	Number of mapped reads	% of mapped reads	Number of duplicates	% of duplicates	Mean depth of coverage
A684C1	CTR_Input_R1	58 309 931	54 632 087	93.69%	5 898 003	11%	1,54
A684C2	CTR_Input_R2	40 480 349	38 257 487	94.51%	3 354 952	9%	1,10
A684C3	HOT_Input_R1	14 849 906	13 742 340	92.54%	1 593 115	12%	0,39
A684C4	HOT_Input_R2	36 966 981	35 458 236	95.92%	3 359 281	9%	1,04
A684C5	HOTΔP_Input_R1	36 656 422	33 957 172	92.64%	2 932 174	9%	0,96
A684C6	HOTΔP_Input_R2	41 571 699	39 544 465	95.12%	3 074 204	8%	1,18
A684C7	HOTΔL_Input_R1	36 521 005	33 135 144	90.73%	3 934 516	12%	0,88
A684C8	HOTΔL_Input_R2	35 517 513	33 006 985	92.93%	3 044 680	9%	0,94
A685C9	CTR_IP-Lsd1_R1	50 076 964	43 581 806	87.03%	9 995 393	23%	1,05
A685C10	CTR_IP-Lsd1_R2	45 743 524	37 829 991	82.70%	10 759 175	28%	0,80
A685C11	HOT_IP-Lsd1_R1	32 091 186	27 305 035	85.09%	5 028 321	18%	0,69
A685C12	HOT_IP-Lsd1_R2	41 916 400	37 622 129	89.76%	6 436 216	17%	0,98
A685C13	HOTΔP_IP-Lsd1_R1	54 162 926	49 688 153	91.74%	7 601 412	15%	1,34
A685C14	HOTΔP_IP-Lsd1_R2	39 678 630	36 393 179	91.72%	6 069 608	17%	0,96
A685C15	HOTΔL_IP-Lsd1_R1	31 419 947	23 283 393	74.10%	6 756 120	29%	0,47
A685C16	HOTΔL_IP-Lsd1_R2	32 129 004	26 531 966	82.58%	5 426 179	20%	0,64

B

Lsd1 peaks	Epi-CTR		Epi-HOT		Epi-HOTΔP		Epi-HOTΔL	
	R1	R2	R1	R2	R1	R2	R1	R2
Total	20 592	14 938	6 557	1 778	1 116	8 079	16 768	15 758
Common R1 & R2	6 803		220		189		5 050	
Common R1 & R2, TSS+/-5kb; (genes)	4 403 (1 165)		165 (57)		87 (30)		4 042 (1 549)	
Common R1 & R2, TSS–TTS; (genes)	5 410 (988)		197 (57)		123 (27)		4 441 (1605)	

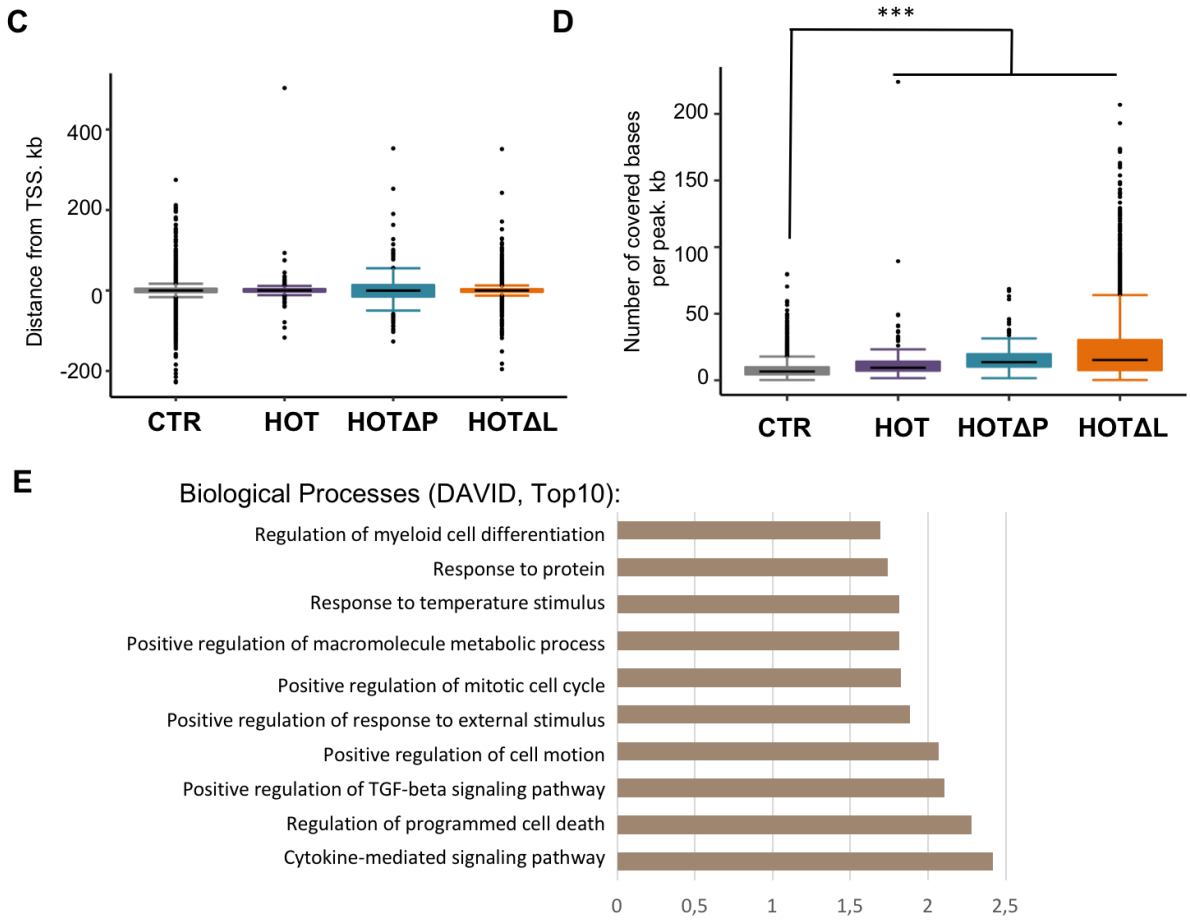


Figure S5. Lsd1 peak features:

(A) ChIP-seq metrics of Input- and IP-DNA sequencing; **(B)** SICER identified Lsd1 peaks in Epi-CTR (n=6803), HOT (n=220), HOTΔP (n=189) and HOTΔL (n=5050) cells, per replicate and common between two replicates; **(C)** Box plot of Lsd1 mead peak distances from genes TSS in Epi-CTR (n=6803), HOT (n=220), HOTΔP (n=189) and HOTΔL (n=5050) cells; **(D)** Box plot of number of bases covered by Lsd1 peaks in Epi-CTR (n=6803), HOT (n=220), HOTΔP (n=189) and HOTΔL (n=5050) cells: *** p-value < 10^{-11}, Wilcoxon test; **(E)** TOP10 hits of biological processes identifies by DAVID as significantly enriched within the set of PCGs presenting Lsd1 peaks within the 5kb window upstream their TSS and within the gene body.

DISCUSSION

The EMT program is proposed as a route for the generation of normal and neoplastic epithelial cells. It enables acquisition of mesenchymal traits promoting migration and invasion, thus, underlying high metastatic potential of tumor tissues. HOTAIR and Lsd1 have been independently studied in a variety of cell-based and clinical settings as factors associated with EMT and cancer metastasis. And if the regulatory function of HOTAIR is unambiguously linked to acquisition of mesenchymal traits as migration and invasion capacities, the role of the Lsd1 histone demethylase is rather context-specific and resumes in positive or negative control of a variety of cell identity programs. Being ubiquitously expressed in both epithelial and mesenchymal cells, it can induce epigenetic changes either locally (enhancers, promoters, gene bodies) or broadly (large chromatin domains, LOCKs) to influence the transcriptional program both way, through repression or activation (Shi et al., 2004b), (Whyte et al., 2012), (McDonald et al., 2011b), (Li et al., 2016b), (Wang et al., 2007a). Lsd1 presence at regions with increased gene expression suggests its positive role through the control of H3K9 methylation status of genes or in tethering of transcription factors promoting transcription initiation (Metzger et al., 2005a), (Yang et al., 2019), (Zhang et al., 2018a), (Zhang et al., 2018a). Another possibility is that Lsd1 regulates co-transcriptional splicing through H3K9 demethylation as it has been shown for CD44 and FGFR2 transcripts (Saint-André et al., 2011), (Gonzalez et al., 2015). Remarkably, in the latter case chromatin modifications have been triggered by the FGFR paired antisense lncRNA (asFGFR2) interacting in *cis* with PRC2 and KDM2a complexes (Gonzalez et al., 2015). Otherwise, we cannot exclude an indirect effect that Lsd1 may exhibit through a control of upstream factors

resulting in up-regulation of LMS genes in Epi-CTR and Epi-HOTΔL cell lines. Further experiments are required to discriminate between these hypotheses.
Lsd1 has also been reported to act independently of its demethylase function at chromatin level and elsewhere (Sehrawat et al., 2018), (Lan et al., 2019), (Carnesecchi et al., 2017). All these mechanistic modalities may be affected by lncRNAs, as described for HOTAIR in breast cancer cells.

The present work identifies HOTAIR as an effector of Lsd1 function as a guardian of epithelial identity. We demonstrated that the 3'-extremity of HOTAIR, which interacts with Lsd1, was essential to promote epithelial cell migration whereas the PRC2-interacting domain was dispensable for this function. Paradoxically and in the light of our results, PRC2 and Lsd1-interacting domains contribute together, but also separately to mechanistically distinct HOTAIR functions. In particular, deletion of the Lsd1-interacting domain still allows cells expressing HOTAIR to maintain a gene expression balance in favor of the epithelial cell identity. This is most likely due to Lsd1 operating independently of HOTAIR. In support of this hypothesis, our Lsd1 chromatin profiling experiments revealed considerable changes in Lsd1 genomic distribution induced by HOTAIR variants with the intact 3'-end sequence enabling its association with Lsd1. The striking correlation between the loss of epithelial traits and changes in Lsd1 landscape strongly supports a pivotal role for Lsd1 as a factor preventing cells from sensing or undergoing an EMT. In the context of effective HOTAIR/Lsd1 association, several molecular scenarios could be considered: (i) HOTAIR may modulate Lsd1 catalytic activity or capacity to interact with its protein partners such as transcription factors or chromatin modifying enzymes (ribo-repressor or activator functions); (ii) HOTAIR may promote the assembly of another specific Lsd1 complex and its tethering to peculiar genomic locations for local chromatin modifications (guide and scaffold functions). Although further studies are required to identify epigenetic changes induced by the Lsd1/HOTAIR complex, to determine other Lsd1/HOTAIR partners and to enlarge this observation to other biological systems, our report revealed an unexpected role of HOTAIR as a molecular toggle switch for Lsd1 function, which may contribute to EMT at the very early steps of transformation of a normal epithelial cell to a neoplastic one.

Intriguingly, several alternative splicing and TSS isoforms of HOTAIR are annotated in the human genome, including those lacking the 3'-terminal sequence interacting

with Lsd1 and variants missing the PRC2- or both PRC2- and Lsd1-interacting domains (Mercer et al., 2012). Even if the clinical relevance of these isoforms has not yet been established, in light of our results, one can anticipate that tumors expressing 3'-end truncated variants of HOTAIR would have a lower metastatic potential, and hence, better prognosis. It will be worth assessing the expression of HOTAIR isoforms in tumors of different grades and prognosis to support our findings.

Chapter 5. Functional discovery of novel lncRNAs in the EMT

1. Introduction

The previous work studying HOTAIR showed our system can be useful to study lncRNAs in EMT, especially the ones associated with the mesenchymal phenotype and their effects on the regulation of epithelial plasticity. From there, the focus of my main project was the discovery and functional characterization of novel lncRNAs involved in EMT. Initiated during my master, the differential analysis of Epi and Mes cells through RNA-seq was first done using Total RNA-seq and showed that many novel lncRNAs could be retrieved in our system.

However, the experience in the Morillon lab in yeast showed that transcription-oriented approaches were good tools for the study of the non-coding transcriptome (Wery et al., 2018a, 2018b), and many lncRNAs in mammals are located to the nucleus where they regulate transcription, epigenetic processes or nuclear architecture (Cabili et al., 2015; Sun et al., 2018). Therefore I first focused on separating nascent and chromatin-associated RNAs from processed cytoplasmic RNAs, more representative of protein-coding steady-state levels, using a subcellular-fractionation approach (Gagnon et al., 2014) coupled to the mammalian method for Native-Elongating-Trancript (NET) Sequencing from the Churchman laboratory (Mayer et al., 2015). Initially designed to study the process of transcription itself, NET-seq relies on the isolation of the 3'-end of nascent transcripts as a snapshot of the location of the polymerase during transcription. Here, I used Total RNA-seq library preparation in order to sequence nuclear chromatin-associated as well as transcriptionally-regulated nascent lncRNAs, although our method cannot separate the two. This method proved very useful for the characterization of the non-coding transcriptome and allowed me to identify many nuclear-enriched lncRNAs which are differentially expressed in Epi and Mes cells.

In order to assess their functional relevance in EMT, I then established a collaboration with the laboratory of Neville Sanjana at the New York Genome Center to apply their CRISPR-based transcriptional activation (CRISPRa) screen method to our system. So far, most CRISPRa screens have been based on stringent phenotypes such as proliferation, cell survival or apoptosis, making up for an easy cell-selection.

However, such phenotypes are not relevant in our system as Epi and Mes cells do not display differences in cell proliferation, cell cycle or apoptosis rates. Instead, I developed two methods of screening which rely on EMT-associated phenotypes. The first method which is the one described in the article below is based on the expression of the Epithelial Cell Adhesion Molecule (EpCAM) surface marker which is typically repressed upon EMT. Epi cells undergoing phenotypic changes can then be isolated through FACS upon the loss of EpCAM expression. The second method which will be discussed next is based on the gain of invasive properties by epithelial cells. For this, I used a Boyden chamber with matrigel to separate the epithelial cells which can pass through the matrigel and membrane, thus isolating cells with increased invasion capacities.

From the differentially-expressed lncRNAs identified in Epi and Mes cells, I designed a library of sgRNAs in the Sanjana laboratory to target the TSS of genes of interest with the CRISPRa machinery and therefore activate transcription. I then cloned the library as a pool of plasmids, each containing one sgRNA targeting either lncRNAs of interest or positive/negative controls, and used it for lentiviral transduction of Epi cells, followed by the screening as mentioned earlier.

The EpCAM CRISPRa-screen allowed me to identify a lncRNA enriched in epithelial cells which have lost EpCAM: MAL-1. This novel lncRNA identified from Mes cells appears to be enriched in the nucleus and associated with the mesenchymal phenotype in other cell lines as well. Interestingly, its *trans* overexpression in Epi cells correlates with a repression of epithelial markers as well as an increase in migration.

The following article is a preliminary draft presenting the results mentioned above. It is not fully complete however, as it is still missing some experiments such as the validation of the single guide RNAs targeting MAL-1 in Epi cells.

In addition, some more experiments were done to study MAL-1, notably siRNA-mediated knock-down as well as the transcriptomic analysis of MAL-1 overexpression and these will be discussed afterward.

For the sake of manuscript clarity, the method section was merged to chapter 3.

2. Publication draft n°2:

CRISPRa screen of chromatin-enriched lncRNAs reveals a new regulator of epithelial identity.

Julien Jarroux[1], Marc Gabriel[1], Rocco Cipolla[1], Meer Mustafa[2], Paola Ortiz-Montero[3], Camille Gautier[1], Arturo Londoño-Valejo[4], Neville Sanjana[5], Marina Pinskaya[1*] & Antonin Morillon[1*]

[1] ncRNA, epigenetic and genome fluidity, CNRS UMR 3244, Sorbonne Université, PSL University, Institut Curie, Centre de Recherche, 26 rue d'Ulm, 75248 Paris, France
[2] Formerly Sanjana lab. Now?
[3] Cellular and Molecular Physiology Group, Faculty of Medicine, Department of Physiological Sciences, National University of Colombia, Bogotá, Colombia.
[4] Telomeres and cancer, CNRS UMR 3244, Sorbonne Université, PSL Université, Institut Curie, Centre de Recherche, 26 rue d'Ulm, 75248 Paris, France
[5] New York Genome Center, New York, NY 10013, USA; Department of Biology, New York University, New York, NY 10003, USA.

Contact:
marina.pinskaya@curie.fr
antonin.morillon@curie.fr

[*] co-corresponding

ABSTRACT

In this study, we focus on identifying functionally relevant long non-coding RNAs (lncNAs) during the epithelial-to-mesenchymal transition (EMT) using the first ever CRISPR-activating (CRISPRa) screen for EMT. We first used *de novo* lncRNA annotation and a subcellular-fractionation approach to RNA-seq in order to characterize the non-coding transcriptome of EMT cells. We showed that chromatin-based RNA-seq allows for a highly sensitive and robust differential analysis of lncRNAs. We then performed a CRISPRa screen to identify several lncRNAs impacting epithelial identity through the loss of the EpCAM surface marker. Interestingly, sgRNAs targeting MAL-1, the most differentially expressed lncRNA in mesenchymal cells from our system, were highly enriched in EpCAM-negative cells. We then showed MAL-1 is a nuclear-enriched transcript associated with the mesenchymal identity which may act in *trans* to induce the repression of epithelial markers such as EpCAM or other junction proteins, and increase the migratory properties of epithelial cells.

INTRODUCTION

The epithelial-to-mesenchymal transition (EMT) is a dynamic biological process which controls the cellular plasticity of epithelial cells in order to repress their specific features and gain a mesenchymal phenotype. Already extensively studied during development, it has also been associated with fibrosis, cancer progression and metastasis (Kalluri and Weinberg, 2009). One of the main markers lost by epithelial cells through EMT are the cell junction and adhesion proteins which are either degraded or relocalized upon EMT (Lamouille et al., 2014b). One of them is the Epithelial Cellular Adhesion Molecule (EpCAM) which is involved in intercellular adhesion and the regulation of proliferation, stemness, as well as invasion and migration (Keller et al., 2019). As a very specific marker of epithelial cells, it is commonly used to separate epithelial and mesenchymal cells through flow-cytometry, and even to isolate circulating tumor cells from blood (Hyun et al., 2016; Latil et al., 2017; Ruscetti et al., 2015). During EMT, the epigenetic landscape and transcriptome have also been shown to be deeply reprogrammed by "EMT drivers", transcription factors which play a pivotal role in the induction of EMT such as the SNAI, ZEB or TWIST families. For example, transcription factors ZEB1 and ZEB2 can repress and activate many target genes to induce EMT, either directly or through the recruitment of epigenetic modifiers such as Lsd1 (Skrypek et al., 2017). Besides protein coding genes (PCGs), EMT has also been shown to be regulated by microRNAs such as the miR-200 family and others (Expósito-Villén et al., 2018; Zaravinos, 2015), and more recently by long non-coding RNAs (lncRNAs) (Gugnoni and Ciarrocchi, 2019).

By definition, lncRNAs are transcripts longer than 200 nucleotides (nt) of low or no coding potential (Quinn and Chang, 2016). Although there are some exceptions, they are typically transcribed by RNA Polymerase II from genomic loci exhibiting similar chromatin features to PCG such as H3K4me3 and H3K27ac around the TSS and H3K36me3 along the gene body (Derrien et al., 2012b; Guttman et al., 2009b; Hnisz et al., 2013). Unlike messenger RNAs which are mostly cytoplasmic, lncRNAs have been found in a variety of subcellular localizations, notably in the nucleus (Cabili et al., 2015b). Nuclear lncRNAs can regulate epigenetic marks or directly transcription, both in *cis* and *trans* (Sun et al., 2018) and several studies specifically characterized nuclear- and chromatin-enriched lncRNAs (Shukla et al., 2018; Werner et al., 2017).

In the context of EMT, the non-coding transcriptome has been shown to be heavily remodeled (Liao et al., 2017), giving rise to many lncRNAs that can either activate (ex: HOTAIR, MALAT1) or hinder (ex: GAS5) the transition. HOTAIR has been shown to promote cancer metastasis through its physical interaction with epigenetic modifiers (Gupta et al., 2010a; Kogo et al., 2011a; Song et al., 2019). By interacting with Ezh2, it acts as a bridge for Snail-mediated repression, and we have recently shown that it can relocate Lsd1 from its inherent genomic loci to repress epithelial identity (Battistelli et al., 2016; Jarroux et al., 2019). Another example would be MALAT1 which has been shown to be upregulated in many types of cancer where it acts as a competing endogenous ceRNA for many miRNAs (miR204, mir205) which target EMT drivers (ZEB1, ZEB2, SNAI2), resulting in their upregulation and the promotion of EMT (Gugnoni and Ciarrocchi, 2019; Zhang et al., 2017). As mentioned, lncRNAs have also been associated with the repression of EMT such as GAS5 which acts as a tumor suppressor ceRNA in osteosarcoma by repressing proliferation, migration and EMT (Ye et al., 2017). However, EMT is a very dynamic process in which lncRNA-based regulations can be very complex with a wide variety of mechanisms of action. Indeed, some lncRNAs were shown to have contradictory actions as is the case of the EMT-activated lncRNA H19 which can promote EMT as well as its counter-process, the mesenchymal-to-epithelial transition, through the regulation of miRNAs, interaction with epigenetic modifiers such as MBD1 and PRC2 or even the tumor suppressor p53 (Matouk et al., 2014; Raveh et al., 2015).

In recent years, siRNA screens have been used to study the role of specific PCG families during EMT (Davis et al., 2014; Pavan et al., 2018), however the new CRISPR-based approaches have yet to be used, especially to study lncRNAs. The versatile nature of these novel CRISPR tools allows for targeted transcriptional activation or repression in a high-throughput manner, which has proven very useful for the identification of functional lncRNAs (Joung et al., 2017a; Liu et al., 2017b; Montalbano et al., 2017).

In this study, we defined a set of transcriptionally regulated lncRNAs through subcellular-fractionation RNA-seq and used CRISPR-based transcriptional activation (CRISPRa) in order to identify lncRNAs which regulate the epithelial identity. Among the functionally relevant lncRNAs which we identified, Mesenchymal identity Associated-LncRNA 1 was the most prominent. Further

characterization showed it is a nuclear-enriched transcript which is associated to mesenchymal identity and can act as a stand-alone lncRNA molecule in *trans* to induce the repression of epithelial markers and increase cell migration.

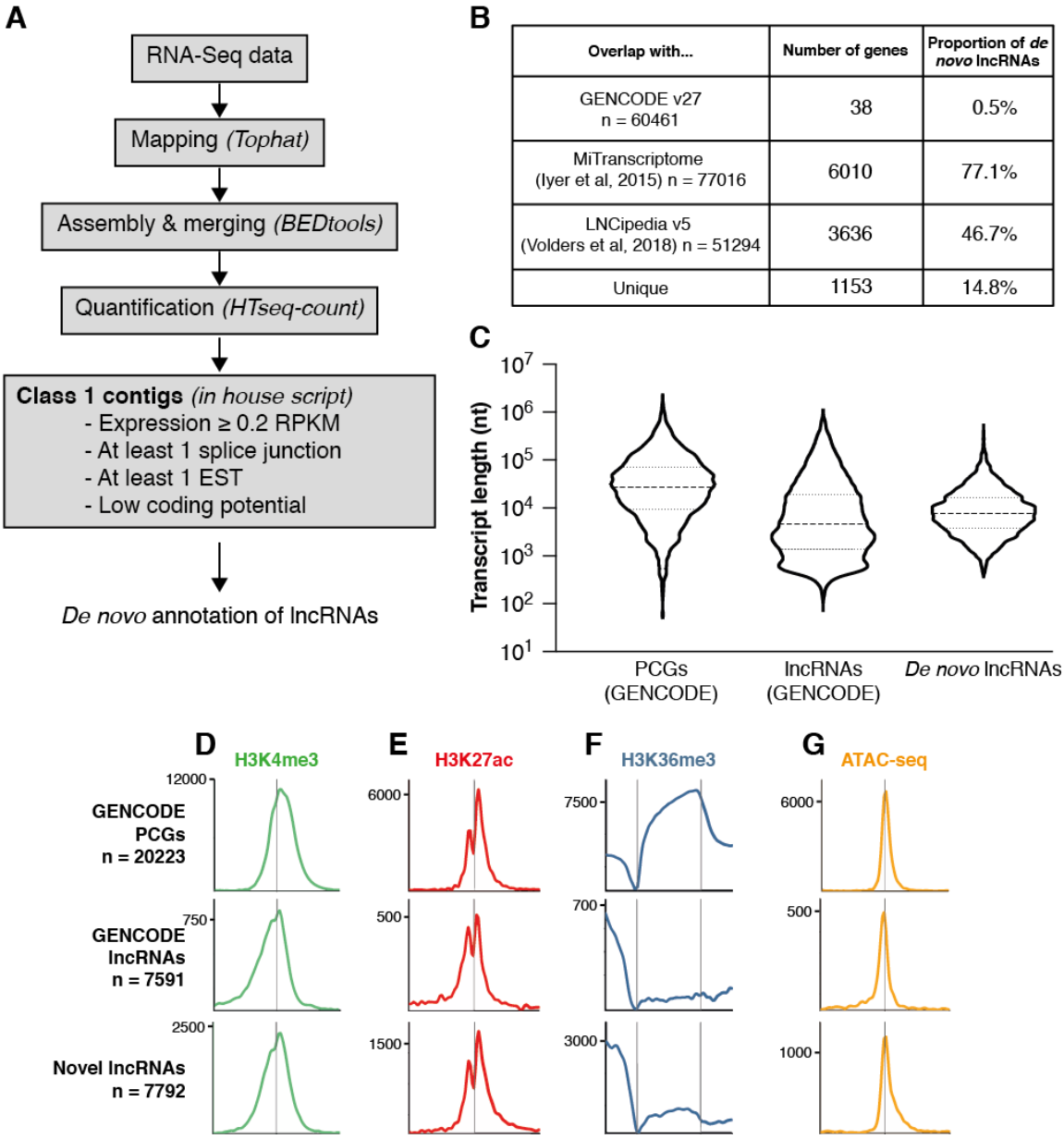
RESULTS

De novo annotation of EMT associated lncRNA reveals new transcripts.

Although most studies rely on the induction of EMT through overexpression of EMT drivers or specific treatments such as TGFβ, meta-analysis have shown strong transcriptomic differences depending on the nature of the induction (Gröger et al., 2012a; Liang et al., 2016). In this study, we take advantage of an original *in vitro* system based on primary Human Epithelial Kidney (HEK) cells which naturally undergo EMT as they enter telomere crisis (Castro-Vega et al., 2013a). This system includes two stable cell lines of epithelial (HA5-Early, here named Epi) or mesenchymal phenotypes (HA5-Late, here named Mes). Comparison between Epi and Mes cells allows the investigation of the EMT program in a stable *in vitro* system, ensuring better insights into the EMT program naturally occurring during malignant transformation. We thus aimed to define a list of differentially expressed (DE) genes in Epi and Mes cells.

First, we used an in-house pipeline (Figure S1A) (Pinskaya et al., 2019) to annotate 7792 lncRNA transcription units from our Epi and Mes datasets. We compared this *de novo* annotation to more recent existing ones with a cutoff of 20% overlap between gene coordinates. It appears only 0.5% overlapped with GENCODE v27 while 77.1% and 46.7% overlapped respectively with the cancer-specific MiTranscriptome (Iyer et al., 2015) and the global lncRNA annotation LNCipedia v5 (Volders et al., 2019) (Figure S1B). In order to check the main characteristics of our *de novo* annotation, we compared them to GENCODE-annotated PCGs and lncRNAs (Figure S1C). First, we checked the size distribution which is slightly higher than annotated lncRNAs ($p < 0.0001$). We also measured the presence of histone marks associated with active transcription around the transcription start site (TSS) (H3K4me3, H3K27ac) and along the gene body (H3K36me3) through Chromatin Immunoprecipitation-seq (ChIPseq), as well as chromatin-accessibility through Assay for Transposase-Accessible Chromatin (ATAC)-seq (Figure S1D-G). Just like GENCODE annotated PCGs and lncRNAs, metagenes for the *de novo* lncRNAs showed

a strong H3K4me3, H3K27ac and ATAC-seq peak around the TSS. Unlike GENCODE-annotated lncRNAs, the *de novo* lncRNAs showed H3K36me3 signal along the gene body, however much lower than for PCGs. Once we checked that the de novo lncRNAs have similar features to existing ones, we aimed to characterize the non-coding transcriptome upon EMT, using the full GENCODE v27 annotation of PCG and lncRNAs as well as our *de novo* lncRNAs.



