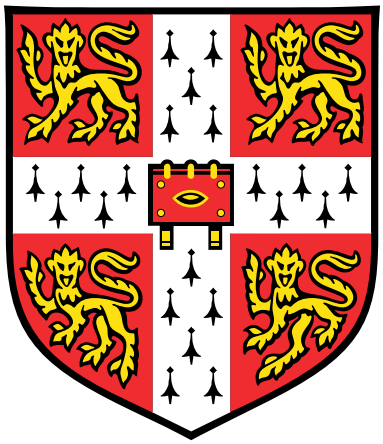
Insights into the function of non-coding RNAs

Tommaso Leonardi



Sidney Sussex College
University of Cambridge

September 2016

A dissertation submitted for the degree of Doctor of Philosophy

Tommaso Leonardi: *Insights into the function of non-coding RNAs*
A dissertation submitted for the degree of Doctor of Philosophy

The books of the great scientists are gathering dust on the shelves of learned libraries. And rightly so. The scientist addresses an infinitesimal audience of fellow composers. His message is not devoid of universality but its universality is disembodied and anonymous. While the artist's communication is linked forever with its original form, that of the scientist is modified, amplified, fused with the ideas and results of others and melts into the stream of knowledge and ideas which forms our culture. The scientist has in common with the artist only this: that he can find no better retreat from the world than his work and also no stronger link with the world than his work.

— Max Delbrück, Nobel Lecture, December 10, 1969

Summary

In the last two decades the development and wide-spread adoption of novel techniques in the field of functional genomics led to the discovery that mammalian genomes produce a large number of RNA molecules which do not encode proteins. A substantial amount of research has been devoted to the identification and characterisation of these non-coding RNAs (ncRNAs), and the picture that has emerged indicates that they represent a broad and heterogeneous group of molecules with diverse roles in the regulation of biological processes. MicroRNAs were one of the first classes of regulatory ncRNAs to be characterised in detail and it emerged that they represent a conserved family of small RNA molecules with important roles in the post-transcriptional regulation of gene expression. More recently, the class of long non-coding RNAs (lncRNAs) has gained interest among the scientific community, and several lncRNAs have been extensively characterised from a biochemical and functional perspective, revealing that they have important roles in chromatin organisation and regulation. The recent upsurge in ncRNA research coincided with a period of renewed interest in the field of extracellular vesicles. Initially considered largely independent, these two fields have come into contact following the discovery that extracellular vesicles contain ncRNAs and mediate their transport from one cell to another.

This thesis will report on three projects that share the common underlying aim of providing new insights into the function of ncRNAs and their role in cell-to-cell communication. The first chapter describes the identification of positionally conserved lncRNAs (pcRNAs), a class of lncRNAs with a conserved genomic position relative to orthologous neighbouring coding genes. pcRNAs are associated with developmental transcription factors and are co-expressed with them, displaying high tissue specificity. Interestingly, over half of the pcRNAs overlap binding sites for the CTCF chromatin organiser and reside on the boundaries of topological anchor points. Further characterisation of these topological anchor point RNAs revealed that they often regulate the expression of the associated coding genes and have similar effects on the metastatic phenotypes of cancer cell lines.

The second chapter explores the process of cell-to-cell communication via the exchange of extracellular vesicles (EVs). We characterised EVs secreted by murine Neural Progenitor Cells (NPCs), finding that they transfer mRNAs and proteins in response to inflammatory cues. Stimulation of NPCs with pro-

inflammatory cytokines induces the secretion within EVs of mRNA and protein components of the IFN-ɣ signalling pathway. IFN-ɣ binds its receptor on the surface of EVs, and is capable of activating an inflammatory response in target cells. These data indicate a novel mechanism by which cells can propagate the activation of a signalling pathway at a distance, highlighting a new level of interaction between stem cells and the immune system.

Lastly, the third chapter focuses on microRNAs in EV-mediated cell-to-cell communication. A specific subset of microRNAs expressed by NPCs is enriched inside EVs, suggesting the existence of an active secretion mechanism for microRNA trafficking. Indeed, the analysis of the sequence of secreted microRNAs revealed the presence of short motifs that act as putative binding sites for carrier proteins. These results shed light on the molecular mechanism responsible for active microRNA secretion in stem cells.