Overlap with...	Number of genes	Proportion of *de novo* lncRNAs
GENCODE v27 n = 60461	38	0.5%
MiTranscriptome (Iyer et al, 2015) n = 77016	6010	77.1%
LNCipedia v5 (Volders et al, 2018) n = 51294	3636	46.7%
Unique	1153	14.8%

Figure S1. *De novo* annotation of EMT-associated transcripts. (A) *De novo* annotation pipeline used in the study. **(B)** Overlap between *de novo* annotation and existing lncRNA annotations. **(C)** Comparison of length and between PCGs, GENCODE-annotated lncRNAs and *de novo* EMT-lncRNAs. **(D-G)** Metagenes of the peak density for **(D)**

H3K4me3, **(E)** H3K27ac, **(F)** H3K36me3 and **(G)** ATAC-seq experiments for GENCODE-annotated PCGs and lncRNAs, as well as *de novo* EMT-lncRNA annotation.

Subcellular fractionation based RNA-seq allows for a deeper characterization of the non-coding transcriptome.

mRNAs are mainly stable transcripts located to the cytoplasm of cells whereas lncRNAs have been reported to be less stable, if not cryptic, and both cytoplasmic and nuclear fractions. In order to characterize more precisely the non-coding transcriptome, we used subcellular fractionation coupled to RNA-seq to separate processed RNAs exported to the cytoplasm which are representative of cytoplasmic steady-state levels (Cyto-seq) from the ones on the chromatin, more representative of transcription levels and association to the chromatin (Chro-seq) (Figure 1A). Although the subcellular fractionation protocol is different (Gagnon et al., 2014), our approach is based on the work of the Churchman laboratory as we used their method to tether the polymerase onto the chromatin with the transcription inhibitor α-amanitin in order to retrieve nascent and/or chromatin-associated RNA molecules (Mayer et al., 2015). Prior to sequencing, the fractionation was validated at both protein and RNA levels (Figure S2A-B). We checked the genome-wide enrichment of premature transcripts in Chro-seq compared to Cyto-seq by measuring the exonic to intronic signal ratio for each expressed gene with at least 2 exons. This was very clear for PCGs with a strong shift toward intronic signal, however GENCODE-annotated lncRNAs had a very slight shift toward intronic signal and a stronger peak toward exonic signal in the Chro-seq than in the Cyto-seq (Figure S2C). This may be due to lncRNAs being less processed or overall enriched on the chromatin as mature transcripts.

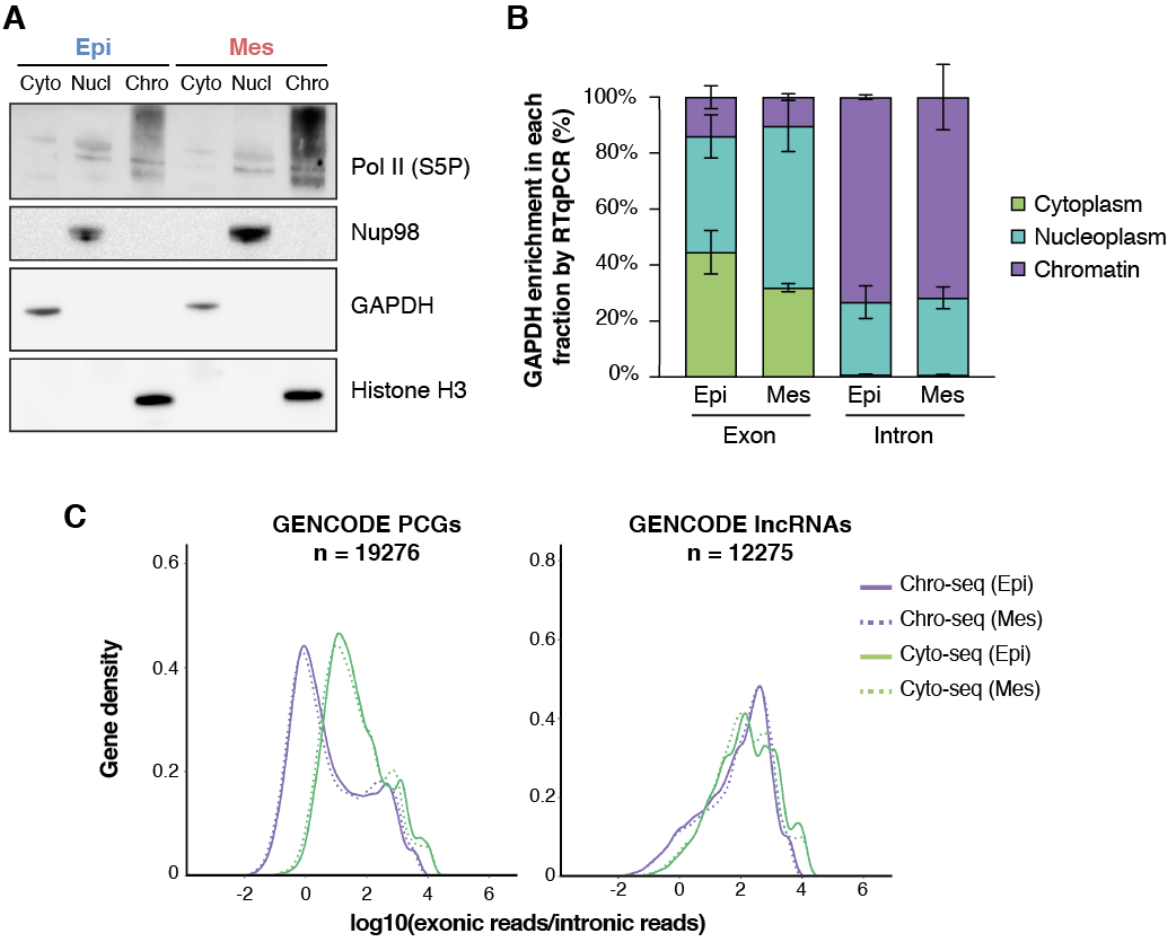
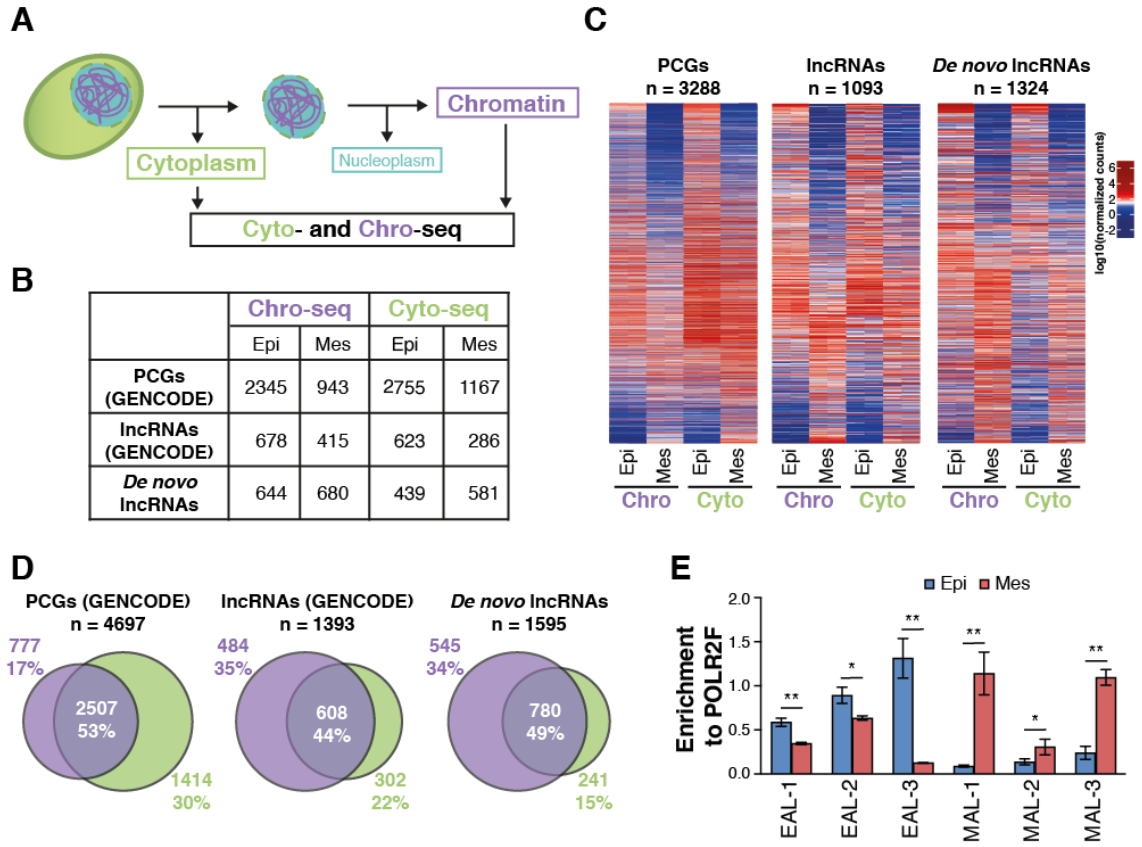


Figure S2. Quality controls for subcellular-fractionation based RNA-seq. (A) Western blot for the control of the subcellular fractionation experiment at the protein level. GAPDH marks the cytoplasmic fraction, the nucleoporin Nup98 marks the nucleoplasm fraction and Histone H3 marks the chromatin fraction. Polymerase II S5P was also assessed to verify its enrichment in the chromatin fraction. **(B)** RTqPCR controls of the subcellular fractionation experiment at the RNA level. "Exon" measurement was done on the junction of exon 3 and 4 of GAPDH while "intron" targeted intron 3 of GAPDH. **(C)** Density plots of the exonic to intronic reads ratio for PCGs and lncRNAs in Chro- and Cyto-seq.

Using Chro- and Cyto-seq, we then defined a list of differentially expressed (DE) genes in Epi and Mes cells (log2(ratio) ≥ ±1 ; p-value < 0.05) (Figure 1B). We observed a higher number of GENCODE-annotated genes downregulated upon EMT in our system, with about 70% of DE-PCGs being enriched in Epi cells compared to Mes cells in Chro-seq (2345 in Epi, 943 in Mes) and in Cyto-seq (2755 in Epi, 1167 in Mes). This proportion was slightly lower for GENCODE-lncRNAs with in Chro-seq (678 in Epi, 415 in Mes) and Cyto-seq (623 in Epi, 286 in Mes). In addition to the GENCODE annotation, our *de novo* gene set allowed the identification of 1565 DE-lncRNAs in

Epi and Mes cells from the two fractions. Interestingly, this study-specific gene set found a higher number of these *de novo* lncRNAs enriched in Mes cells in both Chro- (644 in Epi, 680 in Mes) in Cyto-seq (439 in Epi, 581 in Mes). Although only 14.8% (1153) of these transcripts did not overlap any existing annotation (Figure S1B), the more balanced number of DE-*de novo* lncRNAs between Epi and Mes cells shows that *de novo* annotation may allow for the better definition of a study-specific set of lncRNAs, outside of the rather big and often redundant lncRNA databases which already exist. In order to focus on transcriptionally-regulated lncRNAs, we defined a final list of DE-genes defined from Chro-seq (Figure 1C).



	Chro-seq		Cyto-seq	
	Epi	Mes	Epi	Mes
PCGs (GENCODE)	2345	943	2755	1167
lncRNAs (GENCODE)	678	415	623	286
De novo lncRNAs	644	680	439	581

Figure 1. Chromatin-based characterization of the non-coding transcriptome in EMT. (A) Subcellular fractionation of Epi and Mes cells through sequential lysis to isolate RNAs associated with the cytoplasmic and chromatin fractions for sequencing. **(B)** Table of the differentially expressed PCG, lncRNA and *de novo* lncRNA genes in Epi and Mes cells as defined through DESeq (log2(ratio) > ±1 ; p-value < 0.05). **(C)** Heatmap of the expression profile of DE-genes identified in Chro-seq. **(D)** Venn diagrams representing the overlap between genes defined as differential in Chro- (purple) and Cyto-seq (green). **(E)** RTqPCR validation of some differentially expressed *de novo* lncRNAs in Epi and Mes cells, values are relative to POLR2F, error bars indicate SD. * P<0.05; ** P<0.01.

To further assess the advantages of using Chro-seq for lncRNA differential expression analysis, we compared the overlap between DE genes detected in Chro- and Cyto-seq (Figure 1D). As expected, most of the DE-PCGs could be retrieved in the Cyto-seq while only 17% were exclusively differential in the chromatin fraction. Remarkably, twice as much DE-lncRNAs could be detected using Chro-seq (35%) with an overall overlap between Cyto- and Chro-seq of (44%) lower than for PCGs (53%). Same as GENCODE-annotated lncRNAs, more DE-*de novo* lncRNAs (34%) were found using Chro-seq with only 15% being exclusively retrieved from Cyto-seq. Therefore, it seems nuclear approaches to RNA-seq may be a good tool to define the differential expression of lncRNAs. Indeed, as a way of isolating either chromatin-associated or nascent transcripts, Chro-seq allows for the specific detection of lncRNAs which may be unstable and/or nuclear. Although this is not the focus of this study, together with Cyto-seq as a reflection of the steady-state levels of cytoplasmic mRNAs, subcellular fractionation based RNA-seq may be good tool for characterizing transcripts which are regulated post-transcriptionally, whether coding or non-coding.

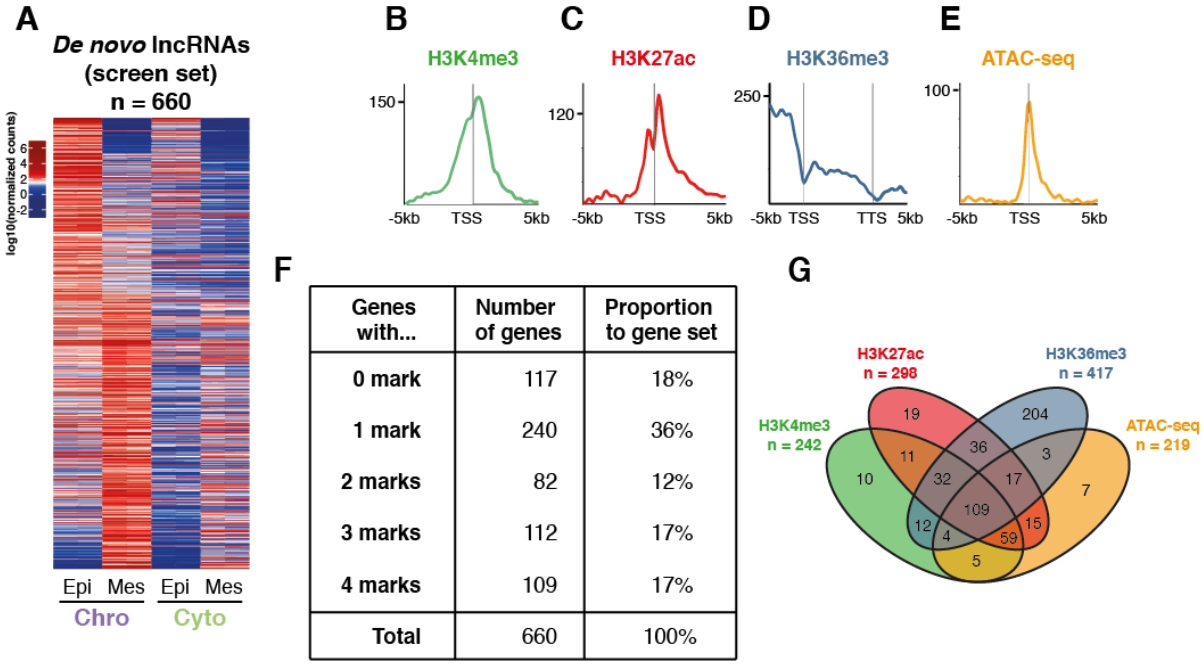
Finally, we confirmed by RTqPCR the differential expression of some *de novo* lncRNA candidates, named here EAL-/MAL- for **E**pithelial/**M**esenchymal identity **A**ssociated **L**ncRNAs (Figure 1E).

CRISPR-based transcriptional activation screen uncovers lncRNAs involved in the regulation of the epithelial identity.

After the identification of the DE-lncRNAs in our system, we wanted to assess their functionality in a high-throughput manner. Given the non-coding nature of lncRNAs, CRISPR techniques bases on Non-Homologous End-Joining are not as effective as for PCGs since mutations may not always impact the expression or functionality of a lncRNA. We therefore decided to use CRISPRa to directly target lncRNA promoter regions as a gain-of-function approach (Joung et al., 2017a).

As CRISPRa directly targets promoter regions upstream of the TSS, we validated the TSS of our DE-*de novo* lncRNA genes. We first defined a reduced screen subset of 660 DE-*de novo* lncRNA genes from Chro-seq in order to keep unique TSS annotations, keeping the longest transcript isoform for each (Figure S3A). As previously, we validated these TSS using ChIP-seq of active chromatin marks and ATAC-seq for

chromatin accessibility (Figure S3B-E). Although not all of these lncRNAs could be associated with a ChIP or ATAC peak, we believe our annotation to be quite robust as 82% of the screen subset lncRNAs had a least one mark and 46% had at least two (Figure S3F-G). Our inability to link the remaining 18% to ChIP/ATAC peaks may solely be due to the overall lower expression of lncRNAs or a lack of sequencing depth in our ChIP- and ATAC-seq experiments.



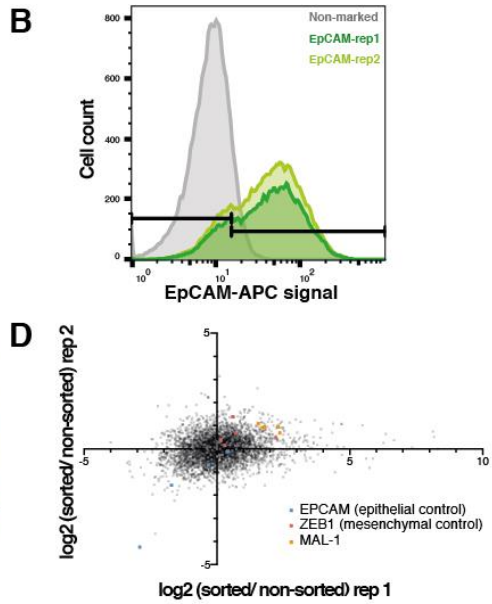
Genes with...	Number of genes	Proportion to gene set
0 mark	117	18%
1 mark	240	36%
2 marks	82	12%
3 marks	112	17%
4 marks	109	17%
Total	660	100%

Figure S3. TSS validation of *de novo* lncRNA annotation by ChIP- and ATAC-seq. (A) Heatmap of lncRNA screen set expression in Epi and Mes cells. **(B-E)** Metagenes of the peak density for **(B)** H3K4me3, **(C)** H3K27ac, **(D)** H3K36me3 and **(E)** ATAC-seq experiments for the screen subset of the *de novo* lncRNA annotation. **(F)** Table and **(G)** Venn diagram showing the number of *de novo* lncRNAs in the subset associated with ChIP- and/or ATAC-seq peaks.

To screen lncRNA functionality, 3 to 5 single guide RNAs (sgRNAs) were designed to target the region upstream of each annotated TSS as suggested in other CRISPRa studies (Joung et al., 2017a; Konermann et al., 2015). In total, our sgRNA library targets the 660 DE-*de novo* lncRNAs as well as 174 DE-GENCODE-annotated lncRNAs. As controls, sgRNA targeting PCG and miRNA genes associated with EMT were also added, such as EMT drivers (ZEB1, ZEB2, SNAI1, SNAI2, TWIST1), EMT inhibitors (ELF3, ELF5), epithelial microRNAs (MIR200 cluster) or genes encoding junction proteins (EPCAM, CTNNB1, TJP3) (Figure S4A).

A

	Number of genes	Number of sgRNAs
Non targetting	-	100
PCG controls	15	75
GENCODE lncRNAs	174	799
De novo lncRNAs	660	3240



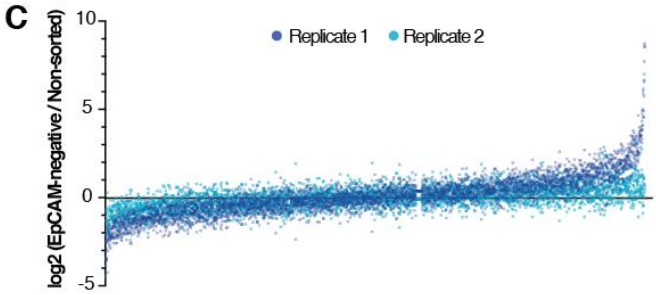
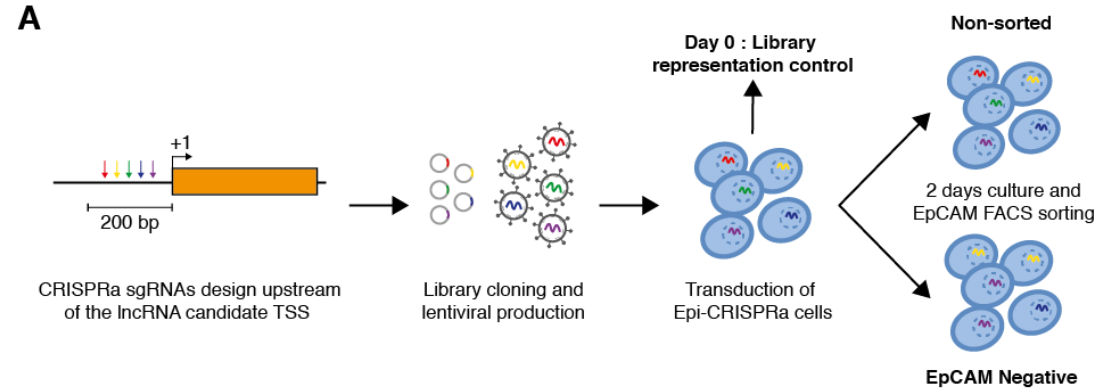
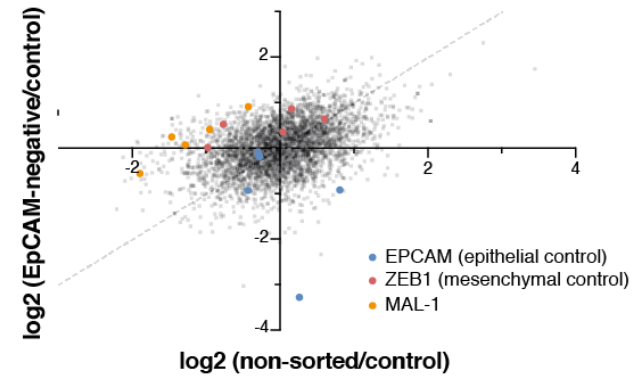


Figure S4. Quality controls of the CRISPRa screen. (A) Composition of the sgRNA library with the number of PCG, annotated lncRNA and *de novo* lncRNA genes targeted as well as the associated number of sgRNAs. **(B)** FACS histogram showing the expression of the epithelial surface marker EpCAM in Epi-CRISPRa cells (dark and light green) as well as non-marked Epi-CRISPRa cells (grey) as control for sorting. The black bars indicate the cells considered EpCAM-negative (left) and EpCAM-positive (right). EpCAM-negative cells were gated as shown and retrieved for further sgRNA analysis. **(C)** Enrichment score for each sgRNA in the two experimental replicates, represented as log2 ratio of readcounts for EpCAM-negative to non-sorted cells. **(D)** Scatterplot of the normalized EpCAM-negative sgRNA counts to non-sorted sgRNA counts for the two experimental replicates.

The sgRNAs library was cloned as previously described (Joung et al., 2017b). Epi cells constitutively expressing the CRISPRa machinery (Epi-CRISPRa) were infected in duplicates with the lentiviral pool and selected for five days, followed by two days of drug-free culture. Using flow cytometry, we then isolated Epi-CRISPRa cells which had lost the expression of EpCAM (Figure S4B). Finally, to assess sgRNA distribution, samples were retrieved after selection as a reference point, as well as the non-sorted and EpCAM-negative sorted cells, followed by genomic DNA extraction and sgRNAs sequencing (Figure 2A). The RNAi Gene Enrichment Ranking (RIGER) (Luo et al., 2008) program was used to correlate sgRNA distribution to the enrichment or depletion of specific target genes.





C

Gene name	Average enrichment score	Average p-value	
MAL-1	1,465	0,00459	Top enriched in EpCAM-neg
EAL-182239	1,555	0,02219	
PRNCR1	1,515	0,02632	
ZEB1	1,435	0,04444	
SNAI1	1,345	0,10663	
EAL-1045149	1,535	0,02505	Top depleted in EpCAM-neg
MAL-1771306	1,425	0,03218	
EAL-550939	1,455	0,03585	
EPCAM	1,67	0,00650	
mir200_locus	1,355	0,09426	

Figure 2. CRISPRa screening of lncRNAs involved in the regulation of the epithelial marker EpCAM. (A) Schematic representation of the CRISPRa screening. sgRNAs were designed upstream of the targeted TSS and pool-cloned into lentiviral vectors. Then Epi-CRISPRa cells (Epi cells with a stable expression of dCas9-VP64 and MS2-P65-HSF1) were transduced at 0.3 MOI, selected by Zeocin for 5 days and cultivated for 2 days to amplify before EpCAM sorting. Cells which lost the expression of EpCAM were retrieved and compared to non-sorted cells. **(B)** Scatterplot of the normalized EpCAM-negative counts to non-sorted counts. All sgRNAs are showed in grey. sgRNAs targeting EpCAM (epithelial positive control) and ZEB1 (mesenchymal positive control) are shown in blue and red respectively. The sgRNAs targeting the top 1 differential lncRNA in Mes cells is shown in orange. The grey dotted lines show the 1:1 ratio line. **(C)** Table with the RIGER average enrichment score and associated average p-value for the top enriched and depleted targeted genes in EpCAM-negative cells. The control genes typically associated with the mesenchymal identity are shown in red and the ones associated with the epithelial identity are shown in blue.

First, comparing EpCAM-sorted to non-sorted cells revealed a depletion of EPCAM-activating sgRNAs as well as an enrichment of ZEB1-activating sgRNAs in EpCAM-negative cells (Figure 2B). These controls validate our approach as CRISPRa transcriptional activation of EPCAM logically prevented the cells from being

EpCAM-negative whereas ZEB1 has been shown to directly bind to the promoter of EPCAM and repress it (Vannier et al., 2013). Some other control genes were also either enriched (SNAI1) or depleted (MIR200) in EpCAM-negative cells but the associated p-value did not pass the 0.05 cutoff (Figure 2C). Although the enrichment for these genes is consistent with what is known about their role in EMT, the p-values score is not significant and the overall number of enriched/depleted genes with a p-value lower than 0.05 was low (8 enriched, 8 depleted).
We focused on lncRNAs and found that the top lncRNA genes enriched in EpCAM-negative cells were not necessarily MAL- lncRNAs, or the opposite for depleted genes (Figure 2C). For example, the lncRNA PRNCR1 is upregulated in Epi cells compared to Mes in our system. However, it has been suggested to promote EMT by downregulating miR-448 (Cheng et al., 2018) therefore its transcriptional activation could very well be linked to a repression of EPCAM. Interestingly, among the enriched genes was the top *de novo* lncRNA overexpressed in Mes cells MAL-1. Quality controls also showed that beside replicate 2 showing a lesser amplitude of distribution than replicate 1 (Figure S4C), the differential distribution of sgRNAs for EPCAM, ZEB1 and MAL-1 were maintained in both replicates (Figure S4D).

As mentioned in the introduction of this chapter, although the necessary validations for the specific MAL-1 targeting sgRNAs were not finished yet at the time this manuscript was written, they are ongoing in order to validate the role of MAL-1 through its activation in cis using CRISPRa outside of a pool-context.

MAL-1 is a nuclear-enriched long non-coding RNA associated with mesenchymal cell identity.

Since this locus seems to be functional in the EMT, we next asked whether MAL-1 simply acts through its transcription or exists as a stand-alone functional RNA molecule. To address this question, we first analyzed the *de novo* annotation as well as the coordinates of the enriched sgRNAs (Figure 3A). This lncRNA is transcribed in Mes cells from the extremity of the short arm of chromosome 6 (6p25.3) as part of a larger locus. Our *de novo* annotation found two variants at this locus, antisense to pseudogenes annotated on the negative strand but which are not transcribed either

in Epi or Mes cells. Considering the sgRNAs enriched in EpCAM-negative cells target the TSS of the shorter transcript, we focused on it.

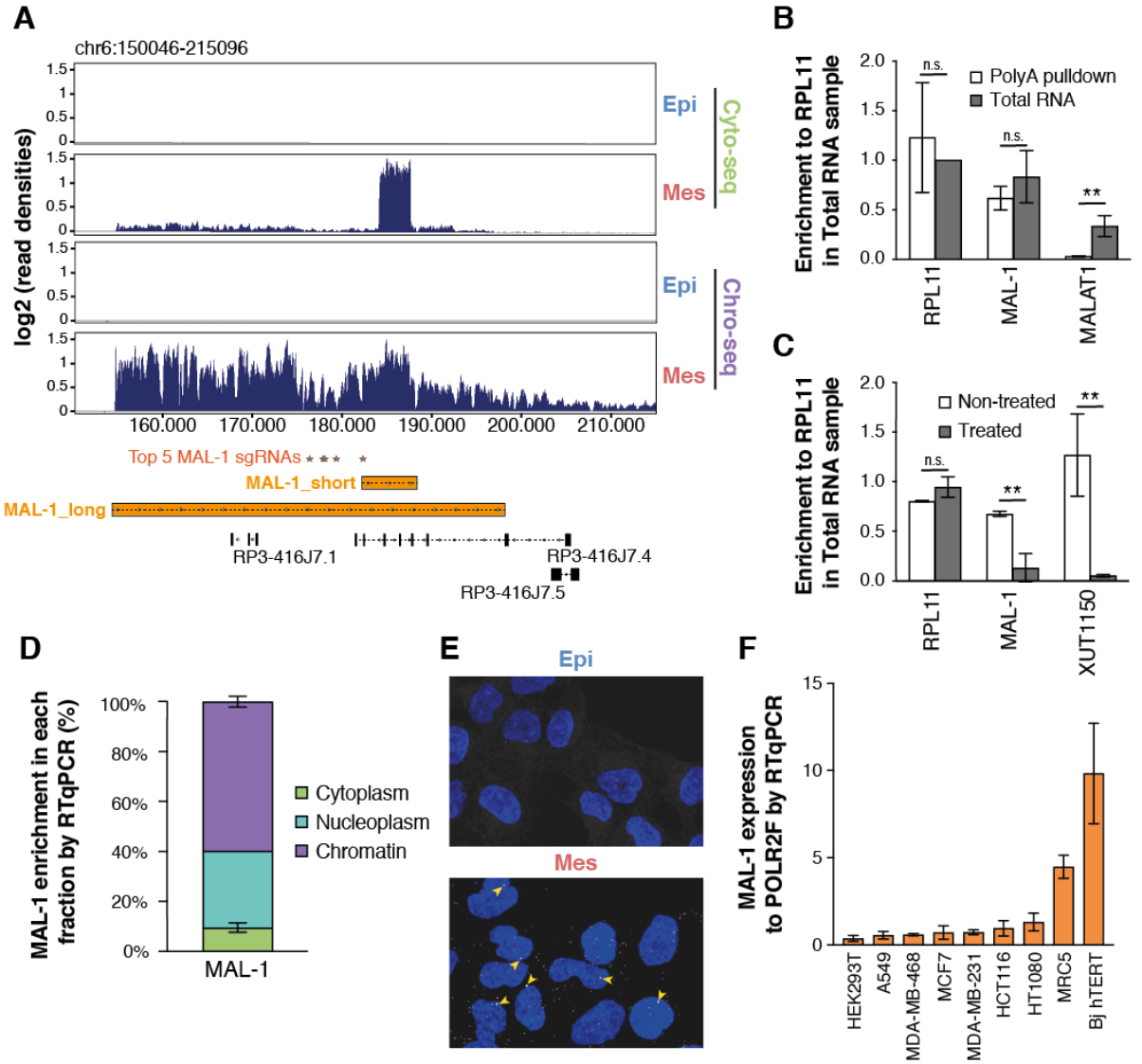


Figure 3. MAL-1 is an uncapped, polyA-tailed and mostly nuclear transcript associated with mesenchymal identity. (A) Cyto- and Chro-seq visualization of the MAL-1 locus in Epi and Mes cells using Ving (Descrimes et al., 2015). **(B)** RTqPCR comparing total RNA and PolyA pull-down experiments for MAL-1, RPL11 (positive control, polyA tail) and MALAT1 (negative control, no polyA tail). **(C)** RTqPCR comparing total RNA samples treated or not with the Terminator 5' Phosphate Exonuclease for MAL-1, RPL11 (positive control) and the yeast lncRNA XUT1150 (negative control). **(D)** RTqPCR of MAL-1 in the cytoplasmic, nucleoplasmic and chromatin fractions of Mes cells. **(E)** smFISH experiment against MAL-1 in Epi and Mes cells (x80). Yellow arrows show foci formations in the nuclei of Mes cells. **(F)** RTqPCR of MAL-1 in different cell lines. Error bars show SD. ** P<0.01.

According to our annotation, as well as polyA-pulldown and Terminator experiments (Figure 3B-C), MAL-1 is a mono-exonic 3828 nt transcript which is poly-adenylated and poorly or not capped. We also confirmed MAL-1 has no or a very low coding potential using tools such as CPC (coding potential score = -0.931), CPAT (coding probability = 0.245) and PORTRAIT (coding probability = 0.447) (Arrial et al., 2009; Kong et al., 2007; Wang et al., 2013). In order to define its localization in the cell, we also performed subcellular fractionation (see Figure 1A) followed by RTqPCR, without tethering the polymerase onto the chromatin as it was done for Chro-seq. Although MAL-1 can be detected in the cytoplasm, it is enriched in the nucleus and particularly on the chromatin (Figure 3D). This was also confirmed by single molecule RNA-FISH in which we see single transcripts in both cytoplasm and nucleus, as well as brighter foci of concentrated transcripts in the nucleus, shown with yellow arrows (Figure 3E).

Finally, we measured MAL-1 expression in various cell lines by RTqPCR and amongst the nine cell lines we tested (Figure 3F), its expression was the highest in MRC5 and Bj hTERT, the only two which are described as fibroblastic by the American Type Culture Collection (ATCC) while the other seven are epithelial. We also checked the expression of the locus in various cancer samples using the TANRIC platform (Li et al., 2015c) and found it to be upregulated in both breast invasive and kidney renal clear cell carcinomas (Figure S5). In addition to MAL-1 being upregulated upon EMT in our system, this suggests its association with mesenchymal cell identity and potentially cancer progression as well.

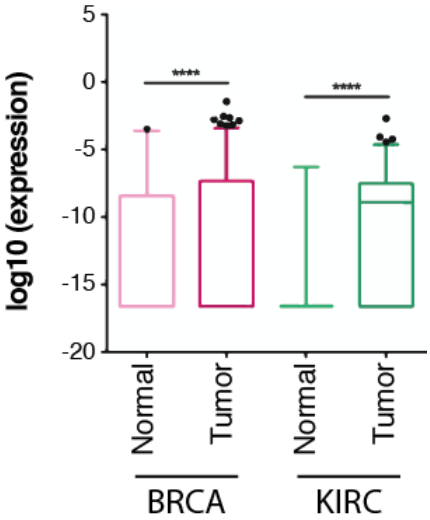


Figure S5. MAL-1 expression in tumor samples using TANRIC. Log10 expression of MAL-1 in different types of cancer from The Cancer Genome Atlas as measured on the

TANRIC platform (Li et al., 2015c), comparing normal and tumor tissues for Breast Invasive Carcinoma (BRCA) and Kidney Renal Clear Cell Carcinoma (KIRC). Error bars display SD; **** P <0.0001.

MAL-1 overexpression in *trans* correlates with a repression of epithelial markers and increase in cell migration.

One of the main advantages of CRISPR-based approaches to modulate transcription is that they directly target loci of interest in *cis*. However, we also investigated whether MAL-1 could act in *trans* as a stand-alone transcript, outside of its genomic context. We thus used the cDNA of MAL-1 cloned from Mes cells to generate the stable Epi_MAL-1 cell line through lentiviral transduction; as well as the Epi_CTR control cell line. MAL-1 expression was checked through RTqPCR and it is expressed over 20 times in Epi_MAL-1 cells compared to the Epi_CTR cell line (Figure 4A). It is also worth noting that the overexpression seems to be at pseudo-physiological levels since it is only 2.5 times stronger than in Mes cells.

First, we checked if MAL-1 overexpression impacted EpCAM levels by FACS as they did in the CRISPRa screen (Figure 4B). Epi_MAL-1 cells showed a decrease of 12% of the average EpCAM signal ($p<0.01$) compared to the Epi_CTR. We then asked if other epithelial markers were repressed and Western blot experiments showed significantly lower levels of β-Catenin ($p = 0.016$) and Claudin ($p = 0.002$) upon MAL-1 overexpression (Figure 4C). These proteins are involved in cell-adhesion and typically expressed in epithelial cells; as shown, they are both repressed during EMT in our system, from Epi to Mes cells. Finally, we assessed if these changes in the expression of epithelial markers could be linked to phenotypic differences. Thus we measured the migratory properties of the cells, a phenotype commonly studied in EMT by wound healing assay. Cells overexpressing MAL-1 displayed a strong increase in migration compared to control, almost as strong as the difference between Epi and Mes cells (Figure 4D-E). We also assessed differences in proliferation and cell cycle but there were no significant differences compared to the control (Figure S6). Altogether, our data shows MAL-1 is involved in the regulation of cell identity in *trans*, correlating with a repression of epithelial protein markers and an increase in migratory properties.

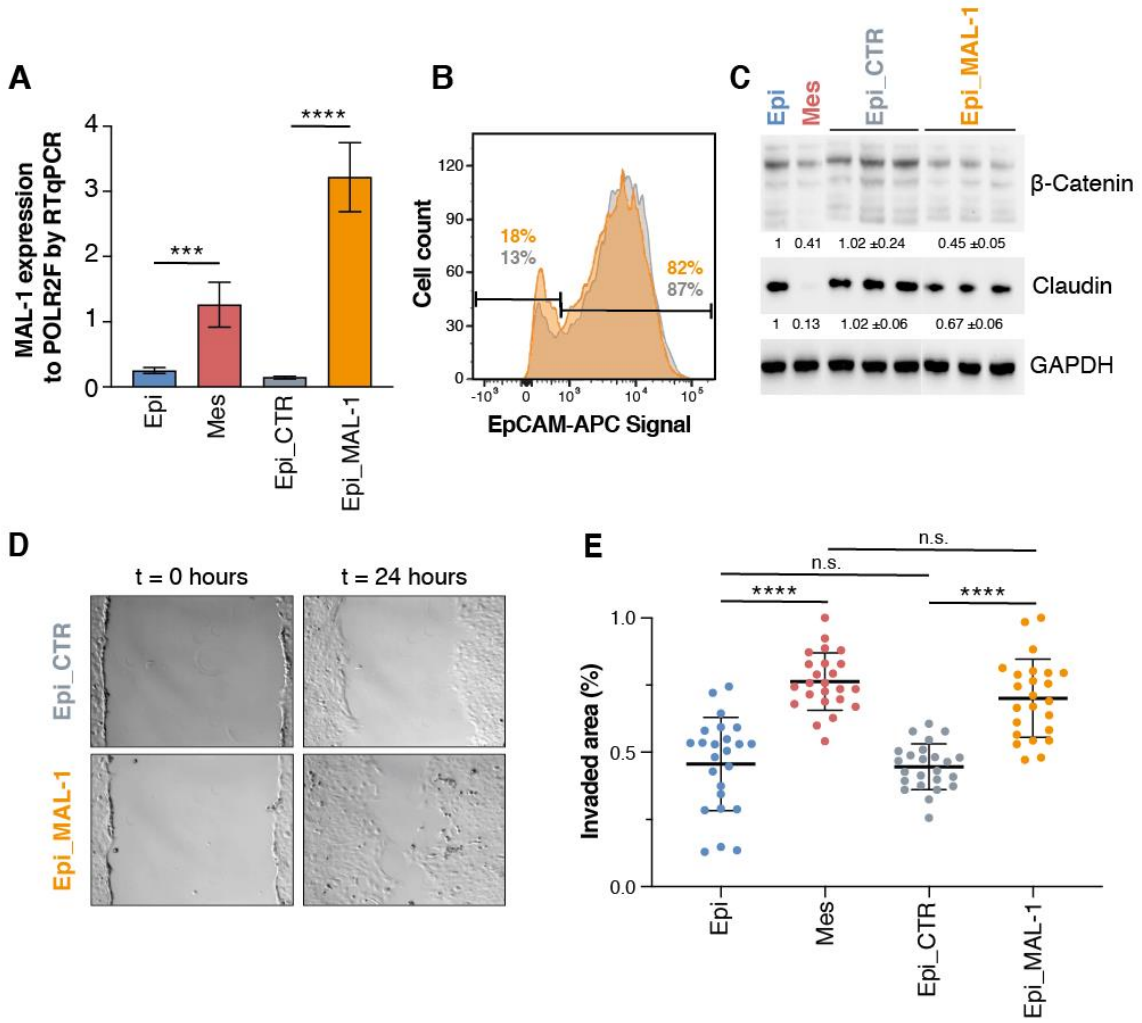


Figure 4. MAL-1 overexpression in *trans* drives a repression of epithelial markers and an increase in migration. (A) RTqPCR of MAL-1 in Epi, Mes, Epi_CTR and Epi_MAL-1 cells. **(B)** FACS histogram showing the expression of the epithelial surface marker EpCAM in Epi_CTR (grey) and Epi_MAL-1 (orange) as well as percentages of EpCAM negative (left) and positive (right) cells. **(C)** Western blot quantification of epithelial proteins β-Catenin and Claudin compared to GAPDH in Epi, Mes, Epi_CTR and Epi_MAL-1 cells. **(D)** Image and **(E)** quantification of the migratory properties of Epi, Mes, Epi_CTR and Epi_MAL-1 cells by wound Healing assay over 24 hours. Error bars display SD; *** P<0.001; **** P<0.0001.

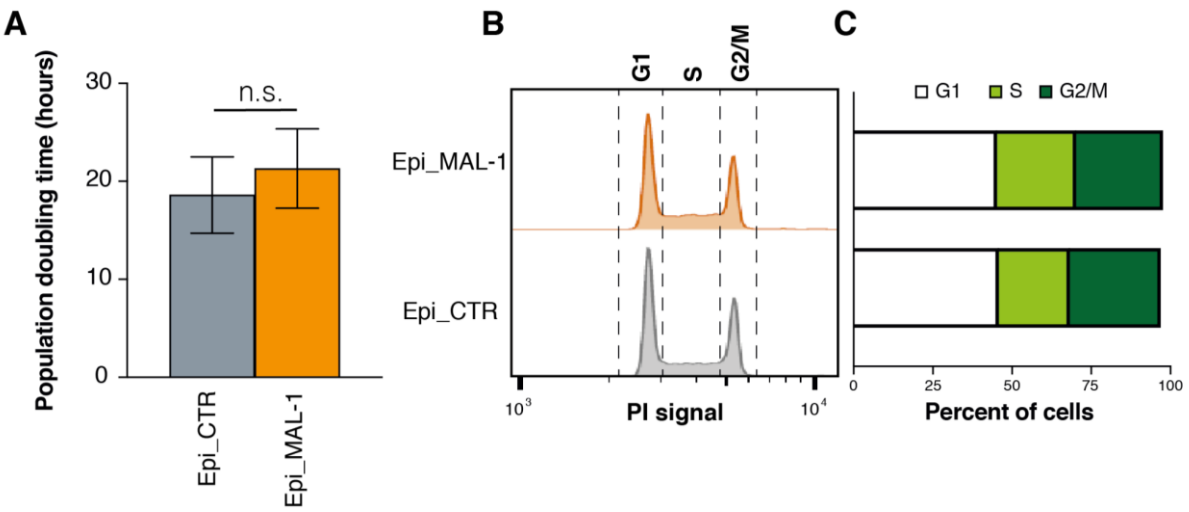


Figure S6. MAL-1 overexpression is not linked to changes in proliferation and cell-cycle progression. (A) Population doubling rate measured in Epi_CTR and Epi_MAL-1 cells. **(B)** FACS histogram and **(C)** quantification showing the distribution of Epi_CTR and Epi_MAL-1 cells to each steps of the cell-cycle through PI-staining.

DISCUSSION

Through the first ever CRISPR-based screen to study the epithelial-to-mesenchymal transition, our study shows the essential role that lncRNAs may have in the regulation of the epithelial phenotype. First, we demonstrated that *de novo* lncRNA assembly coupled to subcellular fractionation based RNAseq are a good tool for the differential analysis of the non-coding transcriptome. Although only 1153 lncRNAs were truly unannotated, using a study-specific set of lncRNAs allowed for a better discovery of differentially expressed lncRNAs compared to existing annotations, especially for the identification of transcripts associated with mesenchymal identity. Using a subcellular fractionation approach to RNA-seq and considering that most lncRNAs are nuclear (Cabili et al., 2015b), our results showed most of the DE-lncRNAs can be identified in the nucleus, contrary to DE-mRNAs which can be retrieved from the cytoplasm. So far, the main limit of Chro-seq is the fact that we cannot separate nascent transcripts from chromatin-associated lncRNAs, such as the ones involved in epigenetic regulation or nuclear architecture (Sun et al., 2018). Coupled to Cyto-seq, it may allow for the discrimination of transcriptionally and post-transcriptionally regulated transcripts, which would not be differential in Chro-seq but would be in Cyto-seq.

For the purpose of this study and to ensure the best efficiency for the CRISPRa screen, we focused on transcriptionally-regulated lncRNAs and defined our list of EMT-associated lncRNA candidates from Chro-seq. We thus proceeded to screen for lncRNAs involved in the loss of EpCAM expression in epithelial cells as a marker of epithelial identity. Our CRISPRa screen showed low amplitude of enrichment and depletion compared to other studies. Indeed, in Epi cells negative for EpCAM, we could only identify 16 functionally relevant lncRNAs (8 enriched, 8 depleted) with a p-value cutoff of 0.05, and only one control gene in each category (ZEB1 enriched, EpCAM depleted). However not significant according to our criteria, some other control genes also seemed enriched (SNAI1) and depleted (MIR200) in the screen. The rather low statistical strength of our analysis could be due to several aspects. First, it could be explained by the small size of our sgRNA library (4214 sgRNAs) compared to other studies, with other CRISPRa libraries typically ranging from approx. 15.000 sgRNAs in a targeted subset, to over 100.000 for genome-wide studies (Horlbeck et al., 2016). Second, unlike most other CRISPR screens which are based on a very stringent phenotype such as drug-resistance, cell proliferation or apoptosis (Gilbert et al., 2014; Joung et al., 2017a; Sanjana et al., 2016), we relied on more subtle changes in cell identity, as measured by the loss of EpCAM.

Among the top enriched lncRNAs in EpCAM-negative Epi cells was the most DE-lncRNA in Mes cells, MAL-1. We showed this lncRNA is enriched in the nucleus of Mes cells where it seems to form foci, although these foci may simply reflect the transcription locus. Its expression seems to be associated to mesenchymal identity as it was found to be highly expressed in fibroblastic cell lines and differentially expressed in some tumor samples, notably breast and kidney. Whether this expression is specifically linked to MAL-1 is however debatable as the RNA-seq data used by the TANRIC platform from The Cancer Genome Atlas project is not stranded, therefore the differential expression could also come from the RP3-416J7.4 pseudogene from the opposite strand. Further experiments are needed to assess MAL-1 expression in various types of cancer. Finally, we assessed the role of MAL-1 through lentiviral overexpression and showed that it can act in *trans* to regulate the epithelial identity. Indeed, upon overexpression in Epi cells, it induces a repression of epithelial proteins involved in the formation of cell junctions. This also correlates with a strong increase in cell migration.

Altogether, our study shows that coupling subcellular fractionation RNA-seq to CRISPRa approaches allows for the identification of functionally relevant lncRNAs in a high-throughput manner. Indeed, the MAL-1 lncRNA found in mesenchymal cells is able to act in the regulation of epithelial identity, inducing a loss of epithelial markers and an increase in cell-migration. More extensive work will be now needed to study the mechanism through which MAL-1 induces such changes and if its role can be generalized to other EMT *in vitro* models as well as tumor samples.

3. Additional data

3.1. CRISPRa-screening of invasion-associated lncRNAs

As mentioned in the introduction of this chapter, the CRISPRa screen of Epi cells was done based on two different phenotypes. The first one is the isolation of Epi cells losing epithelial identity through sorting of EpCAM-negative cells, this is described in the above article. The second one is the isolation of Epi cells with increase migratory properties using transwell inserts also known as Boyden chambers. They consist of plastic inserts added to cell culture dishes at the bottom of which is a PET membrane with 8.0 µm pores through which cells may migrate.

The main idea here was to use the transwell in order to separate cells with increased migratory/invasive properties (at the bottom of the membrane and well) from the rest of the cells. Preliminary experiments showed it was too difficult to find the right time-point for migratory assay over 24 hours alone (with the porous membrane only), for which Epi and Mes cells would have the maximum difference in migratory capacities with this assay (data not shown). Instead, I tested invasion assays by using transwell supplemented with a layer of Matrigel® which cells first have to invade before migrating through the membrane (figure C5-1). I performed this assay over 96 hours with a migration control (without matrigel®) in parallel in order to correct invasion for the amount of cells which could normally go through and then proliferate on the other side of the membrane. At each time point, the top of the transwell was removed and cells at the bottom of the membrane and the culture dish were lifted and counted. Optimal time appeared to be 72 hours as it is when the maximum difference in invasion capacity between Epi and Mes cells can be observed for this assay.

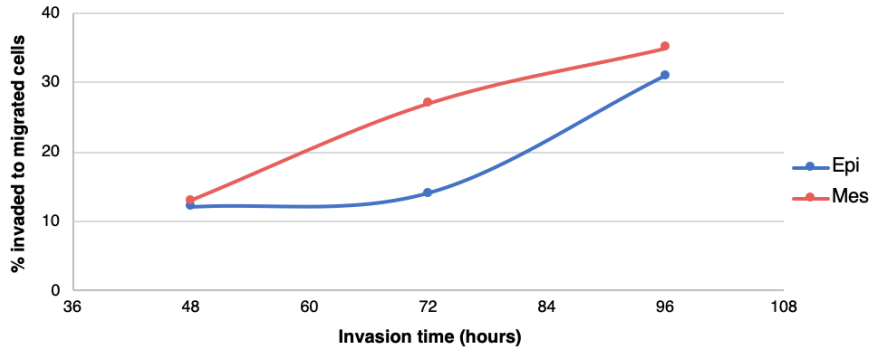


Figure C5-1. Optimization of the invasion assay for the CRISPRa screen over 96 hours. 250.000 cells were seeded on top of the transwell and cells were retrieved by

trypsinization of the bottom of the membrane as well as the cells fallen at the bottom of the culture dish.

From this experiment, the same pool of CRISPRa cells generated for the EpCAM-screen was used in duplicate (figure C5-2A). In order to maintain sgRNA library representation, 2.5 million cells were seeded on top of 10 transwells and incubate for 72 hours before retrieval of the cells with increased invasion capacities (figure C5-2B). In parallel, a portion of the same cells was kept in culture for 72 hours to correct sgRNA distribution for the ones which would impact proliferation. The genomic DNA of the retrieved cells was extracted followed by the sequencing of sgRNAs in the population of cells.

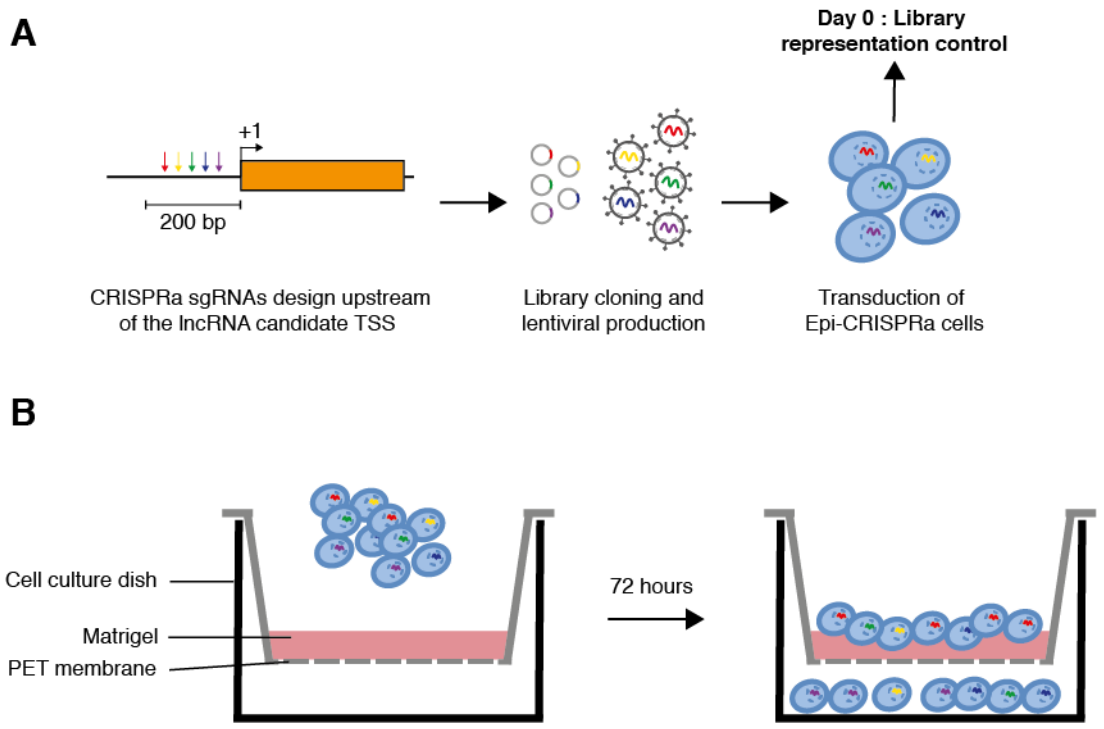


Figure C5-2. CRISPRa screening of lncRNAs involved in the regulation of cell-invasion. (A) Schematic the generation of the CRISPRa cells for screening. sgRNAs were designed upstream of the targeted TSS and pool-cloned into lentiviral vectors. Then Epi-CRISPRa cells (Epi cells with a stable expression of dCas9-VP64 and MS2-P65-HSF1) were transduced at 0.3 MOI, selected by Zeocin for 5 days and cultivated for 2 days to amplify. **(B)** Invasion screening of CRISPRa cells using a transwell with matrigel. Cells were seeded on top of the transwell and incubate for 72 hours before retrieval.

For analysis, each sgRNA was assigned a number of counts and for each duplicate I calculated the sgRNA read count ratios for invasion assay to proliferation assay (figure C5-3A). Although replicates seem less different than for the EpCAM screen,

the amplitude of depletion/enrichment is rather low as log2(ratios) range from -5.227 to 3.464, whereas the EpCAM screen ranged from -4.246 to 8.321.

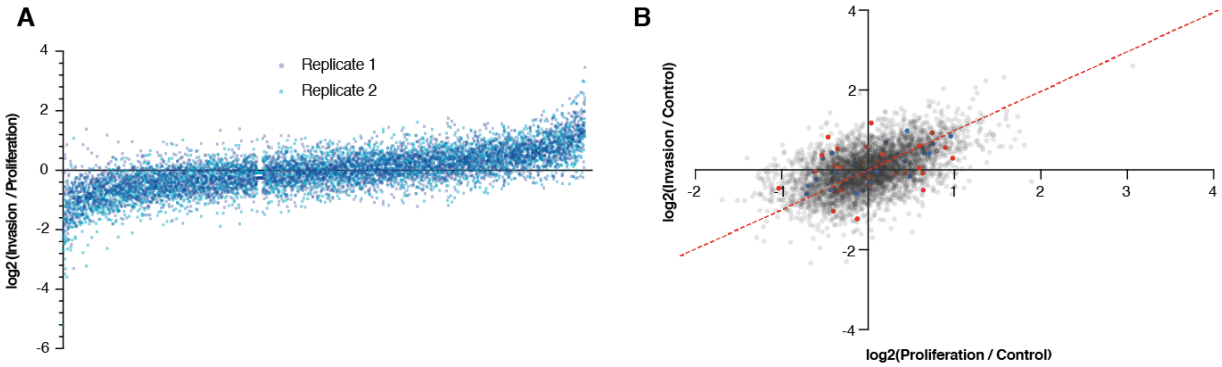


Figure C5-3. sgRNA enrichment analysis for the CRISPRa invasion screen score. **(A)** Enrichment score represented as log2 ratio of readcounts for cells retrieved in the Invasion experiment relative to cells after proliferation. **(B)** Scatterplot of enrichement scores of invasion assay against the proliferation control. Here ratios are calculated to the initial library representation at day 0. sgRNAs in blue and red represent the positive controls associated with the epithelial or mesenchymal state, respectively.