This work contributes towards a better understanding of a still largely uncharacterised fraction of the non-coding genome, suggesting that ncRNAs have important roles in the topological organisation of chromatin and in cell-to-cell communication. Future studies in this direction will reveal the full extent of non-coding RNA function inside the cell and between different cells in an organism.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

This dissertation does not exceed the prescribed word limit of the Degree Committee for the Faculty of Biology

Contents

CAGE	Cap Analysis Gene Expression
CDS	Coding DNA Sequence
CNS	Central Nervous System
CPAT	Coding Potential Assessment Tool
DC	Dendritic Cell
DHS	DNaseI Hypersensitivity Site
DLS	Dynamic Light Scattering
dsRBD	double strand RNA Binding Domain
ECS	Evolutionarily Conserved Structure
ELISA	Enzyme-Linked Immunosorbent Assay
ESCRT	Endosomal Sorting Complex Required for Transport
EV	Extracellular Vesicle
EXO	Exosome
fEGFP	farnesylated Enhanced Green Fluorescent Protein
FISH	Fluorescence *in situ* hybridization
GIS	Gene Identification Signature
GO	Gene Ontology
HGP	Human Genome Project
ICR	Imprinting Control Region
ILV	Intraluminal vesicle
IRES	Internal Ribosome Entry Site
KO	Knock Out
lincRNA	long intervening noncoding RNA
lncRNA	long noncoding RNA
miRNA	microRNA
MVB	Multi Vesicular Body
NAT	Natural Antisense Transcript
NLS	Nuclear Localisation Signal
NPC	Neural Progenitor Cell
NTA	Nanoparticle Tracking Analysis
ORF	Open Reading Frame
RBP	RNA Binding Protein
RFP	Red Fluorescent Protein
RISC	RNA-induced Silencing Complex
RNase P	Ribonuclease P

SEM	Scanning Electron Microscopy
SILAC	Stable Isotope Labelling of Aminoacids in Cell culture
STED	Stimulated Emission Depletion microscopy
SVZ	Subventricular Zone
TAD	Topologically Associating Domain
TEM	Transmission Electron Microscopy
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TSS	Transcriptional Start Site
UCE	Ultraconserved Element
UTR	Unstranslated Region
WB	Western Blot
WT	Wilde Type
XCI	X Chromosome Inactivation
Xi	Inactive X Chromosome
XIC	X Inactivation Center

Part I

Foreword

The completion of the Human Genome Project (HGP) in 2001 marked the beginning of a new era in every biological field. One of the very significant results stemming from the HGP has been the rediscovery that the vast majority of the genome is pervasively transcribed into tens of thousands of non-coding transcripts, a fact that had been observed in the 70's and neglected for several decades (Carninci et al., 2005). The recent development of high-throughput RNA sequencing methods led to an enormous surge in the discovery of non-coding RNAs and recent efforts to annotate the human transcriptome estimate that a multitude of non-coding transcripts are produced across the diversity of human tissues.

Among regulatory non-coding RNAs, microRNAs were one of the first classes to be discovered and extensively characterised leading to a precise understanding of their genomic characteristics, biogenesis and functions (Lee et al., 1993; Wightman et al., 1993). The discovery in the early 2000s that miRNAs control mRNA and protein levels brought popularity to the early concept that non-coding RNAs have widespread regulatory functions. Since then, numerous other classes of regulatory RNAs have been discovered and characterised. Most notably, long non-coding RNAs (lncRNAs) have recently gained attention and the scientific community has devoted substantial efforts toward identifying, cataloguing and characterising them. The function of several lncRNAs has now been extensively characterised, and many of them are emerging as key players in the regulation of a broad and diverse set of biological processes (Amaral et al., 2008).

In the same span of time the field of secreted membrane vesicles has witnessed a similar spark of renewed interest. Following the discovery in 2007 that exosomes are capable of transferring mRNAs and miRNAs between cells (Valadi et al., 2007), the field has seen a substantial growth and various efforts have been made toward establishing their roles and functions. Exosomes and EVs have been detected in virtually every tissue and biological fluid, and their functions have been shown to span from the regulation of immune processes to the spread of cancer (Colombo et al., 2014).

The work that I carried out during my doctorate tried to address some open questions in the field of functional genomics, with the underlying common aim of providing novel insights toward the function of non-coding RNAs and their roles in cell-to-cell communication. In this thesis I will provide a general introduction to non-coding RNAs and extracellular vesicles and describe my contribution to both fields.