Then, I used the RNAi Gene Enrichment Ranking (RIGER) (Luo et al., 2008) program to correlate sgRNA distribution to the enrichment or depletion of specific target genes, with parameters specified in the initial CRISPRa paper from the Zhang laboratory (Joung et al., 2017b). This was done for enrichment and depletion separately and I retrieved a ranked list of genes with an enrichment score and a p-value. However, I did not manage to retrieve any control genes either in the enrichment or depletion analysis (figure C5-3B). Indeed, none of the control genes were enriched or depleted significantly and they seem randomly distributed among other sgRNAs.

Table C5-I. lncRNAs significantly enriched or depleted in the invasion screen.

	Gene id	Average enrichment score	Average p-value
Top 4 enriched	AL161431.1	1,56	0,02570
	EAL-2431056	1,545	0,02716
	EAL-928497	1,43	0,04043
	AL355075.4	1,42	0,04413
Top 4 depleted	EAL-483780	1,58	0,02026
	EAL-668246	1,47	0,03576
	EAL-177130	1,405	0,03716
	EAL-847589	1,39	0,04029

The overall number of genes significantly enriched was very low with 8 lncRNAs with p-value of enrichment scores below 0.05 (see table C5-I). As previously mentioned, this may be due to the small size of our sgRNA library giving low statistical power to the analysis. Also, as most other CRISPR screens which are based on a very stringent phenotype such as drug-resistance, cell proliferation or apoptosis (Gilbert et al., 2014; Joung et al., 2017a; Sanjana et al., 2016), this screen relies on invasion and migration which are much more volatile phenotypes to select. Altogether, I decided not to proceed further with the analysis of this experiment and focused on the EpCAM CRISPRa screen mentioned in the article draft.

3.2. MAL-1 knock-down using siRNAs

As the EpCAM CRISPRa screen was more successful, I decided to focus on the lncRNA MAL-1 which I was already working on prior to its validation in the screen, as it is the most differentially expressed lncRNA in Mes cells compared to Epi. As mentioned previously, I studied MAL-1 overexpression in Epi cells using a lentiviral construct to characterize its effects on epithelial identity. In addition, I also tried knocking it down in Mes cells to measure its role in the maintenance of the mesenchymal phenotype. To do so, I designed 2 siRNAs targeting MAL-1 (siRNA-A and -B), as well as a control siRNA scrambled (siRNA-Scr) for the siRNA-A sequence.

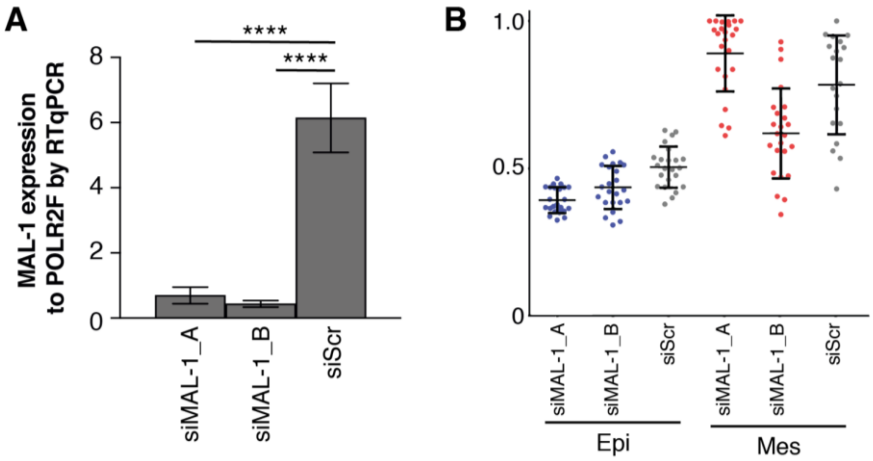


Figure C5-4. siRNA-mediated knock-down of MAL-1 does not affect the migration of Mes cells. (A) RTqPCR of MAL-1 expression in Mes cells upon 24 hours treatment with siRNA against MAL-1 (siMAL-1_A and siMAL-1_B) or a control siRNA (siScr). Values are relative to POLR2F expression. **(B)** Quantification of the migratory

properties of MAL-1 in Epi and Mes cells upon siRNA treatment. Error bars display SD; **** P <0.0001.

I tried several conditions of siRNA treatment, over 24 and 48 hours, which induced in both cases an 80% depletion when measured by RTqPCR (figure C5-4A). Despite the fact that MAL-1 repression was efficient with both siRNA-A and -B, phenotypic assays showed great discrepancies. Indeed, 24 hours after siRNA treatment, I performed wound healing assay on Mes treated cells (as well as Epi cells as a control) and siRNA-B induced a repression of cell migration in Mes cells compared to siRNA-Scr, but siRNA-A did not (figure C5-4B). The decrease in motility associated with siRNA-B may be due to off-target effects, altogether this suggests MAL-1 is not essential for the maintenance of migratory properties in mesenchymal cells.
To check if other changes could be observed upon siRNA treatment, I also measured the expression of some EMT markers by RTqPCR (figure C5-5). The effects of siRNA-A and -B were not consistent with each other as the variations in repression for OCLN and SNAI1 were significantly different between treatments or had completely opposite effects for FN1. Again, this suggests non-specific effects of the siRNA used which may impact other genes involved in the regulation of mesenchymal identity.

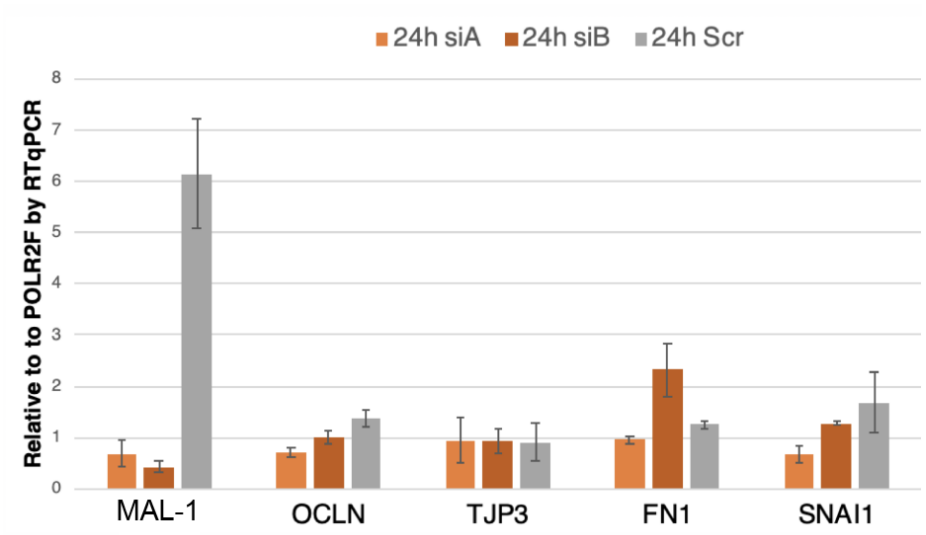


Figure C5-5. RTqPCR of EMT markers in Mes cells upon siRNA treatment against MAL-1 with siRNA-A (light orange), siRNA-B (dark orange) and Scr (grey) over 24 hours. Values are relative to POLR2F.

However, the lack of consistent response despite siRNA-mediated depletion of MAL-1 being very efficient may be due to siRNA degradation mostly happening in

the cytoplasm. The fact that there may still be non-degraded MAL-1 in the nucleus suggests the potential importance of the nuclear localization of MAL-1 for its function, as its cytoplasmic knock-down does not affect its associated phenotype. More experiments need to be performed to confirm this hypothesis. Particularly, I will try depleting MAL-1 using AntiSense Oligonucleotides (ASO) which act throught the cell (including the nucleus) and trigger RNase H degradation of the transcripts. The lackluster effect of siRNA treatment against MAL-1 may also be due to the fact that treatment was only done over 24 hours before RNA-extraction and 48 hours before final wound healing assay readout. In the case MAL-1 is involved directly or indirectly in epigenetic reprograming, the treatment might be too short to measure any effect. Finally, MAL-1 may also not be essential for the maintenance of the mesenchymal identity but rather involved in the initial reprograming of epithelial identity.

3.3. Transcriptomic analysis of MAL-1 overexpression

As MAL-1 is mostly located in the nucleus, it may be involved like many lncRNAs in the regulation of gene expression to repress epithelial identity rather than to maintain mesenchymal identity.

In order to define changes in epithelial identity at the transcriptomic levels upon MAL-1 expression, I performed a Total RNA-seq experiment on Epi_MAL-1 and Epi_CTR cells followed by differential expression analysis.

Compared to the strong changes in migration properties, the number of differentially expressed genes is quite low (table C5-II). Indeed, there is only a total of 375 deregulated genes, coding (325) and non-coding (50), even with a low fold-change cutoff of 1.5 fold. Interestingly, more genes are repressed (262) upon MAL-1 overexpression than upregulated (112).

Table C5-II. Differential expression analysis of MAL-1 overexpression. Analysis was done using DESeq with GENCODE v27 as reference, p-val<0.05, |FC|>1.5)

	PCGs	Non-coding	Total
Upregulated	94	18	112
Downregulated	231	31	262
Total	325	50	375

In order to define more precisely, the processes impacted by MAL-1 expression, I performed gene-ontology analysis of the DE-genes using the GSEA and DAVID online plateforms. Looking at the cellular compartment, the majority of upregulated genes encodes proteins which localize inside of the cell while most of the downregulated ones are associated with the extracellular matrix and the basement membrane.

Quite surprisingly, although no specific pathway seemed upregulated, many downregulated genes were involved in EMT-associated processes and "epithelial-to-mesenchymal transition" was the top hallmark in GSEA (p-value = $1.14e^{-26}$). This seems contradictory and many genes repressed upon MAL-1 expression appear to be mesenchymal markers such as Fibronectin 1 FN1, Metalloproteases MMP2, MMP16 and ADAM12, or Collagens type IV, V and VI; all involved in the formation of the extracellular matrix, typically synthesized by fibroblasts.

However, some repressed genes such as Laminin LAMA1 or Collagens type IV COL4A4 are involved in the formation of the basement membrane, a structure maintained by epithelial cells which prevents them from losing polarity, adhesion and tissue cohesion (Rodriguez-Boulan and Macara, 2014). As previously seen by Western blot and the down-regulation of junction proteins β-Catenin and Claudin, it appears MAL-1 expression affects cell-cell and cell-matrix adhesion properties through the repression (direct or indirect) of genes involved in extracellular matrix organization, cell-cell and cell-membrane adhesion, which could partially explain the increase in migration.

In addition, positional analysis of the differential genes showed enrichment for two specific large loci: 17 downregulated genes are located on chr15q25 and q26 cytobands, as well as 16 upregulated genes on chr6p22 to p25 (figure C5-6). Quite interestingly, the upregulated chromosome 6 domain contains the endogenous MAL-1 locus (chr6p25.3) although MAL-1 overexpression is done here using a lentiviral construct with random genomic integration. In this experiment, we cannot distinguish the lentiviral from the endogenous MAL-1 from RNAseq data but the larger MAL-1 locus did not seem to be upregulated as MAL-1_long expression was unchanged (not shown). In bladder cancer, the chr6p22 locus has been shown to be amplified and could be linked to EMT (Bellmunt, 2018). This has mainly been explained by the fact that it contains the SOX4 gene, a transcription factor which

regulate the expression of EMT drivers such as the SNAI and ZEB families (Lourenço and Coffer, 2017). However, the expression of SOX4 itself is unchanged upon MAL-1 overexpression.

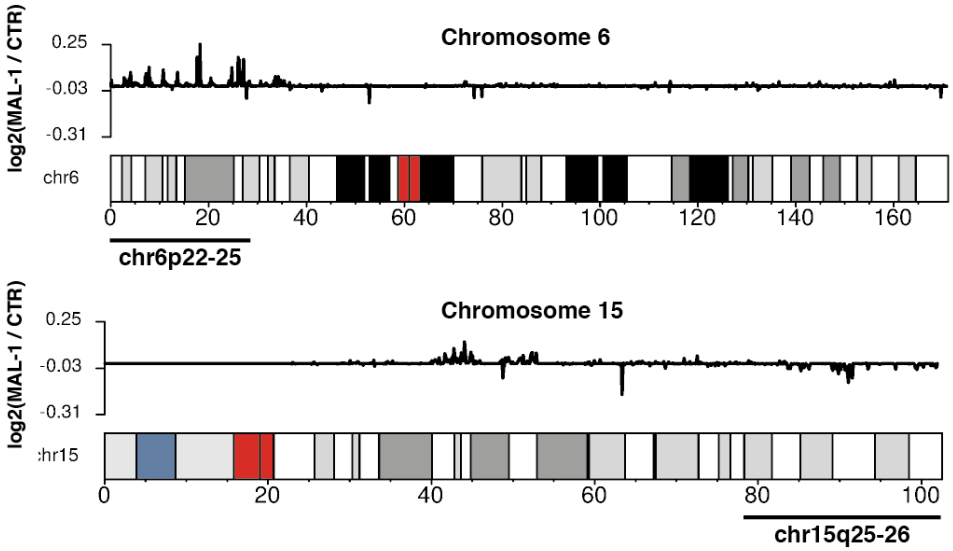


Figure C5-6. Karyoplots of differential gene expression for chromosomes 6 and 15. Values are calculated on 100 kb windows as the normalized read ratios log2(Epi_MAL-1/Epi_CTR) and plotted along the chromosome.

Altogether, MAL-1 is a new lncRNA expressed from mesenchymal cells which is able to increase migration through the repression, direct or indirect, of genes involved in the formation of cell junctions, cell adhesion and the basal membrane of epithelial cells. Its localization being mainly nuclear and its association with the differential expression of large chromosome domains suggest MAL-1 could very well be a new example of lncRNA regulating gene expression. However, we have yet to unravel the mechanism through which it acts to induce such changes in the epithelial cells.

DISCUSSION

CHAPTER 6

Chapter 6. Discussion

As mentioned in the introduction, high-throughput sequencing techniques allowed for the discovery of many new long non-coding RNAs, associated with a variety of cellular processes, in pathological and physiological contexts. The process which I focused on during my PhD is the epithelial-to-mesenchymal transition with the aim of identifying novel lncRNAs involved in its regulation. The non-coding transcriptome is heavily remodeled upon canonical EMT induction through the TGFβ treatment and several lncRNAs were shown to regulate the process itself (Gugnoni and Ciarrocchi, 2019; Liao et al., 2017).

1. Cyto- and Chro-seq as a tool to study the non-coding transcriptome

Many lncRNAs are involved in the regulation of gene expression, whether it is through epigenetic modifications, direct transcription regulation, nuclear assembly or interaction with transcripts of interest. In recent years, some labs have focused on deciphering what underlies the nuclear localization of lncRNAs as well as the identification of chromatin-enriched cheRNAs (Shukla et al., 2018; Werner and Ruthenburg, 2015b). Indeed, Werner and colleagues have shown the majority (60%) of lncRNAs can be found enriched on the chromatin, as nascent transcripts tethered to the transcription locus. CheRNAs were later confirmed to act as activators of transcription for nearby genes and Werner suggested chromatin-enriched RNAs are the most effective chromatin-signature in a very cell-type specific manner (Werner et al., 2017).

In addition, our lab has a lot of experience with transcription-based approaches, notably through Nascent Elongating Transcript (NET)seq in yeast, and we decided to apply a similar methodology on our cells. At the beginning of my PhD, the Proudfoot and Churchman labs had recently published their methods for mammalian NETseq (Mayer et al., 2015; Nojima et al., 2015) and I decided to go with the latter, as it aligned with the work of the Werner group. Both Werner and Churchman group relied on the use of sucrose gradient for subcellular fractionation but I favored another approach which I had already set up in the lab during my master (Gagnon et al., 2014). This subcellular fractionation protocol relies solely on differential lysis to first separate cytoplasm from nucleus through hypotonic lysis,

and then nucleoplasm from chromatin. One of the advantages of this method is that unlike sucrose cushions, it is designed for optimal nuclei isolation without remains from cytoplasmic organelles such as mitochondria or the endoplasmic reticulum.
Therefore, I used this subcellular-fractionation method coupled to the mammalian NETseq from the Churchman laboratory, using the transcription inhibitor drug α-amanitin in the lysis buffer to block the polymerase onto the chromatin and thus enrich for nascent transcripts. For sequencing, NETseq relies on the isolation of the 3'-end of nascent transcripts as a snapshot of the location of the polymerase during transcription. Altogether, this is a much more precise tool than Pol-II ChIP-seq as it is a strand-specific method. In addition, unlike run-on methods such as GRO-seq, the whole process is done on ice and does not rely on a cellular stress such as the block and release of transcription.

Initially designed to study the process of transcription itself, the NETseq library preparation has very low coverage and was not so relevant to this study as I wanted to identify lncRNAs but also visualize their transcription profiles by comparing Cyto-seq and Chro-seq, to identify mature transcript and transcription unit. A good example of this is the visualization of MAL-1 for which Cyto- and Chro-seq allowed the separation of a wide transcription locus on the chromatin to a smaller stand-alone transcript in the cytoplasm (see chapter 5, figure 3). I therefore used a total RNAseq library preparation on the RNAs extracted from cytoplasm and chromatin fractions.
Although this is not the focus of my work, subcellular fractionation based RNAseq could also be used to study post-transcriptionally regulated transcripts. Indeed, one could compare genes with unchanged Chro-seq signal between conditions (same transcription) to differential expression in Cyto-seq (steady-state).

As predicted from the work of the Werner group, Chro-seq allowed the identification of many differential lncRNAs to establish a signature associated with EMT in our system. The most striking observation was that most differential lncRNAs could be identified in the Chro-seq compared to Cyto-seq (chapter 5, figure 1). It also showed more specificity for lncRNAs compared to protein coding genes for which there was a larger overlap of genes detected by Chro- and Cyto-seq.

Altogether, chromatin-based approaches are a good tool for the characterization of the non-coding transcriptome. In our system, it allowed the discovery of many nuclear-enriched lncRNA which are differentially expressed upon EMT. In this next part, the function these lncRNAs may have in the regulation of EMT will be discussed.

2. The role of lncRNAs in the epithelial-to-mesenchymal transition

2.1. HOTAIR as a modulator of Lsd1 function

Much of the work studying HOTAIR has shown its function can be mainly attributed to its interaction with PRC2 (Gupta et al., 2010a; Kogo et al., 2011a). However, recent work has suggested PRC2 could be dispensable for some HOTAIR-mediated regulations (Portoso et al., 2017a). Indeed, Portoso and colleagues showed HOTAIR can regulate chromatin structure and transcription independently of PRC2; suggesting that PRC2 interaction with lncRNAs might serve a function other than guiding it onto specific loci.

In our work, we showed through overexpression of truncated variants of HOTAIR that its overexpression in epithelial cells promotes migration in a Lsd1-dependent manner (chapter 4, figure 1). Interestingly, Lsd1 is a well-known regulator of EMT which has been shown to interact with the SNAI and ZEB EMT-TF to both repress and activate EMT, thus suggesting its context-dependent function (Ferrari-Amorotti et al., 2013; Goossens et al., 2017; Lin et al., 2010a, 2010b; Skrypek et al., 2017; Wang et al., 2007b).

Despite no changes in migratory properties, HOTAIR was however able to induce transcriptomic changes independently of Lsd1 (chapter 4, figure 2 and 3). Most of these changes were however dependent on HOTAIR being able to interact with both Lsd1 and PRC2, each having seemingly distinct roles. The Lsd1-interacting domain of HOTAIR seems to be involved in the regulation of genes associated with cell-junctions and the formation of the extracellular matrix while the PRC2-interacting domain induced differential expression of intracellular signaling pathways.

As the stronger phenotypic changes were associated to Lsd1, we performed the ChIP-seq of Lsd1 in HOTAIR-overexpressing cell lines in order to define its distribution across the genome (chapter 4, figure 4). Strikingly, upon the expression of HOTAIR

variants which can interact with Lsd1, we detected a lot less peaks of Lsd1 association to the chromatin compared to the HOTAIR variant truncated for Lsd1-interacting domain and control. It seems that HOTAIR expression induced the dislocation of Lsd1 from many of its target genes, resulting in the repression of epithelial genes and the promotion of a partial mesenchymal identity and migratory properties.
Further experiments will now be needed to define whether HOTAIR blocks Lsd1 from binding to specific loci or protein partners, or if it actually triggers the assembly of new Lsd1 complexes and/or its recruitment to other genomic regions.

First, we should perform RNA pulldown-based experiments to confirm at RNA levels that HOTAIR is or is not able to interact with Lsd1 depending on the truncated variant, and also at protein levels perform RIP (RNA ImmunoPrecipitation) of Lsd1 to check that if it interacts with HOTAIR or not. We will also perform ChIRP-MS (Chu and Chang, 2018) to identify potential new partners of HOTAIR in this system in a proteome-wide manner. As they were done for the first identification of HOTAIR protein partners PRC2 and Lsd1, pull-down should be performed in our HOTAIR cell lines as well as co-IP experiments with Lsd1 to define new partners and complexes. For example, Lsd1 can form multiple complexes apart for Lsd1-coREST-REST for repression, it can also interact with androgen (AR) and estrogen receptors (ER) to actually activate transcription (Metzger et al., 2005b; Nair et al., 2010). Interestingly, HOTAIR was also shown to interact with both the AR and ER pathways in cancer (Xue et al., 2016; Zhang et al., 2015).
Through its demethylase activity, Lsd1 acts as a repressor with NuRD and REST-coREST, Lsd1 demethylates H3K4; or as an activator of transcription with AR-PKC and ER-PELP1 where it demethylates H3K9. Since our result suggests HOTAIR displaces Lsd1, we should also look at the distribution of various histone modifications in our system such as H3K4me3 or me2, as well as H3K9 methylation and acetylation. The changes in histone marks could very well hint at where Lsd1 goes once displaced by HOTAIR and with which known-partners it may act.
Our work thus identifies HOTAIR as an effector of Lsd1 function as a guardian of epithelial identity. Indeed, we demonstrated that the Lsd1-interacting domain of HOTAIR is essential to promote epithelial cell migration through the displacement of Lsd1 on the genome. As mentioned, Lsd1 can either promote or repress EMT

depending on the cellular context and HOTAIR may very well be providing that context, showing its role as another layer of regulation to modulate the function of Lsd1 in EMT.

2.2. MAL-1, a novel lncRNA repressor of epithelial identity

In addition to HOTAIR, I also identified the novel lncRNA MAL-1 in Mes cells. This lncRNA is the most differentially expressed lncRNA compared to Epi cells in which its locus is seemingly off, in both Chro- and Cyto-seq. It is transcribed from a wide locus (approx. 42 kb) from which emerges a shorter (3.8 kb), monoexonic, and partially cytoplasmic transcript which I named MAL-1 (chapter 5, figure 3, A-C). Although it is unclear how many transcripts are synthesized from this wide locus, the localization of the CRISPRa sgRNA guides suggested it has an independent TSS amid the larger locus. However, we have yet to define the precise structure of transcription on this locus. Indeed, the HoldUp annotation actually shows a TSS slightly upstream of what is observed on the transcription profile and a potential 5' truncation of the MAL-1 transcript could explain the absence of Cap structure, although it is polyadenylated.
Considering its highly differential expression, I aimed at characterizing the features of MAL-1 and first looked at its localization in the cell, as a first hint of its potential function (chapter 5, figure 3, D and E). Subcellular fractionation and RTqPCR showed MAL-1 is highly enriched in the nucleus, and particularly in the chromatin fraction, even without Pol-II tethering. Although only 9.5% of MAL-1 transcripts was detected in the cytoplasm, smFISH experiments suggest this figure is under-estimated, probably due to normalization of the subcellular fractionation. Using FISH, MAL-1 can be seen throughout the cell, but bright foci are found in the nuclei. This explains the high amount of MAL-1 found in the chromatin fraction and is probably due to high transcription levels, the brighter foci being the locus of transcription itself. Although the Cyto-seq transcription profile shows high levels of the short MAL-1 transcript, it is however still present at higher levels in the nucleus.

As I performed the EpCAM-CRISPRa screening to investigate functionally relevant lncRNAs in our system, I found the 5 guides targeting the MAL-1 promoter to be highly enriched in EpCAM-negative cells. This means that transcriptional activation

of MAL-1 in *cis* induced the loss of EpCAM and therefore a partial loss of epithelial identity. As mentioned in the manuscript, the validations of the sgRNAs targeting MAL-1 is still ongoing at the time since the EpCAM-CRISPRa screen was actually performed in my last year of PhD.
In addition to overexpressing it, I also knocked-down MAL-1 in Mes cells (see chapter 5, 3.2). However, there were no phenotypic changes in the cells. As mentioned earlier, this could be due to RNAi pathways mostly degrading transcripts in the cytoplasm, nascent MAL-1 transcription still ongoing. To confirm this, further experiments need to be done to directly knock-down MAL-1 expression as may be done using ASOs. Altogether the nuclear localization of MAL-1 and the effect of its CRISPRa activation in *cis* suggest a potential role in the regulation of gene expression upon EMT.

In order to define if MAL-1 function is linked to its transcription or if it acts as a stand-alone RNA molecule, I cloned the MAL-1 cDNA from Mes cells into a lentiviral vector and transduced Epi cells with it. This generated the Epi_MAL-1 cell lines which overexpressed MAL-1 under a CMV promoter in epithelial cells.
Similarly to the phenotypes seen for HOTAIR in Epi cells, MAL-1 overexpression correlated with a strong increase in cell migration and the repression of epithelial markers at protein levels (EpCAM, Claudin, β-Catenin) (chapter 5, figure 4). However, mesenchymal markers were not upregulated and there were no striking morphological changes upon MAL-1 expression suggesting its expression does not induce a full EMT but rather the repression of epithelial traits in our system.
Surprisingly, the transcriptomic analysis of these cells actually showed the repression of many genes associated with the mesenchymal identity in the literature such as fibronectin FN1 or metalloproteases (MMP2, MMP16, ADAM12) (Lamouille et al., 2014b) (see chapter 5, 3.3). However, in our system MMP2 and MMP16 are actually repressed upon EMT, being more expressed in Epi cells compared to Mes. This is surprising but suggests MAL-1 expression induces changes which go toward a pseudo-mesenchymal identity or intermediate state which may be specific to our system. Among the repressed genes were also components involved in the formation of the basal membrane which are typically synthesized by epithelial cells, such as Laminin LAMA1 or Type IV Collagen COL4A4. This structure specific to epithelial

tissues separates epithelial cells from the conjunctive extracellular matrix which is considered a more mesenchymal tissue. Although it can be maintained around epithelial tumors, it is often disturbed where its loss correlates with irregularities in tumor borders leading to the invasion of the surrounding stroma and metastasis (Kelley et al., 2014; Rodriguez-Boulan and Macara, 2014; Sakr et al., 1987).

As I investigated differential expression in a wider manner, I found two large chromosomic domains of interest: a locus starting from cytobands chr15q25 to the end of chromosome 15 which contains 17 repressed genes, and another from the start of chromosome 6 to cytoband chr6p22 which contains 16 upregulated genes. The latter was particularly interesting as it is where the endogenous MAL-1 is located in the genome and we could imagine its expression being involved in the upregulation of a wider downstream locus, up to 30 Mb long. This has been described for super-enhancers which are regulatory elements that can activate transcription from very long distances. They are typically transcribed into highly unstable lncRNAs which have been suggested to act through transcription itself and rarely as stand-alone RNA molecules (Sengupta and George, 2017). In the case of MAL-1, it seemingly can have a similar effect in *trans*, as a stand-alone transcript and outside of its *cis* genomic context. Indeed, the lentiviral transduction I did leads to random genomic integration and I worked on a population of cells without specific clonal selection: the Epi_MAL-1 cell line is thus a population of Epi cells with randomly distributed MAL-1 in its genome.
So far, a link between chromosome 6p22 amplification and EMT has been established in bladder cancer but was mainly attributed to the localization of the gene encoding stemness and EMT-inducing transcription factor SOX4 in the locus (Bellmunt, 2018). However, SOX4 is not upregulated upon MAL-1 expression. Interestingly, a similar observation was done for the EMT-associated lncRNA PVT1 expressed from the chr8q24 locus which is amplified in many types of cancer. This locus contains the well-known oncogene c-MYC as well as PVT1. Although most of the oncogenic properties of the locus are attributed to c-MYC, it has actually been shown that PVT1 expression is required for high MYC protein levels in 8q24-amplified cancer cells (Tseng et al., 2014). This shows that a specific cancer-

associated region may actually contain more than one oncogene, underlining the role of lncRNAs in complex molecular regulations.
Altogether, our data hints to MAL-1 being involved in the repression of epithelial features in our system, and potentially associated to cancer progression. Now, new experiments need to be done to understand better the function of MAL-1. First we will need to characterize MAL-1 in other EMT systems, notably its expression upon the canonical TGFβ induction of EMT. Considering the MAL-1 locus was found to be upregulated in both breast and kidney tumors, this should also be done in a variety of cancer cell lines as well as tumor samples to assess whether its expression can be specifically linked to tumor development.

As we did for HOTAIR, I also performed an RNA-folding structural analysis of MAL-1 and found 3 distinct structural domains (figure C6-1). We aim to clone the three domains into lentiviral vectors as truncated versions of MAL-1 and investigate their role in EMT regulation, this will be done by our new PhD student Rocco Cipolla.

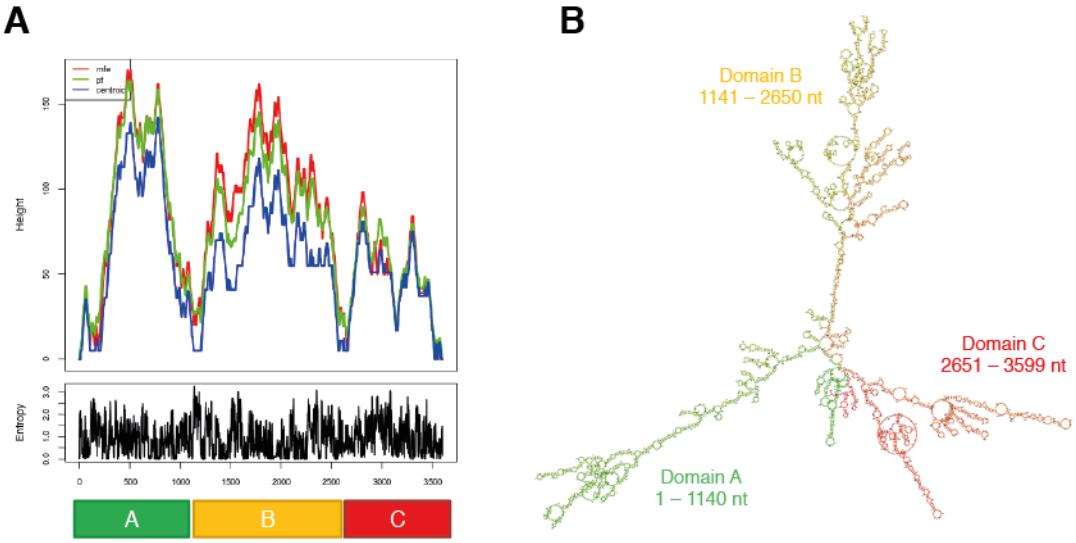


Figure C6-1. Structural analysis of the MAL-1. The analysis was done using Mfold web server (Zuker, 2003). **(A)** Mountain plot measuring the distance between adjacent bases along the transcript coordinate. **(B)** MAL-1 theoretical structure and assigned domains A, B and C.

As most lncRNAs act through their interaction with proteins we should investigate the protein partners with which MAL-1 may act. To do so, several techniques exist such as RNA pulldown-based methods followed by either Western Blot or mass-spectrometry for a proteome-wide analysis. Considering our results on MAL-1, it

could probably interact with transcription factors, epigenetic modifiers or even proteins involved in the formation of nuclear structures. However, performing these experiments on subcellular fractions on the cell could help us understand the potential role of MAL-1 in both the nucleus and cytoplasm where it could associate with different protein partners.

2.3. lncRNAs as regulators of epithelial plasticity

In this work, I studied the role of two lncRNAs found in mesenchymal cells and what phenotypic changes they may induce when expressed in epithelial cells. Although HOTAIR has been said to promote EMT, both HOTAIR and MAL-1 only induced partial EMT in our system, mostly repressing epithelial traits without inducing a mesenchymal phenotype *per se*.

This could be explained by the specific system which may have been strongly set in its identity after immortalization. However, there are many examples in the literature where a lncRNA claimed to be promoting EMT had similar effects to HOTAIR and MAL-1 in the present manuscript. For example, the MEG3 lncRNA was shown to promote EMT in lung cancer as Terashima and colleagues modulated its expression in parallel with TGFβ induction of EMT (Terashima et al., 2017). They showed that MEG3 knock-down by itself did not impact cell identity, but that MEG3 was essential for TGFβ induction of EMT in epithelial cells. However, when overexpressed by itself, MEG3 only repressed epithelial markers such as E-Cadherin without inducing the expression of mesenchymal markers such as Vimentin or Fibronectin, nor any morphological changes. Only when coupled with TGFβ treatment, MEG3 overexpression induced a stronger EMT phenotype compared to TGFβ alone.This goes to show that MEG3, and other lncRNAs are often not sufficient to induce a full EMT without other canonical factors yet still are key players of the EMT phenotype. On the opposit, MEG3 was also shown to actually inhibit EMT in gastric cancer cells as its overexpression correlated with a decrease in cell migration and the repression of mesenchymal markers such as metalloproteases or Snail (Xu et al., 2018). Although it is not mentioned in Xu and colleagues' article, one could see this inhibition of EMT as MEG3 actually promoting its counter process Mesenchymal-To-Epithelial Transition (MET).

In fact, another example which promotes both EMT and MET depending on the cellular context is the lncRNA H19. Indeed, this lncRNA has been linked to many cellular processes including cancer. In a very comprehensive review, Raveh and colleagues suggested H19 supports both EMT and MET to confer high plasticity to the cell, tightly linking it to every stage of tumorigenesis (Raveh et al., 2015).

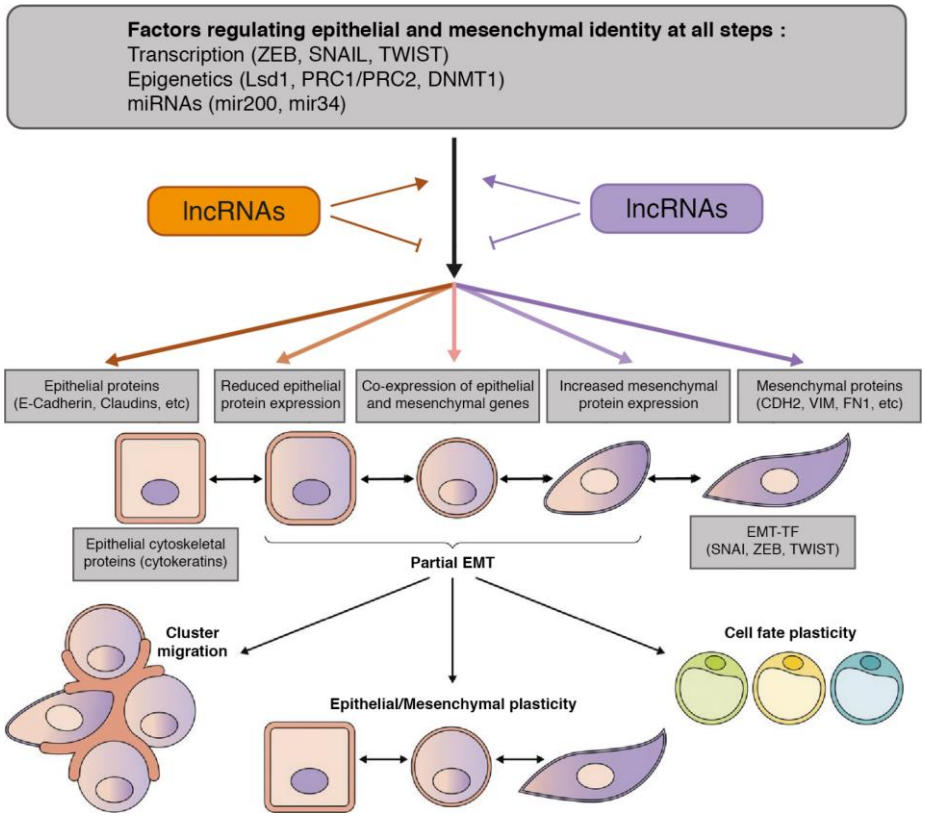


Figure C6-2. lncRNAs as the link between the major regulators of epithelial and mesenchymal identity to induce the hybrid EMT states. Adapted from (Aiello and Kang, 2019).

Altogether, the subtler phenotypes associated with some of these lncRNAs actually correlate with a paradigm shift in the field of the epithelial-to-mesenchymal transition (figure C6-2). Indeed, EMT and MET are often considered as mutually exclusive phenotypes but full EMT does not systematically happen and there are many intermediate states which contribute to cancer heterogeneity (Nieto et al., 2016; Polyak and Weinberg, 2009). It was recently shown that these hybrid states exist in tumors and display strong phenotypic differences as to cellular plasticity, stemness, invasiveness and metastatic potential (Pastushenko et al., 2018). These distinct hybrid states also have different epigenetic landscapes and gene expression

signatures. From both ends of the transition, EMT and MET were both suggested to be important for metastatic progression as mesenchymal cells detach from the primary tumor and then become epithelial again in order to reattach and form a secondary tumor (Banyard and Bielenberg, 2015). Considering this, there is no doubt hybrid states of EMT in tumors participate in tumor progression through heterogeneity, growth and invasion.
The new single-cell (sc)RNA-seq approaches which have been developed recently and used to characterize these hybrid states will no doubt be useful to further characterize the role of lncRNAs in tumor heterogeneity. Although these technologies are improving daily, some challenges still remain as lncRNAs are typically expressed on lower levels than protein coding genes and scRNA-seq approaches tend to have low sequencing coverage. In the upcoming years, the investigation of lncRNAs in tumoral subpopulations as biomarkers for diagnosis and prognosis, but also as effectors of tumorigenesis will certainly be a big focus in cancer research.

As lncRNAs have high specificity of expression, some of them can certainly be found in these intermediate states of EMT, which could explain the subtler changes in phenotype upon ectopic expression of lncRNAs such as HOTAIR, MEG3, H19 or MAL-1: beside promoting EMT or MET as a larger process, they could actually induce changes toward intermediate states, depending on the cellular context or actually providing it. Altogether, our and others work show lncRNAs represent an additional layer of regulation of EMT, as subtle regulators involved in the induction of intermediate EMT states through the coordination of major inducers such as TGFβ, EMT-TF families (SNAI, TWIST, ZEB) and epigenetic modifiers (Lsd1), thus potentially favoring tumor heterogeneity and cancer progression.

REFERENCES

Abdelmohsen, K., Panda, A., Kang, M.-J., Xu, J., Selimyan, R., Yoon, J.-H., Martindale, J.L., De, S., Wood, W.H., Becker, K.G., et al. (2013). Senescence-associated lncRNAs: senescence-associated long noncoding RNAs. Aging Cell *12*, 890–900.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185–2195.

Agarwal, N., and Ansari, A. (2016). Enhancement of Transcription by a Splicing-Competent Intron Is Dependent on Promoter Directionality. PLOS Genetics *12*, e1006047.

Aiello, N.M., and Kang, Y. (2019). Context-dependent EMT programs in cancer metastasis. Journal of Experimental Medicine *216*, 1016–1026.

Ameres, S.L., and Zamore, P.D. (2013). Diversifying microRNA sequence and function. Nature Reviews Molecular Cell Biology *14*, 475–488.

An, Y., Furber, K.L., and Ji, S. (2016). Pseudogenes regulate parental gene expression *via* ceRNA network. Journal of Cellular and Molecular Medicine.

Anandakumar, S., Vijayakumar, S., Centre for Advanced Study in Crystallography and Biophysics, University of Madras, Chennai 600005, Tamil Nadu, India, Arumugam, N., Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600 036, Tamil Nadu, India, Gromiha, M.M., and Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600 036, Tamil Nadu, India (2015). Mammalian Mitochondrial ncRNA Database. Bioinformation *11*, 512–514.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology *11*, R106.

Andrews, S.J., and Rothnagel, J.A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. Nature Reviews Genetics *15*, 193–204.

Antequera, F., and Bird, A. (1993). Number of CpG islands and genes in human and mouse. Proc. Natl. Acad. Sci. U.S.A. *90*, 11995–11999.

Ariel, F., Romero-Barrios, N., Jégu, T., Benhamed, M., and Crespi, M. (2015). Battles and hijacks: noncoding transcription in plants. Trends in Plant Science *20*, 362–371.

Arrial, R.T., Togawa, R.C., and Brigido, M. de M. (2009). Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. BMC Bioinformatics *10*, 239.

Ayupe, A.C., Tahira, A.C., Camargo, L., Beckedorff, F.C., Verjovski-Almeida, S., and Reis, E.M. (2015). Global analysis of biogenesis, stability and sub-cellular localization of lncRNAs mapping to intragenic regions of the human genome. RNA

Biology *12*, 877–892.

Bae, B.-I., Tietjen, I., Atabay, K.D., Evrony, G.D., Johnson, M.B., Asare, E., Wang, P.P., Murayama, A.Y., Im, K., Lisgo, S.N., et al. (2014). Evolutionarily dynamic alternative splicing of GPR56 regulates regional cerebral cortical patterning. Science *343*, 764–768.

Balk, B., Maicher, A., Dees, M., Klermund, J., Luke-Glaser, S., Bender, K., and Luke, B. (2013). Telomeric RNA-DNA hybrids affect telomere-length dynamics and senescence. Nature Structural & Molecular Biology *20*, 1199–1205.

Balk, B., Dees, M., Bender, K., and Luke, B. (2014). The differential processing of telomeres in response to increased telomeric transcription and RNA–DNA hybrid accumulation. RNA Biology *11*, 95–100.

Banfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W.E., Kundaje, A., Gunawardena, H.P., Yu, Y., Xie, L., et al. (2012). Long noncoding RNAs are rarely translated in two human cell lines. Genome Research *22*, 1646–1657.

Banyard, J., and Bielenberg, D.R. (2015). The Role of EMT and MET in Cancer Dissemination . Connect Tissue Res *56*, 403–413.

Barlow, D.P., Stöger, R., Herrmann, B.G., Saito, K., and Schweifer, N. (1991). The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. Nature *349*, 84–87.

Bartolomei, M.S., Zemel, S., and Tilghman, S.M. (1991). Parental imprinting of the mouse H19 gene. Nature *351*, 153–155.

Battistelli, C., Cicchini, C., Santangelo, L., Tramontano, A., Grassi, L., Gonzalez, F.J., de Nonno, V., Grassi, G., Amicone, L., and Tripodi, M. (2016). The Snail repressor recruits EZH2 to specific genomic sites through the enrollment of the lncRNA HOTAIR in epithelial-to-mesenchymal transition. Oncogene.

Bejerano, G. (2004). Ultraconserved Elements in the Human Genome. Science *304*, 1321–1325.

Bellmunt, J. (2018). Stem-Like Signature Predicting Disease Progression in Early Stage Bladder Cancer. The Role of E2F3 and SOX4. Biomedicines *6*.

Berezovska, O.P., Glinskii, A.B., Yang, Z., Li, X.-M., Hoffman, R.M., and Glinsky, G.V. (2006). Essential role for activation of the Polycomb group (PcG) protein chromatin silencing pathway in metastatic prostate cancer. Cell Cycle *5*, 1886–1901.

Berk, A.J. (2016). Discovery of RNA splicing and genes in pieces. Proceedings of the National Academy of Sciences *113*, 801–805.

Bernhardt, H.S. (2012). The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)a. Biology Direct *7*, 23.

Berretta, J., and Morillon, A. (2009). Pervasive transcription constitutes a new level of eukaryotic genome regulation. EMBO Reports *10*, 973–982.

Bierhoff, H., Dammert, M.A., Brocks, D., Dambacher, S., Schotta, G., and Grummt, I. (2014). Quiescence-Induced LncRNAs Trigger H4K20 Trimethylation and Transcriptional Silencing. Molecular Cell *54*, 675–682.

Bird, C.P., Stranger, B.E., Liu, M., Thomas, D.J., Ingle, C.E., Beazley, C., Miller, W., Hurles, M.E., and Dermitzakis, E.T. (2007). Fast-evolving noncoding sequences in the human genome. Genome Biol. *8*, R118.

Blower, M.D. (2016). Centromeric Transcription Regulates Aurora-B Localization and Activation. Cell Reports *15*, 1624–1633.

Bogu, G.K., Vizán, P., Stanton, L.W., Beato, M., Di Croce, L., and Marti-Renom, M.A. (2016). Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse. Molecular and Cellular Biology *36*, 809–819.

Brannan, C.I., Dees, E.C., Ingram, R.S., and Tilghman, S.M. (1990). The product of the H19 gene may function as an RNA. Molecular and Cellular Biology *10*, 28–36.

Broadbent, K.M., Broadbent, J.C., Ribacke, U., Wirth, D., Rinn, J.L., and Sabeti, P.C. (2015). Strand-specific RNA sequencing in Plasmodium falciparum malaria identifies developmentally regulated long non-coding RNA and circular RNA. BMC Genomics *16*.

Burzio, V.A., Villota, C., Villegas, J., Landerer, E., Boccardo, E., Villa, L.L., Martinez, R., Lopez, C., Gaete, F., Toro, V., et al. (2009). Expression of a family of noncoding mitochondrial RNAs distinguishes normal from cancer cells. Proceedings of the National Academy of Sciences *106*, 9430–9434.

Butcher, S.E. (2009). The spliceosome as ribozyme hypothesis takes a second step. Proceedings of the National Academy of Sciences *106*, 12211–12212.

Cabili, M.N., Dunagin, M.C., McClanahan, P.D., Biaesch, A., Padovan-Merhar, O., Regev, A., Rinn, J.L., and Raj, A. (2015). Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. Genome Biol. *16*, 20.

Cajigas, I., Leib, D.E., Cochrane, J., Luo, H., Swyter, K.R., Chen, S., Clark, B.S., Thompson, J., Yates, J.R., Kingston, R.E., et al. (2015). *Evf2* lncRNA/BRG1/DLX1 interactions reveal RNA-dependent inhibition of chromatin remodeling. Development *142*, 2641–2652.

Campeau, E., Ruhl, V.E., Rodier, F., Smith, C.L., Rahmberg, B.L., Fuss, J.O., Campisi, J., Yaswen, P., Cooper, P.K., and Kaufman, P.D. (2009). A versatile viral system for expression and depletion of proteins in mammalian cells. PLoS ONE *4*, e6529.

Cano, A., Pérez-Moreno, M.A., Rodrigo, I., Locascio, A., Blanco, M.J., del Barrio, M.G., Portillo, F., and Nieto, M.A. (2000). The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression. Nat. Cell Biol. *2*, 76–83.

Carlevaro-Fita, J., Rahim, A., Guigo, R., Vardy, L., and Johnson, R. (2015). Widespread localisation of long noncoding RNAs to ribosomes: Distinguishing

features and evidence for regulatory roles.

Carlile, M., Swan, D., Jackson, K., Preston-Fayers, K., Ballester, B., Flicek, P., and Werner, A. (2009). Strand selective generation of endo-siRNAs from the Na/phosphate transporter gene Slc34a1 in murine tissues. Nucleic Acids Research *37*, 2274–2282.

Carnesecchi, J., Cerutti, C., Vanacker, J.-M., and Forcet, C. (2017). ERRα protein is stabilized by LSD1 in a demethylation-independent manner. PLOS ONE *12*, e0188871.

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., Pesce, E., Ferrer, I., Collavin, L., Santoro, C., et al. (2012). Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature *491*, 454–457.

Castel, S.E., and Martienssen, R.A. (2013). RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. Nature Reviews Genetics *14*, 100–112.

Castro-Vega, L.J., Jouravleva, K., Liu, W.-Y., Martinez, C., Gestraud, P., Hupé, P., Servant, N., Albaud, B., Gentien, D., Gad, S., et al. (2013a). Telomere crisis in kidney epithelial cells promotes the acquisition of a microRNA signature retrieved in aggressive renal cell carcinomas. Carcinogenesis *34*, 1173–1180.

Cech, T.R. (2000). STRUCTURAL BIOLOGY: Enhanced: The Ribosome Is a Ribozyme. Science *289*, 878–879.

Chaffer, C.L., and Weinberg, R.A. (2011). A perspective on cancer cell metastasis. Science *331*, 1559–1564.

Chan, F.L., Marshall, O.J., Saffery, R., Won Kim, B., Earle, E., Choo, K.H.A., and Wong, L.H. (2012). Active transcription and essential role of RNA polymerase II at the centromere during mitosis. Proceedings of the National Academy of Sciences *109*, 1979–1984.

Chen, L.-L. (2016). Linking Long Noncoding RNA Localization and Function. Trends in Biochemical Sciences *41*, 761–772.

Chen, S.-X., Yin, J.-F., Lin, B.-C., Su, H.-F., Zheng, Z., Xie, C.-Y., and Fei, Z.-H. (2016). Upregulated expression of long noncoding RNA SNHG15 promotes cell proliferation and invasion through regulates MMP2/MMP9 in patients with GC. Tumour Biol. *37*, 6801–6812.

Cheng, D., Bao, C., Zhang, X., Lin, X., Huang, H., and Zhao, L. (2018). LncRNA PRNCR1 interacts with HEY2 to abolish miR-448-mediated growth inhibition in non-small cell lung cancer. Biomed. Pharmacother. *107*, 1540–1547.

Chinwalla, A.T., Cook, L.L., Delehaunty, K.D., Fewell, G.A., Fulton, L.A., Fulton, R.S., Graves, T.A., Hillier, L.W., Mardis, E.R., McPherson, J.D., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–562.

Choudhry, H., Harris, A.L., and McIntyre, A. (2016). The tumour hypoxia induced non-coding transcriptome. Molecular Aspects of Medicine *47–48*, 35–53.

Chu, C., and Chang, H.Y. (2018). ChIRP-MS: RNA-Directed Proteomic Discovery. Methods Mol. Biol. *1861*, 37–45.

Chu, C., Qu, K., Zhong, F.L., Artandi, S.E., and Chang, H.Y. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Mol. Cell *44*, 667–678.

Clark, M.B., Johnston, R.L., Inostroza-Ponta, M., Fox, A.H., Fortini, E., Moscato, P., Dinger, M.E., and Mattick, J.S. (2012). Genome-wide analysis of long noncoding RNA stability. Genome Research *22*, 885–898.

Clark, M.B., Choudhary, A., Smith, M.A., Taft, R.J., and Mattick, J.S. (2013). The dark matter rises: the expanding world of regulatory RNAs. Essays Biochem. *54*, 1–16.

Cobb, M. (2015). Who discovered messenger RNA? Current Biology *25*, R526–R532.

Cox, D.B.T., Gootenberg, J.S., Abudayyeh, O.O., Franklin, B., Kellner, M.J., Joung, J., and Zhang, F. (2017). RNA editing with CRISPR-Cas13. Science *358*, 1019–1027.

Crick, F.H.C. (1968). The origin of the genetic code. Journal of Molecular Biology *38*, 367–379.

Crollius, H.R. (2000). Characterization and Repeat Analysis of the Compact Genome of the Freshwater Pufferfish Tetraodon nigroviridis. Genome Research *10*, 939–949.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. (2006). A high-resolution map of transcription in the yeast genome. Proceedings of the National Academy of Sciences *103*, 5320–5325.

Davidovich, C., and Cech, T.R. (2015). The recruitment of chromatin modifiers by long noncoding RNAs: lessons from PRC2. RNA *21*, 2007–2022.

Davis, F.M., Azimi, I., Faville, R.A., Peters, A.A., Jalink, K., Putney, J.W., Goodhill, G.J., Thompson, E.W., Roberts-Thomson, S.J., and Monteith, G.R. (2014). Induction of epithelial-mesenchymal transition (EMT) in breast cancer cells is calcium signal dependent. Oncogene *33*, 2307–2316.

Deng, Q., Sun, H., He, B., Pan, Y., Gao, T., Chen, J., Ying, H., Liu, X., Wang, F., Xu, Y., et al. (2014). Prognostic value of long non-coding RNA HOTAIR in various cancers. PLoS ONE *9*, e110059.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012a). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Research *22*, 1775–1789.

Descrimes, M., Zouari, Y.B., Wery, M., Legendre, R., Gautheret, D., and Morillon, A.

(2015). VING: a software for visualization of deep sequencing signals. BMC Res Notes *8*.

Dey, B.K., Pfeifer, K., and Dutta, A. (2014). The H19 long noncoding RNA gives rise to microRNAs miR-675-3p and miR-675-5p to promote skeletal muscle differentiation and regeneration. Genes & Development *28*, 491–501.

Dhir, A., Dhir, S., Proudfoot, N.J., and Jopling, C.L. (2015). Microprocessor mediates transcriptional termination of long noncoding RNA transcripts hosting microRNAs. Nature Structural & Molecular Biology *22*, 319–327.

Di Benedetto, M., Bièche, I., Deshayes, F., Vacher, S., Nouet, S., Collura, V., Seitz, I., Louis, S., Pineau, P., Amsellem-Ouazana, D., et al. (2006). Structural organization and expression of human MTUS1, a candidate 8p22 tumor suppressor gene encoding a family of angiotensin II AT2 receptor-interacting proteins, ATIP. Gene *380*, 127–136.

Diederichs, S. (2014). The four dimensions of noncoding RNA conservation. Trends Genet. *30*, 121–123.

Ding, C., Cheng, S., Yang, Z., Lv, Z., Xiao, H., Du, C., Peng, C., Xie, H., Zhou, L., Wu, J., et al. (2014). Long non-coding RNA HOTAIR promotes cell migration and invasion via down-regulation of RNA binding motif protein 38 in hepatocellular carcinoma cells. Int J Mol Sci *15*, 4060–4076.

Dinger, M.E., Amaral, P.P., Mercer, T.R., and Mattick, J.S. (2009). Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. Briefings in Functional Genomics and Proteomics *8*, 407–423.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Doan, R.N., Bae, B.-I., Cubelos, B., Chang, C., Hossain, A.A., Al-Saad, S., Mukaddes, N.M., Oner, O., Al-Saffar, M., Balkhy, S., et al. (2016). Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. Cell *167*, 341-354.e12.

Dong, L., and Hui, L. (2016). HOTAIR Promotes Proliferation, Migration, and Invasion of Ovarian Cancer SKOV3 Cells Through Regulating PIK3R3. Med. Sci. Monit. *22*, 325–331.

Dugimont, T., Montpellier, C., Adriaenssens, E., Lottin, S., Dumont, L., Iotsova, V., Lagrou, C., Stéhelin, D., Coll, J., and Curgy, J.J. (1998). The H19 TATA-less promoter is efficiently repressed by wild-type tumor suppressor gene product p53. Oncogene *16*, 2395–2401.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., et al. (1999). The DNA sequence of human chromosome 22. Nature *402*, 489–495.

Duret, L. (2006). The Xist RNA Gene Evolved in Eutherians by Pseudogenization of

a Protein-Coding Gene. Science *312*, 1653–1655.

Durruthy-Durruthy, J., Sebastiano, V., Wossidlo, M., Cepeda, D., Cui, J., Grow, E.J., Davila, J., Mall, M., Wong, W.H., Wysocka, J., et al. (2015). The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. Nature Genetics *48*, 44–52.

ENCODE Project Consortium (2012a). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Enuka, Y., Lauriola, M., Feldman, M.E., Sas-Chen, A., Ulitsky, I., and Yarden, Y. (2016). Circular RNAs are long-lived and display only minimal early alterations in response to a growth factor. Nucleic Acids Research *44*, 1370–1383.

Espinoza, C.A., Goodrich, J.A., and Kugel, J.F. (2007). Characterization of the structure, function, and mechanism of B2 RNA, an ncRNA repressor of RNA polymerase II transcription. RNA *13*, 583–596.

Expósito-Villén, A., E Aránega, A., and Franco, D. (2018). Functional Role of Non-Coding RNAs during Epithelial-To-Mesenchymal Transition. Noncoding RNA *4*.

Faghihi, M.A., Modarresi, F., Khalil, A.M., Wood, D.E., Sahagan, B.G., Morgan, T.E., Finch, C.E., St. Laurent III, G., Kenny, P.J., and Wahlestedt, C. (2008). Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase. Nature Medicine *14*, 723–730.

Fassan, M., Dall'Olmo, L., Galasso, M., Braconi, C., Pizzi, M., Realdon, S., Volinia, S., Valeri, N., Gasparini, P., Baffa, R., et al. (2014). Transcribed ultraconserved noncoding RNAs (T-UCR) are involved in Barrett's esophagus carcinogenesis. Oncotarget *5*, 7162–7171.

Feng, J. (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*, 1470–1484.