Part II

Introduction

1

The biology of ncRNAs

Recent results from the GENCODE project (Harrow et al., 2012) show that the human genome is transcribed into ~200 000 transcripts, of which only ~80 000 are protein coding. These results are in line with numerous observations that date back to the early days of molecular biology. In fact, it has been clear since the '60s and '70s that the majority of the genome does not code for protein-coding genes (O'Brien, 1973), although a large portion of it is transcribed into RNA (Comings, 1972). In light of these findings, various early works proposed that non coding transcripts might have a function on their own (Britten and Davidson, 1969; Edelman and Gally, 1970; Holliday, 1970; Orgel and Crick, 1980) and several of these hypothesis have now been proven correct.

1.1 lncRNAs

In the last decade, the amount of research aimed at deciphering the roles and functions of lncRNAs has dramatically increased (**Figure 1.1**). This effort led to a more comprehensive annotation of their genomic locations and features as well as to a better understanding of their role in a variety of biological processes, which span from the regulation of embryonic development to pathological conditions such as cancer.

However, despite the vast research effort invested in lncRNAs and our ever-increasing understanding of their functions, we are still lacking both the capacity to infer their functions from the sequence as well as the ability to divide them into categories of common functionality. With these limitations in mind, I believe it is reasonable to assume that the current definition of lncRNAs – a definition that is usually based solely on the lack of coding potential and the length of the RNA – includes in reality a broad set of molecules with extremely different biochemical properties and functions. For these reasons, any generic introduction to lncRNAs has numerous intrinsic limitations and will likely be quickly outdated with the progression of the field. Nevertheless, in this chapter I will try to give a general introduction on what is currently known on lncRNAs and introduce the repertoire of their possible functions by describing well characterised examples.

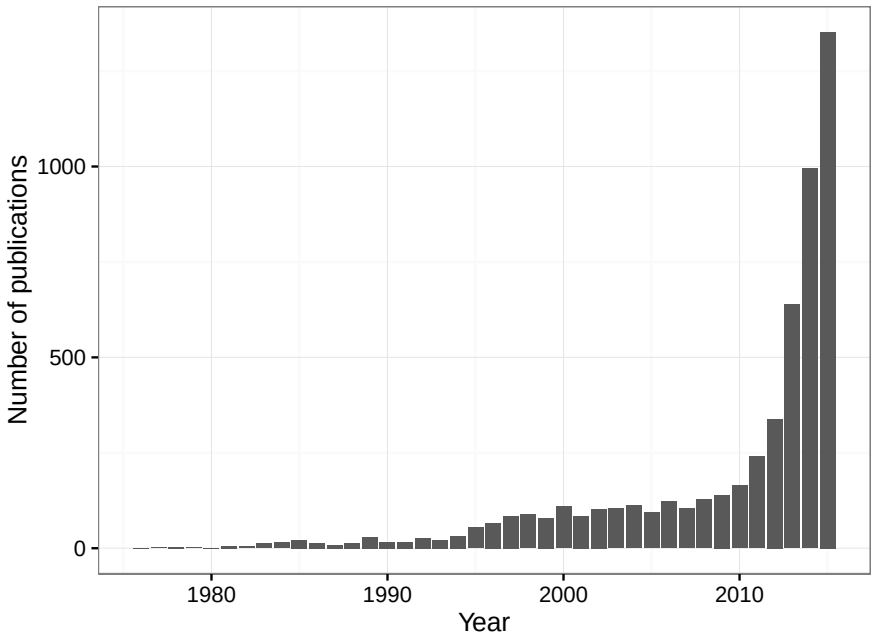


Figure 1.1 *Number of Pubmed results for the keyword "lncRNA". Data accessed on 21st July 2016. The incomplete data for 2016 were omitted from the plot*

1.1.1 *Definition and identification of lncRNAs*

One of the first insights into the transcriptional complexity of mammalian genomes was provided in 2005 by the FANTOM3 project, which used Cap Analysis Gene Expression (CAGE) and Gene Identification Signature (GIS) to precisely map the 5′ and 3′ ends of transcripts from 237 full length cDNA libraries prepared from a broad collection of mouse cells and tissues (Carninci et al., 2005). This work revealed at an unprecedented scale that the majority of the genome is transcribed from both strands, and that the majority of genes are subject to alternative splicing and produce multiple transcripts, many of which lack coding potential. Several other works have subsequently tackled the problem of annotating mammalian lncRNAs, producing increasingly broad and complex catalogues (Harrow et al., 2006; Kapranov et al., 2007; Cabili et al., 2011; Harrow et al., 2012; Guttman et al., 2010; Amaral et al., 2011; Jia et al., 2010; Khalil et al., 2009). In fact, the last release of data from the GENCODE project (version 25; Harrow et al., 2012), annotates in the human genome 27 692 lncRNA transcripts and this estimate seems bound to increase with novel technical advances, such as captureSeq (Clark et al., 2015; Bussotti et al., 2016), that allow the identification of transcripts with low expression. These works resulted in an ever increasing, complex picture of the mammalian transcriptome, where the majority of the genome is transcribed from both strands

to produce intricate networks of often overlapping transcripts, the majority of which lack coding potential.

There are no standard criteria for the identification of lncRNAs, and most works adopt somewhat arbitrary thresholds based on transcript size and lack of Open Reading Frame (ORF) to distinguish them from mRNAs and other classes of small ncRNAs. Typically, lncRNAs are defined as longer than 200nt (Kapranov et al., 2007) with either no or very short ORFs. The most widely used size threshold for operationally defining lncRNAs is 200nt, which serves well to exclude the majority of well known classes of small ncRNAs. In terms of ORFs size, the FANTOM project initially used a cut-off of 300nt (i.e. 100 codons) to distinguish lncRNAs from mRNAs (Okazaki et al., 2002). This arbitrary threshold was based on the observation that the majority of proteins in Swiss-Prot and the International Protein Index are longer than 100 amino acids. This length is also conveniently ~2 standard deviations above the mean ORF size in 1000nt of random sequence, and was already chosen as a size threshold for the identification of protein coding genes during the sequencing of the yeast genome (Oliver et al., 1992). However, these criteria, although very practical and easy to apply, are subject to numerous false positive and false negative classification errors. For example, the murine *XIST* lncRNA is approximately 15kb in size and contains an ORF of 298 amino acids, which mistakenly led to its classification as a protein coding gene (Borsani et al., 1991; Dinger et al., 2008a). Analogously, relying exclusively on these thresholds would misclassify genes that code for small proteins shorter than 100 amino acids, such as the hormone peptide HEPCIDIN (84aa). Various approaches can be applied to mitigate these problems. For example, a common strategy is to align the peptide sequence encoded by the ORFs of the putative lncRNAs against a database of know protein and/or domain sequences. Other methods, such as Phylogenetic Codon Substitution Frequencies (PhyloCSF), examine multiple phylogenetic alignments of the putative lncRNAs in order to identify characteristics typical of protein coding genes, such as high frequency of synonymous substitutions and low frequency of missense or non-sense substitutions (Lin et al., 2011). However, these methods rely on multi-species alignments, which are often hard to obtain for poorly conserved or lineage specific lncRNAs, and are biased by the fact that a large fraction of lncRNAs overlap either in the sense or antisense orientation with isoforms of coding genes, thus skewing the conservation results. More recent methods to assess the coding potential of RNAs use machine learning algorithms to automatically discern features that separate coding genes from lncRNAs. For example, Coding Potential Assessment Tool (CPAT) estimates four features of an RNA sequence (ORF size, ORF coverage, Fickett TESTCODE statistic and hexamer usage bias) and trains a logistic re-

gression on known mRNAs and lncRNAs (Wang et al., 2013). These methods have been shown to have high sensitivity and specificity and to be computationally efficient, however they rely on a training dataset of known lncRNAs which could skew the results in favour of the methods used to generate them. The task of separating mRNAs from lncRNAs is further complicated by the fact that there isn't necessarily a clear distinction between the two classes. In fact, several well-known RNAs function both as a non-coding RNA molecule as well as an mRNA which is translated into a functional protein (Dinger et al., 2011). For example, the steroid receptor RNA activator (*SRA*) is a lncRNA that increases the activity of the steroid receptors in activating target genes. However, *SRA* was also shown to be translated both *in vitro* and *in vivo* into a functional protein which is conserved among vertebrates (Chooniedass-Kothari et al., 2004).