Feng, J., Xu, G., Liu, J., Zhang, N., Li, L., Ji, J., Zhang, J., Zhang, L., Wang, G., Wang, X., et al. (2016). Phosphorylation of LSD1 at Ser112 is crucial for its function in induction of EMT and metastasis in breast cancer. Breast Cancer Res. Treat. *159*, 443–456.

Ferdin, J., Nishida, N., Wu, X., Nicoloso, M.S., Shah, M.Y., Devlin, C., Ling, H., Shimizu, M., Kumar, K., Cortez, M.A., et al. (2013). HINCUTs in cancer: hypoxia-induced noncoding ultraconserved transcripts. Cell Death and Differentiation *20*, 1675–1687.

Ferrari-Amorotti, G., Fragliasso, V., Esteki, R., Prudente, Z., Soliera, A.R., Cattelani,

S., Manzotti, G., Grisendi, G., Dominici, M., Pieraccioli, M., et al. (2013). Inhibiting interactions of lysine demethylase LSD1 with snail/slug blocks cancer cell invasion. Cancer Res. *73*, 235–245.

Feuerhahn, S., Iglesias, N., Panza, A., Porro, A., and Lingner, J. (2010). TERRA biogenesis, turnover and implications for function. FEBS Letters *584*, 3812–3818.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., and Merrick, J.M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496–512.

Flynn, R.A., Almada, A.E., Zamudio, J.R., and Sharp, P.A. (2011). Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. Proceedings of the National Academy of Sciences *108*, 10460–10465.

Fort, A., Hashimoto, K., Yamada, D., Salimullah, M., Keya, C.A., Saxena, A., Bonetti, A., Voineagu, I., Bertin, N., Kratz, A., et al. (2014). Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nature Genetics *46*, 558–566.

Gagnon, K.T., Li, L., Janowski, B.A., and Corey, D.R. (2014). Analysis of nuclear RNA interference in human cells by subcellular fractionation and Argonaute loading. Nat Protoc *9*, 2045–2060.

Ganesh, S., and Svoboda, P. (2016). Retrotransposon-associated long non-coding RNAs in mice and men. Pflügers Archiv - European Journal of Physiology *468*, 1049–1060.

Gebäck, T., Schulz, M.M.P., Koumoutsakos, P., and Detmar, M. (2009). TScratch: a novel and simple software tool for automated analysis of monolayer wound healing assays. BioTechniques *46*, 265–274.

Geng, Y.J., Xie, S.L., Li, Q., Ma, J., and Wang, G.Y. (2011). Large intervening non-coding RNA HOTAIR is associated with hepatocellular carcinoma progression. J. Int. Med. Res. *39*, 2119–2128.

Giannakakis, A., Zhang, J., Jenjaroenpun, P., Nama, S., Zainolabidin, N., Aau, M.Y., Yarmishyn, A.A., Vaz, C., Ivshina, A.V., Grinchuk, O.V., et al. (2015). Contrasting expression patterns of coding and noncoding parts of the human genome upon oxidative stress. Scientific Reports *5*, 9737.

Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., et al. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell *159*, 647–661.

Giovarelli, M., Bucci, G., Ramos, A., Bordo, D., Wilusz, C.J., Chen, C.-Y., Puppo, M., Briata, P., and Gherzi, R. (2014). H19 long noncoding RNA controls the mRNA decay promoting function of KSRP. Proceedings of the National Academy of Sciences *111*, E5023–E5028.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. Science *274*, 546, 563–567.

Gomez, J.A., Wapinski, O.L., Yang, Y.W., Bureau, J.-F., Gopinath, S., Monack, D.M., Chang, H.Y., Brahic, M., and Kirkegaard, K. (2013). The NeST Long ncRNA Controls Microbial Susceptibility and Epigenetic Activation of the Interferon-γ Locus. Cell *152*, 743–754.

Gonzalez, D.M., and Medici, D. (2014). Signaling mechanisms of the epithelial-mesenchymal transition. Sci Signal *7*, re8.

Gonzalez, I., Munita, R., Agirre, E., Dittmer, T.A., Gysling, K., Misteli, T., and Luco, R.F. (2015). A lncRNA regulates alternative splicing via establishment of a splicing-specific chromatin signature. Nature Structural & Molecular Biology.

Goodman, A.J., Daugharthy, E.R., and Kim, J. (2013). Pervasive Antisense Transcription Is Evolutionarily Conserved in Budding Yeast. Molecular Biology and Evolution *30*, 409–421.

Goossens, S., Peirs, S., Van Loocke, W., Wang, J., Takawy, M., Matthijssens, F., Sonderegger, S.E., Haigh, K., Nguyen, T., Vandamme, N., et al. (2017). Oncogenic ZEB2 activation drives sensitivity toward KDM1A inhibition in T-cell acute lymphoblastic leukemia. Blood *129*, 981–990.

Grandér, D., and Johnsson, P. (2015). Pseudogene-Expressed RNAs: Emerging Roles in Gene Regulation and Disease. In Long Non-Coding RNAs in Human Disease, K.V. Morris, ed. (Cham: Springer International Publishing), pp. 111–126.

Greenwood, J., and Cooper, J.P. (2012). Non-coding telomeric and subtelomeric transcripts are differentially regulated by telomeric and heterochromatin assembly factors in fission yeast. Nucleic Acids Research *40*, 2956–2963.

Griffiths, A.J., Miller, J.H., Suzuki, D.T., Lewontin, R.C., and Gelbart, W.M. (2000). Transcription.

Gröger, C.J., Grubinger, M., Waldhör, T., Vierlinger, K., and Mikulits, W. (2012a). Meta-analysis of gene expression signatures defining the epithelial to mesenchymal transition during cancer progression. PLoS ONE *7*, e51136.

Guenzl, P.M., and Barlow, D.P. (2012). Macro lncRNAs: A new layer of *cis* -regulatory information in the mammalian genome. RNA Biology *9*, 731–741.

Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell *35*, 849–857.

Gugnoni, M., and Ciarrocchi, A. (2019). Long Noncoding RNA and Epithelial Mesenchymal Transition in Cancer. Int J Mol Sci *20*.

Guo, H., Zhao, L., Shi, B., Bao, J., Zheng, D., Zhou, B., and Shi, J. (2018). GALNT5

uaRNA promotes gastric cancer progression through its interaction with HSP90. Oncogene *37*, 4505–4517.

Guo, X., Gao, L., Wang, Y., Chiu, D.K.Y., Wang, T., and Deng, Y. (2015). Advances in long noncoding RNAs: identification, structure prediction and function annotation. Briefings in Functional Genomics.

Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.-C., Hung, T., Argani, P., Rinn, J.L., et al. (2010a). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature *464*, 1071–1076.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009a). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

Ha, H., Song, J., Wang, S., Kapusta, A., Feschotte, C., Chen, K.C., and Xing, J. (2014). A comprehensive analysis of piRNAs from adult human testis and their relationship with genes and mobile elements. BMC Genomics *15*, 545.

Hacisuleyman, E., Shukla, C.J., Weiner, C.L., and Rinn, J.L. (2016). Function and evolution of local repeats in the Firre locus. Nature Communications *7*, 11021.

Hadjiargyrou, M., and Delihas, N. (2013). The intertwining of transposable elements and non-coding RNAs. Int J Mol Sci *14*, 13307–13328.

Hamazaki, N., Uesaka, M., Nakashima, K., Agata, K., and Imamura, T. (2015). Gene activation-associated long noncoding RNAs function in mouse preimplantation development. Development *142*, 910–920.

Han, P., and Chang, C.-P. (2015). Long non-coding RNA and chromatin remodeling. RNA Biology *12*, 1094–1098.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*, 646–674.

Hansen, T.B., Jensen, T.I., Clausen, B.H., Bramsen, J.B., Finsen, B., Damgaard, C.K., and Kjems, J. (2013). Natural RNA circles function as efficient microRNA sponges. Nature *495*, 384–388.

Harris, W.J., Huang, X., Lynch, J.T., Spencer, G.J., Hitchin, J.R., Li, Y., Ciceri, F., Blaser, J.G., Greystoke, B.F., Jordan, A.M., et al. (2012). The histone demethylase KDM1A sustains the oncogenic potential of MLL-AF9 leukemia stem cells. Cancer Cell *21*, 473–487.

He, L., and Hannon, G.J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. Nature Reviews Genetics *5*, 522–531.

van Heesch, S., van Iterson, M., Jacobi, J., Boymans, S., Essers, P.B., de Bruijn, E., Hao, W., MacInnes, A.W., Cuppen, E., and Simonis, M. (2014). Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes.

Genome Biology *15*, R6.

Heinen, T.J.A.J., Staubach, F., Häming, D., and Tautz, D. (2009). Emergence of a New Gene from an Intergenic Region. Current Biology *19*, 1527–1531.

Hendrickson, D.G., Kelley, D.R., Tenen, D., Bernstein, B., and Rinn, J.L. (2016). Widespread RNA binding by chromatin-associated proteins. Genome Biol. *17*, 28.

Hezroni, H., Koppstein, D., Schwartz, M.G., Avrutin, A., Bartel, D.P., and Ulitsky, I. (2015). Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. Cell Reports *11*, 1110–1122.

Hino, S., Kohrogi, K., and Nakao, M. (2016). Histone demethylase LSD1 controls the phenotypic plasticity of cancer cells. Cancer Sci. *107*, 1187–1192.

Hirata, H., Hinoda, Y., Shahryari, V., Deng, G., Nakajima, K., Tabatabai, Z.L., Ishii, N., and Dahiya, R. (2015). Long Noncoding RNA MALAT1 Promotes Aggressive Renal Cell Carcinoma through Ezh2 and Interacts with miR-205. Cancer Res. *75*, 1322–1331.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

Hoffmann, M., Dehn, J., Droop, J., Niegisch, G., Niedworok, C., Szarvas, T., and Schulz, W. (2015). Truncated Isoforms of lncRNA ANRIL Are Overexpressed in Bladder Cancer, But Do Not Contribute to Repression of INK4 Tumor Suppressors. Non-Coding RNA *1*, 266–284.

Horlbeck, M.A., Gilbert, L.A., Villalta, J.E., Adamson, B., Pak, R.A., Chen, Y., Fields, A.P., Park, C.Y., Corn, J.E., Kampmann, M., et al. (2016). Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. Elife *5*.

Housman, G., and Ulitsky, I. (2016). Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms *1859*, 31–40.

Hu, H., He, L., and Khaitovich, P. (2014a). Deep sequencing reveals a novel class of bidirectional promoters associated with neuronal genes. BMC Genomics *15*, 457.

Hu, P., Yang, J., Hou, Y., Zhang, H., Zeng, Z., Zhao, L., Yu, T., Tang, X., Tu, G., Cui, X., et al. (2014b). LncRNA expression signatures of twist-induced epithelial-to-mesenchymal transition in MCF10A cells. Cell. Signal. *26*, 83–93.

Huang, H.-H., Chen, F.-Y., Chou, W.-C., Hou, H.-A., Ko, B.-S., Lin, C.-T., Tang, J.-L., Li, C.-C., Yao, M., Tsay, W., et al. (2019). Long non-coding RNA HOXB-AS3 promotes myeloid cell proliferation and its higher expression is an adverse prognostic marker in patients with acute myeloid leukemia and myelodysplastic syndrome. BMC Cancer *19*, 617.

Huang, J., Sengupta, R., Espejo, A.B., Lee, M.G., Dorsey, J.A., Richter, M., Opravil, S., Shiekhattar, R., Bedford, M.T., Jenuwein, T., et al. (2007). p53 is regulated by the lysine demethylase LSD1. Nature *449*, 105–108.

Huang, J.-Z., Chen, M., Chen, D., Gao, X.-C., Zhu, S., Huang, H., Hu, M., Zhu, H., and Yan, G.-R. (2017). A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. Mol. Cell *68*, 171-184.e6.

Huarte, M., and Marín-Béjar, O. (2015). Long noncoding RNAs: from identification to functions and mechanisms. Advances in Genomics and Genetics 257.

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M.J., Kenzelmann-Broz, D., Khalil, A.M., Zuk, O., Amit, I., Rabani, M., et al. (2010). A Large Intergenic Noncoding RNA Induced by p53 Mediates Global Gene Repression in the p53 Response. Cell *142*, 409–419.

Human Genome Sequencing Consortium, I. (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Hung, T., Wang, Y., Lin, M.F., Koegel, A.K., Kotake, Y., Grant, G.D., Horlings, H.M., Shah, N., Umbricht, C., Wang, P., et al. (2011). Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nature Genetics *43*, 621–629.

Hyun, K.-A., Koo, G.-B., Han, H., Sohn, J., Choi, W., Kim, S.-I., Jung, H.-I., and Kim, Y.-S. (2016). Epithelial-to-mesenchymal transition leads to loss of EpCAM and different physical properties in circulating tumor cells from metastatic breast cancer. Oncotarget 7, 24677–24687.

Indrieri, A., Grimaldi, C., Zucchelli, S., Tammaro, R., Gustincich, S., and Franco, B. (2016). Synthetic long non-coding RNAs [SINEUPs] rescue defective gene expression in vivo. Scientific Reports *6*, 27315.

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet. *47*, 199–208.

Jarroux, J., Morillon, A., and Pinskaya, M. (2017). History, Discovery, and Classification of lncRNAs. Adv. Exp. Med. Biol. *1008*, 1–46.

Jarroux, J., Bertrand, C., Marc, G., Foretek, D., Saci, Z., Vallejo, J.A.L., Pinskaya, M., and Morillon, A. (2019). HOTAIR promotes an epithelial-to-mesenchymal transition through relocation of the histone demethylase Lsd1. BioRxiv 724948.

Ji, Z., Song, R., Regev, A., and Struhl, K. (2015). Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. ELife *4*.

Jiang, Li, Zhao, and Lu (2016). Identifying and functionally characterizing tissue-specific and ubiquitously expressed human lncRNAs. Oncotarget.

Johnson, R., and Guigo, R. (2014). The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. RNA *20*, 959–976.

Jolly, M.K., Boareto, M., Huang, B., Jia, D., Lu, M., Ben-Jacob, E., Onuchic, J.N., and Levine, H. (2015). Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. Front Oncol *5*, 155.

Jolly, M.K., Ware, K.E., Gilja, S., Somarelli, J.A., and Levine, H. (2017). EMT and MET: necessary or permissive for metastasis? Mol Oncol *11*, 755–769.

Joung, J., Engreitz, J.M., Konermann, S., Abudayyeh, O.O., Verdine, V.K., Aguet, F., Gootenberg, J.S., Sanjana, N.E., Wright, J.B., Fulco, C.P., et al. (2017a). Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. Nature *548*, 343–346.

Joung, J., Konermann, S., Gootenberg, J.S., Abudayyeh, O.O., Platt, R.J., Brigham, M.D., Sanjana, N.E., and Zhang, F. (2017b). Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. Nat Protoc *12*, 828–863.

Kalluri, R., and Weinberg, R.A. (2009). The basics of epithelial-mesenchymal transition. J. Clin. Invest. *119*, 1420–1428.

Kambara, H., Gunawardane, L., Zebrowski, E., Kostadinova, L., Jobava, R., Krokowski, D., Hatzoglou, M., Anthony, D.D., and Valadkhan, S. (2015). Regulation of Interferon-Stimulated Gene BST2 by a lncRNA Transcribed from a Shared Bidirectional Promoter. Frontiers in Immunology *5*.

Kaneko, S., Son, J., Shen, S.S., Reinberg, D., and Bonasio, R. (2013). PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. Nat. Struct. Mol. Biol. *20*, 1258–1264.

Kaneko, S., Son, J., Bonasio, R., Shen, S.S., and Reinberg, D. (2014). Nascent RNA interaction keeps PRC2 activity poised and in check. Genes & Development *28*, 1983–1988.

Kapranov, P. (2002). Large-Scale Transcriptional Activity in Chromosomes 21 and 22. Science *296*, 916–919.

Kapranov, P. (2005). Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. Genome Research *15*, 987–997.

Kapranov, P., Willingham, A.T., and Gingeras, T.R. (2007). Genome-wide transcription and the implications for genomic organization. Nature Reviews Genetics *8*, 413–423.

Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C. (2013). Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. PLoS Genetics *9*, e1003470.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., et al. (2005). Antisense transcription in the mammalian transcriptome. Science *309*, 1564–1566.

Keller, L., Werner, S., and Pantel, K. (2019). Biology and clinical relevance of EpCAM. Cell Stress *3*, 165–180.

Kelley, L.C., Lohmer, L.L., Hagedorn, E.J., and Sherwood, D.R. (2014). Traversing the basement membrane in vivo: a diversity of strategies. J. Cell Biol. *204*, 291–302.

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A., 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 *106*, 11667–11672.

Kim, K., Jutooru, I., Chadalapaka, G., Johnson, G., Frank, J., Burghardt, R., Kim, S., and Safe, S. (2013). HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. Oncogene *32*, 1616–1625.

Kino, T., Hurt, D.E., Ichijo, T., Nader, N., and Chrousos, G.P. (2010). Noncoding RNA Gas5 Is a Growth Arrest- and Starvation-Associated Repressor of the Glucocorticoid Receptor. Science Signaling *3*, ra8–ra8.

Kogo, R., Shimamura, T., Mimori, K., Kawahara, K., Imoto, S., Sudo, T., Tanaka, F., Shibata, K., Suzuki, A., Komune, S., et al. (2011a). Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. Cancer Res. *71*, 6320–6326.

Konermann, S., Brigham, M.D., Trevino, A.E., Joung, J., Abudayyeh, O.O., Barcena, C., Hsu, P.D., Habib, N., Gootenberg, J.S., Nishimasu, H., et al. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature *517*, 583–588.

Kong, Q., and Qiu, M. (2018). Long noncoding RNA SNHG15 promotes human breast cancer proliferation, migration and invasion by sponging miR-211-3p. Biochem. Biophys. Res. Commun. *495*, 1594–1600.

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. *35*, W345-349.

Kornienko, A.E., Dotter, C.P., Guenzl, P.M., Gisslinger, H., Gisslinger, B., Cleary, C., Kralovics, R., Pauler, F.M., and Barlow, D.P. (2016). Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. Genome Biology *17*.

Kramer, M.C., Liang, D., Tatomer, D.C., Gold, B., March, Z.M., Cherry, S., and Wilusz, J.E. (2015). Combinatorial control of *Drosophila* circular RNA expression by intronic repeats, hnRNPs, and SR proteins. Genes & Development *29*, 2168–2182.

Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., Gottschling, D.E., and Cech, T.R. (1982). Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. Cell *31*, 147–157.

Kumar, V., Westra, H.-J., Karjalainen, J., Zhernakova, D.V., Esko, T., Hrdlickova, B., Almeida, R., Zhernakova, A., Reinmaa, E., Võsa, U., et al. (2013). Human Disease-Associated Genetic Variation Impacts Large Intergenic Non-Coding RNA Expression. PLoS Genetics *9*, e1003201.

Kwapisz, M., Ruault, M., van Dijk, E., Gourvennec, S., Descrimes, M., Taddei, A., and Morillon, A. (2015). Expression of Subtelomeric lncRNAs Links Telomeres Dynamics to RNA Decay in S. cerevisiae. Non-Coding RNA *1*, 94–126.

Lai, F., Orom, U.A., Cesaroni, M., Beringer, M., Taatjes, D.J., Blobel, G.A., and Shiekhattar, R. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature *494*, 497–501.

Lamouille, S., Subramanyam, D., Blelloch, R., and Derynck, R. (2013). Regulation of epithelial-mesenchymal and mesenchymal-epithelial transitions by microRNAs. Curr. Opin. Cell Biol. *25*, 200–207.

Lamouille, S., Xu, J., and Derynck, R. (2014a). Molecular mechanisms of epithelial-mesenchymal transition. Nat. Rev. Mol. Cell Biol. *15*, 178–196.

Lan, F., Collins, R.E., De Cegli, R., Alpatov, R., Horton, J.R., Shi, X., Gozani, O., Cheng, X., and Shi, Y. (2007a). Recognition of unmethylated histone H3 lysine 4 links BHC80 to LSD1-mediated gene repression. Nature *448*, 718–722.

Lan, F., Zaratiegui, M., Villén, J., Vaughn, M.W., Verdel, A., Huarte, M., Shi, Y., Gygi, S.P., Moazed, D., Martienssen, R.A., et al. (2007b). S. pombe LSD1 homologs regulate heterochromatin propagation and euchromatic gene transcription. Mol. Cell *26*, 89–101.

Lan, H., Tan, M., Zhang, Q., Yang, F., Wang, S., Li, H., Xiong, X., and Sun, Y. (2019). LSD1 destabilizes FBXW7 and abrogates FBXW7 functions independent of its demethylase activity. Proceedings of the National Academy of Sciences *116*, 12311–12320.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Landerer, E., Villegas, J., Burzio, V.A., Oliveira, L., Villota, C., Lopez, C., Restovic, F., Martinez, R., Castillo, O., and Burzio, L.O. (2011). Nuclear localization of the mitochondrial ncRNAs in normal and cancer cells. Cellular Oncology *34*, 297–305.

Latil, M., Nassar, D., Beck, B., Boumahdi, S., Wang, L., Brisebarre, A., Dubois, C., Nkusi, E., Lenglez, S., Checinska, A., et al. (2017). Cell-Type-Specific Chromatin States Differentially Prime Squamous Cell Carcinoma Tumor-Initiating Cells for Epithelial to Mesenchymal Transition. Cell Stem Cell *20*, 191-204.e5.

Lazorthes, S., Vallot, C., Briois, S., Aguirrebengoa, M., Thuret, J.-Y., Laurent, G.St., Rougeulle, C., Kapranov, P., Mann, C., Trouche, D., et al. (2015). A vlincRNA participates in senescence maintenance by relieving H2AZ-mediated repression at the INK4 locus. Nature Communications *6*, 5971.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843–854.

Lee, S., Kopp, F., Chang, T.-C., Sataluri, A., Chen, B., Sivakumar, S., Yu, H., Xie, Y., and Mendell, J.T. (2016). Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. Cell *164*, 69–80.

Lepoivre, C., Belhocine, M., Bergon, A., Griffon, A., Yammine, M., Vanhille, L., Zacarias-Cabeza, J., Garibal, M.-A., Koch, F., Maqbool, M., et al. (2013). Divergent transcription is associated with promoters of transcriptional regulators. BMC Genomics *14*, 914.

Lewis, J.B., Atkins, J.F., Anderson, C.W., Baum, P.R., and Gesteland, R.F. (1975). Mapping of late adenovirus genes by cell-free translation of RNA selected by hybridization to specific DNA fragments. PNAS *72*, 1344–1348.

Li, C.H., and Chen, Y. (2013). Targeting long non-coding RNAs in cancers: progress and prospects. Int. J. Biochem. Cell Biol. *45*, 1895–1910.

Li, F., Xiao, Y., Huang, F., Deng, W., Zhao, H., Shi, X., Wang, S., Yu, X., Zhang, L., Han, Z., et al. (2015a). Spatiotemporal-specific lncRNAs in the brain, colon, liver and lung of macaque during development. Mol. BioSyst. *11*, 3253–3263.

Li, H.-M., Yang, H., Wen, D.-Y., Luo, Y.-H., Liang, C.-Y., Pan, D.-H., Ma, W., Chen, G., He, Y., and Chen, J.-Q. (2017a). Overexpression of LncRNA HOTAIR is Associated with Poor Prognosis in Thyroid Carcinoma: A Study Based on TCGA and GEO Data. Horm. Metab. Res.

Li, J., Yang, J., Zhou, P., Le, Y., Zhou, C., Wang, S., Xu, D., Lin, H.-K., and Gong, Z. (2015b). Circular RNAs in cancer: novel insights into origins, properties, functions and implications. American Journal of Cancer Research *5*, 472–480.

Li, J., Han, L., Roebuck, P., Diao, L., Liu, L., Yuan, Y., Weinstein, J.N., and Liang, H. (2015c). TANRIC: An Interactive Open Platform to Explore the Function of lncRNAs in Cancer. Cancer Res. *75*, 3728–3737.

Li, L., Liu, B., Wapinski, O.L., Tsai, M.-C., Qu, K., Zhang, J., Carlson, J.C., Lin, M., Fang, F., Gupta, R.A., et al. (2013a). Targeted disruption of Hotair leads to homeotic transformation and gene derepression. Cell Rep *5*, 3–12.

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A.Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., et al. (2013b). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature *498*, 516–520.

Li, W., Notani, D., and Rosenfeld, M.G. (2016a). Enhancers as non-coding RNA

transcription units: recent insights and future perspectives. Nature Reviews Genetics *17*, 207–223.

Li, W., Zhang, Z., Liu, X., Cheng, X., Zhang, Y., Han, X., Zhang, Y., Liu, S., Yang, J., Xu, B., et al. (2017b). The FOXN3-NEAT1-SIN3A repressor complex promotes progression of hormonally responsive breast cancer. J. Clin. Invest. *127*, 3421–3440.

Li, Y., Wang, Z., Shi, H., Li, H., Li, L., Fang, R., Cai, X., Liu, B., Zhang, X., and Ye, L. (2016b). HBXIP and LSD1 Scaffolded by lncRNA Hotair Mediate Transcriptional Activation by c-Myc. Cancer Res. *76*, 293–304.

Li, Z., Huang, C., Bao, C., Chen, L., Lin, M., Wang, X., Zhong, G., Yu, B., Hu, W., Dai, L., et al. (2015d). Exon-intron circular RNAs regulate transcription in the nucleus. Nature Structural & Molecular Biology *22*, 256–264.

Liang, H., Yu, T., Han, Y., Jiang, H., Wang, C., You, T., Zhao, X., Shan, H., Yang, R., Yang, L., et al. (2018). LncRNA PTAR promotes EMT and invasion-metastasis in serous ovarian cancer by competitively binding miR-101-3p to regulate ZEB1 expression. Molecular Cancer *17*.

Liang, L., Sun, H., Zhang, W., Zhang, M., Yang, X., Kuang, R., and Zheng, H. (2016). Meta-Analysis of EMT Datasets Reveals Different Types of EMT. PLoS ONE *11*, e0156839.

Liao, J.-Y., Wu, J., Wang, Y.-J., He, J.-H., Deng, W.-X., Hu, K., Zhang, Y.-C., Zhang, Y., Yan, H., Wang, D.-L., et al. (2017). Deep sequencing reveals a global reprogramming of lncRNA transcriptome during EMT. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research *1864*, 1703–1713.

Lim, S., Janzer, A., Becker, A., Zimmer, A., Schüle, R., Buettner, R., and Kirfel, J. (2010). Lysine-specific demethylase 1 (LSD1) is highly expressed in ER-negative breast cancers and a biomarker predicting aggressive biology. Carcinogenesis *31*, 512–520.

Lin, T., Ponn, A., Hu, X., Law, B.K., and Lu, J. (2010a). Requirement of the histone demethylase LSD1 in Snai1-mediated transcriptional repression during epithelial-mesenchymal transition. Oncogene *29*, 4896–4904.

Lin, Y., Wu, Y., Li, J., Dong, C., Ye, X., Chi, Y.-I., Evers, B.M., and Zhou, B.P. (2010b). The SNAG domain of Snail1 functions as a molecular hook for recruiting lysine-specific demethylase 1. EMBO J. *29*, 1803–1816.

Liu, B., Wu, S., Ma, J., Yan, S., Xiao, Z., Wan, L., Zhang, F., Shang, M., and Mao, A. (2018). lncRNA GAS5 Reverses EMT and Tumor Stem Cell-Mediated Gemcitabine Resistance and Metastasis by Targeting miR-221/SOCS3 in Pancreatic Cancer. Mol Ther Nucleic Acids *13*, 472–482.

Liu, K., Hou, Y., Liu, Y., and Zheng, J. (2017a). LncRNA SNHG15 contributes to proliferation, invasion and autophagy in osteosarcoma cells by sponging miR-141. J. Biomed. Sci. *24*, 46.

Liu, S.J., Horlbeck, M.A., Cho, S.W., Birk, H.S., Malatesta, M., He, D., Attenello, F.J., Villalta, J.E., Cho, M.Y., Chen, Y., et al. (2017b). CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science *355*.

Lobos-González, L., Silva, V., Araya, M., Restovic, F., Echenique, J., Oliveira-Cruz, L., Fitzpatrick, C., Briones, M., Villegas, J., Villota, C., et al. (2016). Targeting antisense mitochondrial ncRNAs inhibits murine melanoma tumor growth and metastasis through reduction in survival and invasion factors. Oncotarget.

Lourenço, A.R., and Coffer, P.J. (2017). SOX4: Joining the Master Regulators of Epithelial-to-Mesenchymal Transition? Trends Cancer *3*, 571–582.

Louro, R., El-Jundi, T., Nakaya, H.I., Reis, E.M., and Verjovski-Almeida, S. (2008). Conserved tissue expression signatures of intronic noncoding RNAs transcribed from human and mouse loci. Genomics *92*, 18–25.

Luo, B., Cheung, H.W., Subramanian, A., Sharifnia, T., Okamoto, M., Yang, X., Hinkle, G., Boehm, J.S., Beroukhim, R., Weir, B.A., et al. (2008). Highly parallel identification of essential genes in cancer cells. Proc. Natl. Acad. Sci. U.S.A. *105*, 20380–20385.

Luo, M., Li, Z., Wang, W., Zeng, Y., Liu, Z., and Qiu, J. (2013). Long non-coding RNA H19 increases bladder cancer metastasis by associating with EZH2 and inhibiting E-cadherin expression. Cancer Lett. *333*, 213–221.

Magistri, M., Faghihi, M.A., St Laurent, G., and Wahlestedt, C. (2012). Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts. Trends in Genetics *28*, 389–396.

Maiques-Diaz, A., Spencer, G.J., Lynch, J.T., Ciceri, F., Williams, E.L., Amaral, F.M.R., Wiseman, D.H., Harris, W.J., Li, Y., Sahoo, S., et al. (2018). Enhancer Activation by Pharmacologic Displacement of LSD1 from GFI1 Induces Differentiation in Acute Myeloid Leukemia. Cell Rep *22*, 3641–3659.

Marques, A.C., and Ponting, C.P. (2014). Intergenic lncRNAs and the evolution of gene expression. Current Opinion in Genetics & Development *27*, 48–53.

Marques, A.C., Hughes, J., Graham, B., Kowalczyk, M.S., Higgs, D.R., and Ponting, C.P. (2013). Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. Genome Biology *14*, R131.

Massone, S., Ciarlo, E., Vella, S., Nizzari, M., Florio, T., Russo, C., Cancedda, R., and Pagano, A. (2012). NDM29, a RNA polymerase III-dependent non coding RNA, promotes amyloidogenic processing of APP and amyloid β secretion. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research *1823*, 1170–1177.

Matouk, I.J., DeGroot, N., Mezan, S., Ayesh, S., Abu-lail, R., Hochberg, A., and Galun, E. (2007). The H19 non-coding RNA is essential for human tumor growth. PLoS ONE *2*, e845.

Matouk, I.J., Raveh, E., Abu-lail, R., Mezan, S., Gilon, M., Gershtain, E., Birman, T., Gallula, J., Schneider, T., Barkali, M., et al. (2014). Oncofetal H19 RNA promotes tumor metastasis. Biochim. Biophys. Acta *1843*, 1414–1426.

Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R., Fung, J., Monteleone, E., Saghatelian, A., Nakayama, K.I., Clohessy, J.G., and Pandolfi, P.P. (2017). mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. Nature *541*, 228–232.

Mattick, J.S. (2001). Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep. *2*, 986–991.

Mattick, J.S. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. BioEssays *25*, 930–939.

Mattick, J.S., and Gagen, M.J. (2001). The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. Mol. Biol. Evol. *18*, 1611–1630.

Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J.A., and Churchman, L.S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell *161*, 541–554.

McDonald, O.G., Wu, H., Timp, W., Doi, A., and Feinberg, A.P. (2011a). Genome-scale epigenetic reprogramming during epithelial-to-mesenchymal transition. Nat. Struct. Mol. Biol. *18*, 867–874.

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S.D., Gregersen, L.H., Munschauer, M., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. Nature *495*, 333–338.

Mercer, T.R., Wilhelm, D., Dinger, M.E., Solda, G., Korbie, D.J., Glazov, E.A., Truong, V., Schwenke, M., Simons, C., Matthaei, K.I., et al. (2011). Expression of distinct RNAs from 3' untranslated regions. Nucleic Acids Research *39*, 2393–2403.

Mercer, T.R., Gerhardt, D.J., Dinger, M.E., Crawford, J., Trapnell, C., Jeddeloh, J.A., Mattick, J.S., and Rinn, J.L. (2012). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nature Biotechnology *30*, 99–104.

Meseure, D., Vacher, S., Lallemand, F., Alsibai, K.D., Hatem, R., Chemlali, W., Nicolas, A., De Koning, L., Pasmant, E., Callens, C., et al. (2016). Prognostic value of a newly identified MALAT1 alternatively spliced transcript in breast cancer. British Journal of Cancer *114*, 1395–1404.

Mestdagh, P., Fredlund, E., Pattyn, F., Rihani, A., Van Maerken, T., Vermeulen, J., Kumps, C., Menten, B., De Preter, K., Schramm, A., et al. (2010). An integrative genomics screen uncovers ncRNA T-UCR functions in neuroblastoma tumours. Oncogene *29*, 3583–3592.

Metzger, E., Wissmann, M., Yin, N., Müller, J.M., Schneider, R., Peters, A.H.F.M.,

Günther, T., Buettner, R., and Schüle, R. (2005a). LSD1 demethylates repressive histone marks to promote androgen-receptor-dependent transcription. Nature *437*, 436–439.

Milligan, M.J., and Lipovich, L. (2015). Pseudogene-derived lncRNAs: emerging regulators of gene expression. Frontiers in Genetics *5*.

Mise, N., Savai, R., Yu, H., Schwarz, J., Kaminski, N., and Eickelberg, O. (2012). Zyxin is a transforming growth factor-β (TGF-β)/Smad3 target gene that regulates lung cancer cell motility via integrin α5β1. J. Biol. Chem. *287*, 31393–31405.

Mondal, T., Rasmussen, M., Pandey, G.K., Isaksson, A., and Kanduri, C. (2010). Characterization of the RNA content of chromatin. Genome Research *20*, 899–907.

Mondal, T., Subhash, S., Vaid, R., Enroth, S., Uday, S., Reinius, B., Mitra, S., Mohammed, A., James, A.R., Hoberg, E., et al. (2015). MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA–DNA triplex structures. Nature Communications *6*, 7743.

Montalbano, A., Canver, M.C., and Sanjana, N.E. (2017). High-Throughput Approaches to Pinpoint Function within the Noncoding Genome. Mol. Cell *68*, 44–59.

Montgomery, M.K. (2004). RNA Interference. In RNA Interference, Editing, and Modification, J.M. Gott, ed. (Totowa, NJ: Humana Press), pp. 3–21.

Morlando, M., Ballarino, M., Fatica, A., and Bozzoni, I. (2014). The role of long noncoding RNAs in the epigenetic control of gene expression. ChemMedChem *9*, 505–510.

Morris, K.V., Santoso, S., Turner, A.-M., Pastori, C., and Hawkins, P.G. (2008). Bidirectional Transcription Directs Both Transcriptional Gene Activation and Suppression in Human Cells. PLoS Genetics *4*, e1000258.

Mouse ENCODE Consortium, Stamatoyannopoulos, J.A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D.M., Groudine, M., Bender, M., Kaul, R., et al. (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biology *13*, 418.

Nair, S.S., Nair, B.C., Cortez, V., Chakravarty, D., Metzger, E., Schüle, R., Brann, D.W., Tekmal, R.R., and Vadlamudi, R.K. (2010). PELP1 is a reader of histone H3 methylation that facilitates oestrogen receptor-alpha target gene activation by regulating lysine demethylase 1 specificity. EMBO Rep. *11*, 438–444.

Nakaya, H.I., Amaral, P.P., Louro, R., Lopes, A., Fachel, A.A., Moreira, Y.B., El-Jundi, T.A., da Silva, A.M., Reis, E.M., and Verjovski-Almeida, S. (2007). Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. Genome Biology *8*, R43.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker,

J.C., Grützner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature *505*, 635–640.

Nelson, B.R., Makarewich, C.A., Anderson, D.M., Winders, B.R., Troupes, C.D., Wu, F., Reese, A.L., McAnally, J.R., Chen, X., Kavalali, E.T., et al. (2016). A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. Science *351*, 271–275.

Nieto, M.A., Huang, R.Y.-J., Jackson, R.A., and Thiery, J.P. (2016). EMT: 2016. Cell *166*, 21–45.

Noh, J.H., Kim, K.M., Abdelmohsen, K., Yoon, J.-H., Panda, A.C., Munk, R., Kim, J., Curtis, J., Moad, C.A., Wohler, C.M., et al. (2016). HuR and GRSF1 modulate the nuclear export and mitochondrial localization of the lncRNA *RMRP*. Genes & Development.

Nojima, T., Gomes, T., Grosso, A.R.F., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N.J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. Cell *161*, 526–540.

Ntini, E., Järvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jørgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P.R., et al. (2013). Polyadenylation site–induced decay of upstream transcripts enforces promoter directionality. Nature Structural & Molecular Biology *20*, 923–928.

Ochoa, S. (1980). A Pursuit of a Hobby. Annual Review of Biochemistry *49*, 1–31.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature *420*, 563–573.

Ombrato, L., and Malanchi, I. (2014). The EMT universe: space between cancer cell dissemination and metastasis initiation. Crit Rev Oncog *19*, 349–361.

Orgel, L.E., and Crick, F.H.C. (1980). Selfish DNA: the ultimate parasite. Nature *284*, 604–607.

Ørom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., et al. (2010). Long Noncoding RNAs with Enhancer-like Function in Human Cells. Cell *143*, 46–58.

Pachnis, V., Belayew, A., and Tilghman, S.M. (1984). Locus unlinked to alpha-fetoprotein under the control of the murine raf and Rif genes. Proc. Natl. Acad. Sci. U.S.A. *81*, 5523–5527.

Pádua Alves, C., Fonseca, A.S., Muys, B.R., de Barros E Lima Bueno, R., Bürger, M.C., de Souza, J.E.S., Valente, V., Zago, M.A., and Silva, W.A. (2013). Brief report: The lincRNA Hotair is required for epithelial-to-mesenchymal transition and stemness maintenance of cancer cell lines. Stem Cells *31*, 2827–2832.

Pastushenko, I., Brisebarre, A., Sifrim, A., Fioramonti, M., Revenco, T., Boumahdi,

S., Van Keymeulen, A., Brown, D., Moers, V., Lemaire, S., et al. (2018). Identification of the tumour transition states occurring during EMT. Nature *556*, 463–468.

Pavan, S., Meyer-Schaller, N., Diepenbruck, M., Kalathur, R.K.R., Saxena, M., and Christofori, G. (2018). A kinome-wide high-content siRNA screen identifies MEK5-ERK5 signaling as critical for breast cancer cell EMT and metastasis. Oncogene *37*, 4197–4213.

Peng, L., Yuan, X., and Li, G. (2015). The emerging landscape of circular RNA ciRS-7 in cancer (Review). Oncology Reports.

Pickard, M.R., and Williams, G.T. (2015). Molecular and Cellular Mechanisms of Action of Tumour Suppressor GAS5 LncRNA. Genes (Basel) *6*, 484–499.

Pinskaya, M., Saci, Z., Gallopin, M., Nguyen, N.H., Gabriel, M., Firlej, V., Descrimes, M., Taille, A. de la, Londoño-Vallejo, A., Allory, Y., et al. (2019). Blind exploration of the unreferenced transcriptome reveals novel RNAs for prostate cancer diagnosis. BioRxiv 644104.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J., and Pandolfi, P.P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature *465*, 1033–1038.

Pollard, K.S., Salama, S.R., King, B., Kern, A.D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J.S., Bejerano, G., Baertsch, R., et al. (2006a). Forces shaping the fastest evolving regions in the human genome. PLoS Genet. *2*, e168.

Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. (2006b). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167–172.

Polyak, K., and Weinberg, R.A. (2009). Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. Nat. Rev. Cancer *9*, 265–273.

Ponting, C.P., Oliver, P.L., and Reik, W. (2009). Evolution and Functions of Long Noncoding RNAs. Cell *136*, 629–641.

Porro, A., Feuerhahn, S., Reichenbach, P., and Lingner, J. (2010). Molecular Dissection of Telomeric Repeat-Containing RNA Biogenesis Unveils the Presence of Distinct and Multiple Regulatory Pathways. Molecular and Cellular Biology *30*, 4808–4817.

Portoso, M., Ragazzini, R., Brenčič, Ž., Moiani, A., Michaud, A., Vassilev, I., Wassef, M., Servant, N., Sargueil, B., and Margueron, R. (2017a). PRC2 is dispensable for HOTAIR-mediated transcriptional repression. EMBO J. *36*, 981–994.

Postepska-Igielska, A., Giwojna, A., Gasri-Plotnitsky, L., Schmitt, N., Dold, A., Ginsberg, D., and Grummt, I. (2015). LncRNA Khps1 Regulates Expression of the Proto-oncogene SPHK1 via Triplex-Mediated Changes in Chromatin Structure. Molecular Cell *60*, 626–636.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters. Science *322*, 1851–1854.

Preker, P., Almvig, K., Christensen, M.S., Valen, E., Mapendano, C.K., Sandelin, A., and Jensen, T.H. (2011). PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. Nucleic Acids Research *39*, 7179–7193.

Prensner, J.R., Iyer, M.K., Balbin, O.A., Dhanasekaran, S.M., Cao, Q., Brenner, J.C., Laxman, B., Asangani, I.A., Grasso, C.S., Kominsky, H.D., et al. (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nature Biotechnology *29*, 742–749.

Qiu, J., Lin, Y., Ye, L., Ding, J., Feng, W., Jin, H., Zhang, Y., Li, Q., and Hua, K. (2014). Overexpression of long non-coding RNA HOTAIR predicts poor patient prognosis and promotes tumor metastasis in epithelial ovarian cancer. Gynecol. Oncol. *134*, 121–128.

Qu, X., Alsager, S., Zhuo, Y., and Shan, B. (2019). HOX transcript antisense RNA (HOTAIR) in cancer. Cancer Lett. *454*, 90–97.

Quénet, D., and Dalal, Y. (2014). A long non-coding RNA is required for targeting centromeric protein A to the human centromere. ELife *3*.

Quinn, J.J., and Chang, H.Y. (2016). Unique features of long non-coding RNA biogenesis and function. Nat. Rev. Genet. *17*, 47–62.

Rackham, O., Shearwood, A.-M.J., Mercer, T.R., Davies, S.M.K., Mattick, J.S., and Filipovska, A. (2011). Long noncoding RNAs are generated from the mitochondrial genome and regulated by nuclear-encoded proteins. RNA *17*, 2085–2093.

Rands, C.M., Meader, S., Ponting, C.P., and Lunter, G. (2014). 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. PLoS Genetics *10*, e1004525.

Raveh, E., Matouk, I.J., Gilon, M., and Hochberg, A. (2015). The H19 Long non-coding RNA in cancer initiation, progression and metastasis - a proposed unifying theory. Mol. Cancer *14*, 184.

Richards, E.J., Zhang, G., Li, Z.-P., Permuth-Wey, J., Challa, S., Li, Y., Kong, W., Dan, S., Bui, M.M., Coppola, D., et al. (2015). Long non-coding RNAs (LncRNA) regulated by transforming growth factor (TGF) β: LncRNA-hit-mediated TGFβ-induced epithelial to mesenchymal transition in mammary epithelia. J. Biol. Chem. *290*, 6857–6867.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., et al. (2007a). Functional

demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311–1323.

Rodriguez-Boulan, E., and Macara, I.G. (2014). Organization and execution of the epithelial polarity programme. Nat. Rev. Mol. Cell Biol. *15*, 225–242.

Rošić, S., Köhler, F., and Erhardt, S. (2014). Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. The Journal of Cell Biology *207*, 335–349.

Roth, M., Wang, Z., and Chen, W.Y. (2016). SIRT1 and LSD1 competitively regulate KU70 functions in DNA repair and mutation acquisition in cancer cells. Oncotarget *7*, 50195–50214.

Ruiz-Orera, J., Messeguer, X., Subirana, J.A., and Alba, M.M. (2014). Long non-coding RNAs as a source of new peptides. ELife *3*.

Ruscetti, M., Quach, B., Dadashian, E.L., Mulholland, D.J., and Wu, H. (2015). Tracking and Functional Characterization of Epithelial-Mesenchymal Transition and Mesenchymal Tumor Cells during Prostate Cancer Metastasis. Cancer Res. *75*, 2749–2759.

Rybak-Wolf, A., Stottmeister, C., Glažar, P., Jens, M., Pino, N., Giusti, S., Hanan, M., Behm, M., Bartok, O., Ashwal-Fluss, R., et al. (2015). Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. Molecular Cell *58*, 870–885.

Saeinasab, M., Bahrami, A.R., González, J., Marchese, F.P., Martinez, D., Mowla, S.J., Matin, M.M., and Huarte, M. (2019). SNHG15 is a bifunctional MYC-regulated noncoding locus encoding a lncRNA that promotes cell proliferation, invasion and drug resistance in colorectal cancer by interacting with AIF. J. Exp. Clin. Cancer Res. *38*, 172.

Saint-André, V., Batsché, E., Rachez, C., and Muchardt, C. (2011). Histone H3 lysine 9 trimethylation and HP1γ favor inclusion of alternative exons. Nat. Struct. Mol. Biol. *18*, 337–344.

Sakr, W.A., Zarbo, R.J., Jacobs, J.R., and Crissman, J.D. (1987). Distribution of basement membrane in squamous cell carcinoma of the head and neck. Human Pathology *18*, 1043–1050.

Sanjana, N.E., Wright, J., Zheng, K., Shalem, O., Fontanillas, P., Joung, J., Cheng, C., Regev, A., and Zhang, F. (2016). High-resolution interrogation of functional elements in the noncoding genome. Science *353*, 1545–1549.

Santamaría, P.G., Moreno-Bueno, G., and Cano, A. (2019). Contribution of Epithelial Plasticity to Therapy Resistance. J Clin Med *8*.

Scaruffi, P., Stigliani, S., Moretti, S., Coco, S., De Vecchi, C., Valdora, F., Garaventa, A., Bonassi, S., and Tonini, G.P. (2009). Transcribed-ultra conserved region expression is associated with outcome in high-risk neuroblastoma. BMC Cancer *9*.

Schenk, T., Chen, W.C., Göllner, S., Howell, L., Jin, L., Hebestreit, K., Klein, H.-U., Popescu, A.C., Burnett, A., Mills, K., et al. (2012). Inhibition of the LSD1 (KDM1A) demethylase reactivates the all-trans-retinoic acid differentiation pathway in acute myeloid leukemia. Nat. Med. *18*, 605–611.

Scoumanne, A., and Chen, X. (2007). The lysine-specific demethylase 1 is required for cell proliferation in both p53-dependent and -independent manners. J. Biol. Chem. *282*, 15471–15475.

Scruggs, B.S., Gilchrist, D.A., Nechaev, S., Muse, G.W., Burkholder, A., Fargo, D.C., and Adelman, K. (2015). Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. Molecular Cell *58*, 1101–1112.

Sehrawat, A., Gao, L., Wang, Y., Bankhead, A., McWeeney, S.K., King, C.J., Schwartzman, J., Urrutia, J., Bisson, W.H., Coleman, D.J., et al. (2018). LSD1 activates a lethal prostate cancer gene network independently of its demethylase function. Proceedings of the National Academy of Sciences *115*, E4179–E4188.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent Transcription from Active Promoters. Science *322*, 1849–1851.

Sengupta, S., and George, R.E. (2017). Super-enhancer-driven Transcriptional Dependencies in Cancer. Trends Cancer *3*, 269–281.

Shahryari, A., Jazi, M.S., Samaei, N.M., and Mowla, S.J. (2015). Long non-coding RNA SOX2OT: expression signature, splicing patterns, and emerging roles in pluripotency and tumorigenesis. Frontiers in Genetics *6*.

Shi, S.-J., Wang, L.-J., Yu, B., Li, Y.-H., Jin, Y., and Bai, X.-Z. (2015). LncRNA-ATB promotes trastuzumab resistance and invasion-metastasis cascade in breast cancer. Oncotarget *6*, 11652–11663.

Shi, Y., Lan, F., Matson, C., Mulligan, P., Whetstine, J.R., Cole, P.A., Casero, R.A., and Shi, Y. (2004a). Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. Cell *119*, 941–953.

Shimojo, M., and Hersh, L.B. (2006). Characterization of the REST/NRSF-interacting LIM domain protein (RILP): localization and interaction with REST/NRSF. Journal of Neurochemistry *96*, 1130–1138.

Shukla, C.J., McCorkindale, A.L., Gerhardinger, C., Korthauer, K.D., Cabili, M.N., Shechner, D.M., Irizarry, R.A., Maass, P.G., and Rinn, J.L. (2018). High-throughput identification of RNA nuclear enrichment sequences. EMBO J. *37*.

Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., et al. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proceedings of the National Academy of Sciences *110*, 2876–2881.

Singh, D.K., and Prasanth, K.V. (2013). Functional insights into the role of nuclear-

retained long noncoding RNAs in gene expression control in mammalian cells. Chromosome Research *21*, 695–711.

Skrypek, N., Goossens, S., De Smedt, E., Vandamme, N., and Berx, G. (2017). Epithelial-to-Mesenchymal Transition: Epigenetic Reprogramming Driving Cellular Plasticity. Trends Genet. *33*, 943–959.

Somarowthu, S., Legiewicz, M., Chillón, I., Marcia, M., Liu, F., and Pyle, A.M. (2015). HOTAIR Forms an Intricate and Modular Secondary Structure. Molecular Cell *58*, 353–361.

Song, X., Wang, X., Arai, S., and Kurokawa, R. (2012). Promoter-Associated Noncoding RNA from the CCND1 Promoter. In Transcriptional Regulation, A. Vancura, ed. (New York, NY: Springer New York), pp. 609–622.

Song, Y., Wang, R., Li, L.-W., Liu, X., Wang, Y.-F., Wang, Q.-X., and Zhang, Q. (2019). Long non-coding RNA HOTAIR mediates the switching of histone H3 lysine 27 acetylation to methylation to promote epithelial-to-mesenchymal transition in gastric cancer. Int. J. Oncol. *54*, 77–86.

Spurlock, C.F., Tossberg, J.T., Guo, Y., Collier, S.P., Crooke, P.S., and Aune, T.M. (2015). Expression and functions of long noncoding RNAs during human T helper cell differentiation. Nature Communications *6*, 6932.

St Laurent, G., Shtokalo, D., Tackett, M.R., Yang, Z., Eremina, T., Wahlestedt, C., Urcuqui-Inchima, S., Seilheimer, B., McCaffrey, T.A., and Kapranov, P. (2012). Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells. BMC Genomics *13*, 504.

St Laurent, G., Shtokalo, D., Dong, B., Tackett, M.R., Fan, X., Lazorthes, S., Nicolas, E., Sang, N., Triche, T.J., McCaffrey, T.A., et al. (2013). VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. Genome Biology *14*, R73.

Su, W.-Y., Xiong, H., and Fang, J.-Y. (2010). Natural antisense transcripts regulate gene expression in an epigenetic manner. Biochemical and Biophysical Research Communications *396*, 177–181.

Sun, Y., and Ma, L. (2019). New Insights into Long Non-Coding RNA MALAT1 in Cancer and Metastasis. Cancers *11*, 216.

Sun, M., Nie, F., Wang, Y., Zhang, Z., Hou, J., He, D., Xie, M., Xu, L., De, W., Wang, Z., et al. (2016). LncRNA HOXA11-AS Promotes Proliferation and Invasion of Gastric Cancer by Scaffolding the Chromatin Modification Factors PRC2, LSD1, and DNMT1. Cancer Res. *76*, 6299–6310.

Sun, Q., Hao, Q., and Prasanth, K.V. (2018). Nuclear Long Noncoding RNAs: Key Regulators of Gene Expression. Trends Genet. *34*, 142–157.

Sun, X., Bai, Y., Yang, C., Hu, S., Hou, Z., and Wang, G. (2019). Long noncoding RNA SNHG15 enhances the development of colorectal carcinoma via functioning as a

ceRNA through miR-141/SIRT1/Wnt/β-catenin axis. Artif Cells Nanomed Biotechnol *47*, 2536–2544.

Szcześniak, M.W., and Makałowska, I. (2016). lncRNA-RNA Interactions across the Human Transcriptome. PLOS ONE *11*, e0150353.

Tay, Y., Rinn, J., and Pandolfi, P.P. (2014). The multilayered complexity of ceRNA crosstalk and competition. Nature *505*, 344–352.

Terashima, M., Tange, S., Ishimura, A., and Suzuki, T. (2017). MEG3 Long Noncoding RNA Contributes to the Epigenetic Regulation of Epithelial-Mesenchymal Transition in Lung Cancer Cell Lines. J. Biol. Chem. *292*, 82–99.

The FANTOM Consortium (2005a). The Transcriptional Landscape of the Mammalian Genome. Science *309*, 1559–1563.

Thiery, J.P. (2002). Epithelial-mesenchymal transitions in tumour progression. Nat. Rev. Cancer *2*, 442–454.

Thomson, D.W., and Dinger, M.E. (2016). Endogenous microRNA sponges: evidence and controversy. Nature Reviews Genetics *17*, 272–283.

Tong, J., Ma, X., Yu, H., and Yang, J. (2019). SNHG15: a promising cancer-related long noncoding RNA. Cancer Manag Res *11*, 5961–5969.

Tornavaca, O., Chia, M., Dufton, N., Almagro, L.O., Conway, D.E., Randi, A.M., Schwartz, M.A., Matter, K., and Balda, M.S. (2015). ZO-1 controls endothelial adherens junctions, cell–cell tension, angiogenesis, and barrier formation. The Journal of Cell Biology *208*, 821–838.

Trofimova, I., Chervyakova, D., and Krasikova, A. (2015). Transcription of subtelomere tandemly repetitive DNA in chicken embryogenesis. Chromosome Research *23*, 495–503.

Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. (2010a). Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. Science *329*, 689–693.

Tsai, M.-C., Spitale, R.C., and Chang, H.Y. (2011). Long intergenic noncoding RNAs: new links in cancer progression. Cancer Res. *71*, 3–7.

Tseng, Y.-Y., Moriarity, B.S., Gong, W., Akiyama, R., Tiwari, A., Kawakami, H., Ronning, P., Reuland, B., Guenther, K., Beadnell, T.C., et al. (2014). PVT1 dependence in cancer with MYC copy-number increase. Nature *512*, 82–86.

Uesaka, M., Nishimura, O., Go, Y., Nakashima, K., Agata, K., and Imamura, T. (2014). Bidirectional promoters are the major source of gene activation-associated non-coding RNAs in mammals. BMC Genomics *15*, 35.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution. Cell *147*, 1537–1550.

Vannier, C., Mock, K., Brabletz, T., and Driever, W. (2013). Zeb1 regulates E-cadherin and Epcam (epithelial cell adhesion molecule) expression to control cell behavior in early zebrafish development. J. Biol. Chem. *288*, 18643–18659.

Venter, J.C. (2001). The Sequence of the Human Genome. Science *291*, 1304–1351.

Vidaurre, S., Fitzpatrick, C., Burzio, V.A., Briones, M., Villota, C., Villegas, J., Echenique, J., Oliveira-Cruz, L., Araya, M., Borgna, V., et al. (2014). Down-regulation of the Antisense Mitochondrial Non-coding RNAs (ncRNAs) Is a Unique Vulnerability of Cancer Cells and a Potential Target for Cancer Therapy. Journal of Biological Chemistry *289*, 27182–27198.

Volders, P.-J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P., and Vandesompele, J. (2019). LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Res. *47*, D135–D139.

Wang, J., Scully, K., Zhu, X., Cai, L., Zhang, J., Prefontaine, G.G., Krones, A., Ohgi, K.A., Zhu, P., Garcia-Bassets, I., et al. (2007a). Opposing LSD1 complexes function in developmental gene activation and repression programmes. Nature *446*, 882–887.

Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.-P., and Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. *41*, e74.

Wang, P., Xue, Y., Han, Y., Lin, L., Wu, C., Xu, S., Jiang, Z., Xu, J., Liu, Q., and Cao, X. (2014). The STAT3-Binding Long Noncoding RNA lnc-DC Controls Human Dendritic Cell Differentiation. Science *344*, 310–313.

Wang, X., Arai, S., Song, X., Reichart, D., Du, K., Pascual, G., Tempst, P., Rosenfeld, M.G., Glass, C.K., and Kurokawa, R. (2008). Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. Nature *454*, 126–130.

Wang, Y., Devereux, W., Woster, P.M., Stewart, T.M., Hacker, A., and Casero, R.A. (2001). Cloning and characterization of a human polyamine oxidase that is inducible by polyamine analogue exposure. Cancer Res. *61*, 5370–5373.

Wang, Y., Zhang, H., Chen, Y., Sun, Y., Yang, F., Yu, W., Liang, J., Sun, L., Yang, X., Shi, L., et al. (2009). LSD1 Is a Subunit of the NuRD Complex and Targets the Metastasis Programs in Breast Cancer. Cell *138*, 660–672.

Ward, M., McEwan, C., Mills, J.D., and Janitz, M. (2015). Conservation and tissue-specific transcription patterns of long noncoding RNAs. Journal of Human Transcriptome *1*, 2–9.

Washietl, S., Kellis, M., and Garber, M. (2014). Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. Genome Research *24*, 616–628.

Wassenegger, M., Heimes, S., Riedel, L., and Sänger, H.L. (1994). RNA-directed de novo methylation of genomic sequences in plants. Cell *76*, 567–576.

Waterston, R., and Sulston, J. (1995). The genome of Caenorhabditis elegans. Proc. Natl. Acad. Sci. U.S.A. *92*, 10836–10840.

Watters, K.M., Bryan, K., Foley, N.H., Meehan, M., and Stallings, R.L. (2013). Expressional alterations in functional ultra-conserved non-coding rnas in response to all-transretinoic acid - induced differentiation in neuroblastoma cells. BMC Cancer *13*.

Wei, W., Pelechano, V., Järvelin, A.I., and Steinmetz, L.M. (2011). Functional consequences of bidirectional promoters. Trends in Genetics *27*, 267–276.

Weinberg, R.A., and Penman, S. (1968). Small molecular weight monodisperse nuclear RNA. Journal of Molecular Biology *38*, 289–304.

Werner, A. (2013). Biological functions of natural antisense transcripts. BMC Biology *11*, 31.

Werner, M.S., and Ruthenburg, A.J. (2015a). Nuclear Fractionation Reveals Thousands of Chromatin-Tethered Noncoding RNAs Adjacent to Active Genes. Cell Reports *12*, 1089–1098.

Werner, M.S., Sullivan, M.A., Shah, R.N., Nadadur, R.D., Grzybowski, A.T., Galat, V., Moskowitz, I.P., and Ruthenburg, A.J. (2017). Chromatin-enriched lncRNAs can act as cell-type specific activators of proximal gene transcription. Nat. Struct. Mol. Biol. *24*, 596–603.

Wery, M., Descrimes, M., Vogt, N., Dallongeville, A.-S., Gautheret, D., and Morillon, A. (2016). Nonsense-Mediated Decay Restricts LncRNA Levels in Yeast Unless Blocked by Double-Stranded RNA Structure. Molecular Cell *61*, 379–392.

Wery, M., Gautier, C., Descrimes, M., Yoda, M., Migeot, V., Hermand, D., and Morillon, A. (2018a). Bases of antisense lncRNA-associated regulation of gene expression in fission yeast. PLoS Genet. *14*, e1007465.

Wery, M., Gautier, C., Descrimes, M., Yoda, M., Vennin-Rendos, H., Migeot, V., Gautheret, D., Hermand, D., and Morillon, A. (2018b). Native elongating transcript sequencing reveals global anti-correlation between sense and antisense nascent transcription in fission yeast. RNA *24*, 196–208.

Whyte, W.A., Bilodeau, S., Orlando, D.A., Hoke, H.A., Frampton, G.M., Foster, C.T., Cowley, S.M., and Young, R.A. (2012). Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. Nature.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell *75*, 855–862.

Willingham, A.T., and Gingeras, T.R. (2006). TUF Love for "Junk" DNA. Cell *125*, 1215–1220.

Wilusz, J.E., Freier, S.M., and Spector, D.L. (2008). 3′ End Processing of a Long

Nuclear-Retained Noncoding RNA Yields a tRNA-like Cytoplasmic RNA. Cell *135*, 919–932.

Wilusz, J.E., JnBaptiste, C.K., Lu, L.Y., Kuhn, C.-D., Joshua-Tor, L., and Sharp, P.A. (2012). A triple helix stabilizes the 3' ends of long noncoding RNAs that lack poly(A) tails. Genes & Development *26*, 2392–2407.

Wissmann, M., Yin, N., Müller, J.M., Greschik, H., Fodor, B.D., Jenuwein, T., Vogler, C., Schneider, R., Günther, T., Buettner, R., et al. (2007). Cooperative demethylation by JMJD2C and LSD1 promotes androgen receptor-dependent gene expression. Nat. Cell Biol. *9*, 347–353.

Wong, L.H., Brettingham-Moore, K.H., Chan, L., Quach, J.M., Anderson, M.A., Northrop, E.L., Hannan, R., Saffery, R., Shaw, M.L., Williams, E., et al. (2007). Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere. Genome Research *17*, 1146–1160.

Wood, E.J., Chin-Inmanu, K., Jia, H., and Lipovich, L. (2013). Sense-antisense gene pairs: sequence, transcription, and structure are not conserved between human and mouse. Frontiers in Genetics *4*.

Wu, C.-L., Wang, Y., Jin, B., Chen, H., Xie, B.-S., and Mao, Z.-B. (2015). Senescence-associated Long Non-coding RNA (*SALNR*) Delays Oncogene-induced Senescence through NF90 Regulation. Journal of Biological Chemistry *290*, 30175–30192.

Wu, D.-D., Irwin, D.M., and Zhang, Y.-P. (2011). De Novo Origin of Human Protein-Coding Genes. PLoS Genetics *7*, e1002379.

Wu, L., Murat, P., Matak-Vinkovic, D., Murrell, A., and Balasubramanian, S. (2013). Binding Interactions between Long Noncoding RNA HOTAIR and PRC2 Proteins. Biochemistry *52*, 9519–9527.

Xu, G., Meng, L., Yuan, D., Li, K., Zhang, Y., Dang, C., and Zhu, K. (2018). MEG3/miR- 21 axis affects cell mobility by suppressing epithelial- mesenchymal transition in gastric cancer. Oncology Reports *40*, 39–48.

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Münster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L.M. (2009). Bidirectional promoters generate pervasive transcription in yeast. Nature *457*, 1033–1037.

Xu, Z.-Y., Yu, Q.-M., Du, Y.-A., Yang, L.-T., Dong, R.-Z., Huang, L., Yu, P.-F., and Cheng, X.-D. (2013). Knockdown of Long Non-coding RNA HOTAIR Suppresses Tumor Invasion and Reverses Epithelial-mesenchymal Transition in Gastric Cancer. International Journal of Biological Sciences *9*, 587–597.

Xue, X., Yang, Y.A., Zhang, A., Fong, K.-W., Kim, J., Song, B., Li, S., Zhao, J.C., and Yu, J. (2016). LncRNA HOTAIR enhances ER signaling and confers tamoxifen resistance in breast cancer. Oncogene *35*, 2746–2755.

Yang, L., Duff, M.O., Graveley, B.R., Carmichael, G.G., and Chen, L.-L. (2011). Genomewide characterization of non-polyadenylated RNAs. Genome Biology *12*,

R16.

Yang, Y.-T., Wang, X., Zhang, Y.-Y., and Yuan, W.-J. (2019). The histone demethylase LSD1 promotes renal inflammation by mediating TLR4 signaling in hepatitis B virus-associated glomerulonephritis. Cell Death & Disease *10*, 278.

Yao, H., Brick, K., Evrard, Y., Xiao, T., Camerini-Otero, R.D., and Felsenfeld, G. (2010). Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. Genes & Development *24*, 2543–2555.

Ye, X., and Weinberg, R.A. (2015). Epithelial-Mesenchymal Plasticity: A Central Regulator of Cancer Progression. Trends Cell Biol. *25*, 675–686.

Ye, K., Wang, S., Zhang, H., Han, H., Ma, B., and Nan, W. (2017). Long Noncoding RNA GAS5 Suppresses Cell Growth and Epithelial-Mesenchymal Transition in Osteosarcoma by Regulating the miR-221/ARHI Pathway. J. Cell. Biochem. *118*, 4772–4781.

Yin, Q.-F., Yang, L., Zhang, Y., Xiang, J.-F., Wu, Y.-W., Carmichael, G.G., and Chen, L.-L. (2012). Long Noncoding RNAs with snoRNA Ends. Molecular Cell *48*, 219–230.

Ying, L., Chen, Q., Wang, Y., Zhou, Z., Huang, Y., and Qiu, F. (2012). Upregulated MALAT-1 contributes to bladder cancer cell migration by inducing epithelial-to-mesenchymal transition. Mol Biosyst *8*, 2289–2294.

Yoon, J.-H., Abdelmohsen, K., Kim, J., Yang, X., Martindale, J.L., Tominaga-Yamanaka, K., White, E.J., Orjalo, A.V., Rinn, J.L., Kreft, S.G., et al. (2013a). Scaffold function of long non-coding RNA HOTAIR in protein ubiquitination. Nature Communications *4*.

Young, R.S., and Ponting, C.P. (2013). Identification and function of long non-coding RNAs. Essays In Biochemistry *54*, 113–126.

Yuan, C., Wang, J., Harrison, A.P., Meng, X., Chen, D., and Chen, M. (2015a). Genome-wide view of natural antisense transcripts in Arabidopsis thaliana. DNA Research *22*, 233–243.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D., et al. (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature *515*, 355–364.

Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics *25*, 1952–1958.

Zaravinos, A. (2015). The Regulatory Role of MicroRNAs in EMT and Cancer. J Oncol *2015*, 865816.

Zeisberg, M., and Neilson, E.G. (2009). Biomarkers for epithelial-mesenchymal transitions. J. Clin. Invest. *119*, 1429–1437.

Zeng, X., Jedrychowski, M.P., Chen, Y., Serag, S., Lavery, G.G., Gygi, S.P., and Spiegelman, B.M. (2016). Lysine-specific demethylase 1 promotes brown adipose tissue thermogenesis via repressing glucocorticoid activation. Genes Dev. *30*, 1822–1836.

Zhang, A., Zhao, J.C., Kim, J., Fong, K.-W., Yang, Y.A., Chakravarti, D., Mo, Y.-Y., and Yu, J. (2015). LncRNA HOTAIR Enhances the Androgen-Receptor-Mediated Transcriptional Program and Drives Castration-Resistant Prostate Cancer. Cell Rep *13*, 209–221.

Zhang, B., Gunawardane, L., Niazi, F., Jahanbani, F., Chen, X., and Valadkhan, S. (2014). A Novel RNA Motif Mediates the Strict Nuclear Localization of a Long Noncoding RNA. Molecular and Cellular Biology *34*, 2318–2329.

Zhang, L., Carnesecchi, J., Cerutti, C., Tribollet, V., Périan, S., Forcet, C., Wong, J., and Vanacker, J.-M. (2018a). LSD1-ERRα complex requires NRF1 to positively regulate transcription and cell invasion. Scientific Reports *8*.

Zhang, X., Hamblin, M.H., and Yin, K.-J. (2017). The long noncoding RNA Malat1: Its physiological and pathophysiological functions. RNA Biol *14*, 1705–1714.

Zhang, X., Feng, W., Zhang, J., Ge, L., Zhang, Y., Jiang, X., Peng, W., Wang, D., Gong, A., and Xu, M. (2018b). Long non- coding RNA PVT1 promotes epithelial- mesenchymal transition via the TGF- β/Smad pathway in pancreatic cancer cells. Oncol. Rep. *40*, 1093–1102.

Zhang, Y., Zhang, X.-O., Chen, T., Xiang, J.-F., Yin, Q.-F., Xing, Y.-H., Zhu, S., Yang, L., and Chen, L.-L. (2013). Circular Intronic Long Noncoding RNAs. Molecular Cell *51*, 792–806.

Zhao, W., An, Y., Liang, Y., and Xie, X.-W. (2014). Role of HOTAIR long noncoding RNA in metastatic progression of lung cancer. Eur Rev Med Pharmacol Sci *18*, 1930–1936.

Zheng, D., Frankish, A., Baertsch, R., Kapranov, P., Reymond, A., Choo, S.W., Lu, Y., Denoeud, F., Antonarakis, S.E., Snyder, M., et al. (2007). Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution. Genome Research *17*, 839–851.

Zheng, R., Shen, Z., Tripathi, V., Xuan, Z., Freier, S.M., Bennett, C.F., Prasanth, S.G., and Prasanth, K.V. (2010). Polypurine-repeat-containing RNAs: a novel class of long non-coding RNA in mammalian cells. Journal of Cell Science *123*, 3734–3744.

Zheng, S., Vuong, B.Q., Vaidyanathan, B., Lin, J.-Y., Huang, F.-T., and Chaudhuri, J. (2015). Non-coding RNA Generated following Lariat Debranching Mediates Targeting of AID to DNA. Cell *161*, 762–773.

Zhuang, X., liu, Y., and Li, J. (2019). Overexpression of long noncoding RNA HOXB-AS3 indicates an unfavorable prognosis and promotes tumorigenesis in epithelial ovarian cancer via Wnt/β-catenin signaling pathway. Biosci Rep *39*.

Zieve, G., and Penman, S. (1976). Small RNA species of the HeLa cell: Metabolism and subcellular localization. Cell *8*, 19–31.

Zucchelli, S., Fasolo, F., Russo, R., Cimatti, L., Patrucco, L., Takahashi, H., Jones, M.H., Santoro, C., Sblattero, D., Cotella, D., et al. (2015). SINEUPs are modular antisense long non-coding RNAs that increase synthesis of target proteins in cells. Frontiers in Cellular Neuroscience *9*.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406–3415.

Résumé en français

1) Introduction (voir chapitres 1 et 2)

Jusqu'au début des années 2000, le dogme central de la biologie moléculaire consistait en un flux d'information allant de l'ADN, à l'ARN-messager (ARNm) puis aux protéines, comme déterminants principaux de l'identité cellulaire. Cependant, les nouvelles techniques de séquençage à haut-débit ont révélées que parmi les 3 milliards de bases qui composent le génome humain, seules 2% codent pour des protéines. En revanche, 97% du génome humain sont transcrits en ARN, dont la grande majorité provient ainsi de régions dites « non-codantes » (nc) (Djebali et al, 2012). Ces ncARN sont divisés en deux catégories en fonction de leur taille : les petits (< 200 nucléotides) et les longs (≥ 200 nucléotides) ARN non codants (lncARN). Ces lncARN sont transcrits par l'ARN Polymérase II et subissent une maturation similaire à celle des ARNm puisqu'ils sont généralement coiffés, épissés et poly-adénylés. Ils sont impliqués dans de nombreux processus biologiques (inactivation du chromosome X, marquage parental, régulation transcriptionnelle (Guttman et al, 2011 ; Rinn et al, 2012)) et agissent par de nombreux mécanismes encore peu caractérisés. Leurs profils d'expression sont spécifiques au tissu, étape de développement ou à des variations pathologiques. C'est notamment le cas du cancer où les lncARN sont fortement dérégulés (Brunner et al, 2008). Grace à leur expression hautement spécifique, ces lncARN ont été proposés comme des biomarqueurs de diagnostic et classification (Li et al, 2013) ou même comme des acteurs de la cancérogénèse (Schmitt et Chang, 2016).

Le travail présenté ici se concentre donc sur ces lncARN liés au cancer et plus spécifiquement sur leur association à la transition épithélio-mésenchymateuse (TEM). Ce processus biologique est particulièrement important dans le cancer puisque récemment admis comme étant associé à la formation de métastases. En effet, lors de la TEM les cellules perdent leur identité épithéliale (une forte cohésion entre cellules et avec la matrice extracellulaire, par exemple) et deviennent mésenchymateuses : elles changent alors de morphologie et acquièrent des propriétés accrues d'invasion et de migration cellulaire. Ces nouvelles capacités permettent aux cellules cancéreuses de migrer jusqu'au système sanguin et de se disséminer dans l'organisme, formant ainsi des tumeurs secondaires. Ces dernières réduisent fortement le taux de survie chez les patients : les métastases sont associées à 90 % des décès liés au cancer (Mehlen et

Puisieux, 2006). Récemment, le changement strict de l'identité épithéliale à mésenchymateuse a été remis en question et des états hybrides/intermédiaires ont été identifiés, avec notamment des phénotypes variables de caractère « souche », de plasticité cellulaire ou de capacité d'invasion et migration. Ces traits sont particulièrement importants dans le développement tumoral et sont associés à la résistance aux traitements, aux métastase et à la récurrence tumorale. Bien que les gènes codants impliqués dans la TEM aient été caractérisés dans le développement et plus récemment dans le cancer, le rôle des lncARN n'a que très peu été décrit. Il est cependant indéniable puisque dans les cinq dernières années, une dizaine de transcrits (HOTAIR, MALAT1, CCAT2, etc.) ont été associés à plusieurs niveaux de régulation de la TEM : épigénétique, transcriptionnelle et post-transcriptionnelle (Dhamija et Diederichs, 2016).

Durant ma thèse, j'ai étudié le rôle des lncARN dans la régulation de la TEM en identifiant les lncARN différentiellement exprimés entre les cellules épithéliales et mésenchymateuses. J'ai d'abord caractérisé le rôle du lncARN connu HOTAIR dans la TEM (chapitre 4) puis j'ai identifié de nouveaux candidats en définissant ceux qui étaient fonctionnels au travers d'un criblage génétique par CRISPR (chapitre 5). Dans ce but, j'ai utilisé un système HEK-TEM développé dans le laboratoire d'A. Londoño (voir chapitre 3.1) qui reposent sur un modèle original provenant de cellules primaires humaines d'épithélium de rein (HEK) et dont ont été tirées une lignée cellulaire épithéliale « Epi » et une lignée mésenchymateuse « Mes ».

2) Le rôle de HOTAIR dans la TEM (voir chapitre 4)

Avant mon arrivée au laboratoire, l'ancienne doctorante Claire Bertrand a identifié le lncARN connu HOTAIR comme étant surexprimé dans les cellules Mes. Comme première preuve de concept de l'étude de lncARN dans le système HEK-TEM, je me suis concentré sur le rôle de HOTAIR dans la régulation de la TEM, ce qui a amené une première publication (Jarroux et al, 2019).

Il a été montré que HOTAIR recrute des protéines de modifications de la chromatine, pour réprimer l'expression génique, grâce à des domaines structuraux situées à ses extrémités en 5' et 3'. Ces domaines interagissent avec les complexes PRC2 et Lsd1-CoREST-REST, respectivement. Nous avons d'abord défini l'importance de ces deux domaines dans la régulation de la TEM en surexprimant des variants tronqués de HOTAIR dans les cellules Epi. Nous avons montré que le domaine d'interaction avec Lsd1

est essentiel pour l'activation de la migration cellulaire, notamment au travers de la répression transcriptionnelle de gènes impliqués dans la formation d'adhésion focales et les interactions avec la matrice extracellulaire. Bien que la présence du domaine d'interaction avec PRC2 n'était pas essentiel à l'activation du phénotype migratoire, sa surexpression induit d'autres changements transcriptomiques.

Compte tenu de l'importance du domaine d'interaction avec Lsd1, nous avons souhaité examiner son recrutement sur le génome par ChIP-seq, lorsqu'HOTAIR est surexprimé. Contrairement au modèle suggéré par la littérature selon lequel HOTAIR induirait le recrutement de Lsd1 sur de nouveaux promoteurs pour réguler leur expression, nous avons observé une perte du recrutement de Lsd1 lorsque HOTAIR est surexprimé, tandis que le niveau moyen de Lsd1 reste constant. Il semblerait donc que HOTAIR promeuve la délocalisation de Lsd1 du promoteur de ses cibles habituelles, induisant ainsi une perte partielle du phénotype épithélial mais pas l'activation d'un phénotype mésenchymateux à proprement parler. En fonction du contexte cellulaire, il a été montré que Lsd1 peut induire ou réprimer la TEM et il semblerait que HOTAIR puisse fournir ce contexte, en modulant l'action de Lsd1 pour induire une TEM partielle.

3) Identification fonctionnelle de nouveaux lncARN régulant le phénotype épithélial au cours de la TEM (voir chapitre 5)

En plus d'étudier le mécanisme au travers duquel HOTAIR régule la TEM, mon projet principal a été d'identifier et caractériser de nouveaux candidats lncARN exprimés dans la TEM, et en particuliers ceux associés au phénotype mésenchymateux et leurs effets sur la régulation de l'identité épithéliale.

D'abord, j'ai souhaité décrire le transcriptome non-codant associé à la TEM en utilisant une approche d'annotation *de novo* de transcrits (Pinskaya et al, 2019) couplée à un séquençage d'ARN à haut-débit basé sur le fractionnement subcellulaire des cellules. Ainsi, j'ai séparé les ARN néotranscrits et associés à la chromatine (Chro-seq) des ARN maturés présents dans le cytoplasme (Cyto-seq). Le Cyto-seq est un bon reflet de l'expression des gènes codants tandis que le Chro-seq s'est révélé particulièrement utile pour caractériser l'expression de transcrits non-codants qui sont généralement enrichis dans le noyau. L'expression différentielle sur le Chro-seq permet également d'identifier les gènes régulés transcriptionnellement. Ensemble, ces méthodes m'ont permis d'identifier prés de 3000 transcrits non-codants différentiellement exprimés.

Ensuite, dans le but d'estimer la pertinence biologique de ces lncARN dans la régulation de la TEM, j'ai établis un crible génétique en collaboration avec le laboratoire de Neville Sanjana au New York Genome Center. Ce crible innovant repose sur l'utilisation d'une technologie dérivée de CRISPR-Cas9 pour activer l'expression de gènes de manière ciblée (CRISPRa) (voir chapitre 3.2). Jusqu'ici, la plupart des cribles CRISPRa publiées ont été basés sur des phénotypes stringents tels que la prolifération, la survie ou l'apoptose comme sélection. Or, ces phénotypes sont peu pertinents dans le système de TEM puisque les cellules Epi et Mes ont des propriétés similaires sur ces aspects. En revanche, j'ai développé deux méthodes de criblage pour mesurer la perte de l'identité épithéliale, au travers de la perte du marqueur de surface EpCAM (Epithelial Cell Adhesion Molecule) par FACS ; ou un gain de propriétés plus mésenchymateuses grâce à l'utilisation de chambre de Boyden pour séparer les cellules épithéliales pouvant traverser une matrice et membrane, sélectionnant ainsi les cellules ayant acquis des capacités accrues de migration et d'invasion.

A partir de la liste de lncARN différentiellement exprimés entre les cellules Epi et Mes, j'ai designé une banque d'ARN-guides CRISPR pour cibler le promoteur de ces gènes grâce à la machinerie CRISPRa, induisant ainsi leur expression. J'ai ensuite cloné cette banque d'ARN-guides contre ces nouveaux lncARN ainsi qu'une série de lncARN déjà annotés et des contrôles positifs et négatifs. Tandis que le crible basé sur l'invasion n'a pas fonctionné, l'utilisation de l'expression d'EpCAM comme marqueur de l'identité épithéliale s'est monté fructueuse. Ainsi, après insertion des guides CRISPRa dans les cellules Epi, j'ai isolé celles qui avaient perdu l'expression de EpCAM et donc potentiellement subi une répression (au moins partielle) de leur identité épithéliale. Ce crible a permis d'identifier le nouveau lncARN que j'ai baptisé MAL-1 (Mesenchymal identity Associated LncRNA 1) qui était le transcrit le plus différentiellement exprimé dans les cellules Mes.

Au moment de la rédaction de ce manuscrit, les validations d'activation de MAL-1 en *cis* grâce au système CRISPRa sont encore en cours. Cependant, j'ai tout de même caractérisé l'expression de MAL-1 dans le système HEK-TEM. Exprimé à partir d'un large locus dans les cellules Mes (qui est inactif dans les cellules Epi), MAL-1 est un lncARN monoexonique, poly-adénylé mais peu coiffé, qui est majoritairement localisé dans le noyau des cellules mésenchymateuses. Ce lncARN est aussi exprimé dans des lignées cellulaires d'origine fibroblastiques telles que MRC5 ou Bj-hTERT mais pas dans des lignées épithéliales. Il est également surexprimé dans certains types de carcinomes tels que dans le sein ou le rein.

Afin d'estimer si MAL-1 peut également agir en *trans* en tant que molécule d'ARN seule, je l'ai cloné et surexprimé dans les cellules Epi par vecteur lentiviral. Dans le système HEK-TEM, la surexpression de MAL-1 corrèle avec une répression de marqueurs protéiques épithéliaux tels qu'EpCAM (FACS) ou des protéines de jonctions cellulaires (Western Blot). Ces cellules ont aussi des capacités de migration cellulaire très fortement accrues et comparables à celles des cellules Mes. De plus, l'analyse transcriptomique de ces cellules a aussi montré que malgré un faible nombre de gènes différentiels, l'expression en *trans* de MAL-1 corrèle avec une activation partielle de son locus endogène sur le chromosome 6. L'ensemble de ces données suggère que MAL-1 pourrait être un nouvel exemple de lncARN capable de réguler l'expression génique pour réprimer l'identité épithéliale de cellules, en faisant un potentiel acteur du développement tumoral.

4) Conclusion (voir chapitre 6)

Dans le présent manuscrit, j'ai exploré le rôle de deux lncARN identifiés dans des cellules mésenchymateuses et les changements de phénotype qu'ils induisent dans les cellules épithéliales. Bien que HOTAIR ait été décrit comme un inducteur de la TEM, HOTAIR et MAL-1 ne semblent promouvoir qu'une TEM partielle, en réprimant des traits épithéliaux sans induire l'expression de marqueurs mésenchymateux. Des observations similaires ont déjà été rapportées pour d'autres lncARN tels que MEG3 ou H19, qui semblent pouvoir induire ou réprimer la TEM en fonction du contexte cellulaire. Or, les lncARN ont une forte spécificité d'expression et certains d'entre eux peuvent certainement être liés à des états intermédiaires de la TEM. Cela pourrait expliquer les changements subtils de phénotype lors de leur surexpression ectopique : en plus de promouvoir la TEM ou son inverse la transition mésenchymo-épithéliale, ils pourraient induire ces états intermédiaires en fonction du contexte cellulaire, ou en créant ce contexte (comme dans le cas de HOTAIR et Lsd1).

Dans l'ensemble, les lncARN semblent représenter un niveau additionnel de régulation de la TEM, comme effecteurs subtils de l'induction d'états intermédiaires. Ils pourraient ainsi coordonner l'action de facteurs de transcription ou complexes de modification de la chromatine, favorant ainsi l'hétérogénéité et la progression tumorale.

RÉSUMÉ

Ces dix dernières années, les longs ARN non-codants (lncARN) ont été un focus majeur de la recherche en biologie. Leur expression est particulièrement spécifique de l'identité cellulaire ou de variations pathologiques comme le cancer. Cependant, l'étude de leurs mécanismes dans le développement cancéreux est encore à un stage précoce. Dans ce manuscrit, je décris le rôle des lncARN et leur association à la transition épithélio-mésenchymateuse (TEM), un processus biologique lié à la métastase et la progression tumorale. D'abord, j'ai étudié le rôle du lncARN HOTAIR et en particulier son interaction avec le régulateur épigénétique Lsd1 dans la régulation de la TEM. Ensuite, je me suis concentré sur la découverte de nouveaux lncARN régulateurs et leur impact sur le phénotype de TEM en utilisant des techniques de pointe telles qu'un crible d'activation transcriptionnelle par CRISPR. Par ce biais, j'ai ainsi identifié MAL-1, un nouveau lncARN qui réprime l'identité épithélial et promeut la migration cellulaire.
Pour conclure, les résultats de ma thèse consolident le rôle clé des lncARN dans la régulation de la TEM, et particulièrement en lumière des états hybrides de la transition.

MOTS CLÉS

Longs ARN non-codants, transition épithélio-mésenchymateuse, crible CRISPR, cellules humaines, cancer, plasticité cellulaire

ABSTRACT

In the last decade, long non-coding (lnc)RNAs have been a new focus for research in biology. They are very specific to tissues, developmental stages and pathological variations like cancer. However, their functional characterization in the promotion of cancer is still in early steps. In this manuscript, I investigated the role of lncRNAs and their association to the epithelial-to-mesenchymal transition (EMT), a biological process which has been linked to metastasis and cancer progression. First I studied the role of lncRNA HOTAIR and especially its interaction with the epigenetic modifier Lsd1 in the regulation of EMT. Then, I focused on the discovery of functionally relevant new lncRNAs and their impact on the EMT phenotype using cutting-edge technologies such as CRISPR-based transcriptional activation screening. Through this, I identified the new lncRNA MAL-1 which represses epithelial identity to promote cell migration.
In conclusion, the results of my thesis consolidate the role of lncRNAs as key players in the promotion and regulation of EMT, especially in regard to hybrid states of the transition.

KEYWORDS

Long non-coding RNA, epithelial-to-mesenchymal transition, CRISPR screening, human cells, cancer, cellular plasticity