From an experimental perspective, there are various techniques to investigate the coding potential of an RNA molecule. For example, *in vitro* translation assays allow to determine whether a putative protein-coding RNA is translated into a polypeptide, whereas ribosome profiling allows to identify RNAs which are bound to ribosomes (Ingolia et al., 2014; Guttman et al., 2013). These techniques, although very useful at providing an indication on the coding/non-coding nature of an RNA molecule, often do not provide a final answer. In fact, there is evidence that spurious ORFs can be translated *in vitro* (Dinger et al., 2008a) and that the majority of cytoplasmic lncRNAs bind to ribosomes in human cells (Carlevaro-Fita et al., 2016).

In conclusion, the task of defining and annotating lncRNAs is complex and suffers from the lack of exclusive defining criteria. The methods presented above only provide an estimate of the likelihood that an RNA sequence is coding or non-coding, while such a dichotomous distinction might have little biological relevance. A more realistic view might in fact be that RNAs with an exclusively coding or non-coding function are only the two extremes of a continuous spectrum of functionality where each gene independently evolves functions at both the protein and RNA level (Mercer et al., 2009). For these reasons, definitive answers on the coding/non-coding nature of RNAs can only be obtained by assessing them experimentally on a case by case basis.

1.1.2 *Transcription and processing of lncRNAs*

Recent efforts to characterise the mammalian transcriptome led to a substantial revision in our understanding of transcription. It is now apparent that the vast majority of genes are embedded in complex transcriptional loci that produce myriads of isoforms, many of which are non-coding (Gerstein et al.,

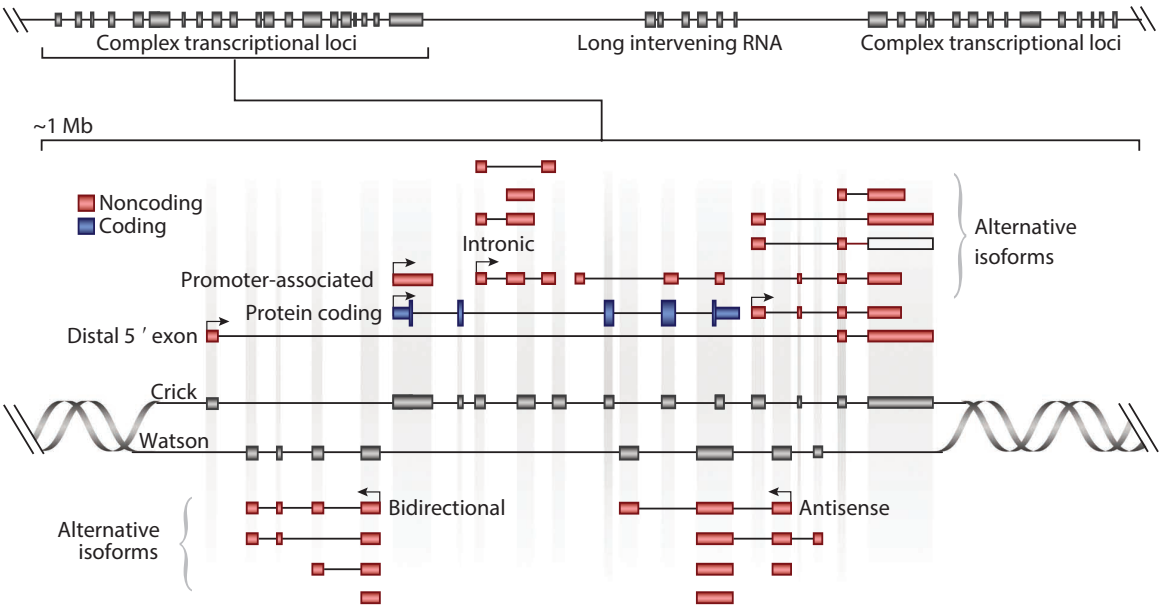


Figure 1.2 *Schematic representation of complex transcriptional loci, where protein coding transcripts (in blue) are intertwined with numerous non-coding transcripts. Adapted from Mercer and Mattick, 2013.*

2007). In fact, most lncRNAs are found in close proximity to protein coding genes (Bertone et al., 2004; Ponjavic et al., 2007) and their genomic architectures are often intertwined (**Figure 1.2**). These complex protein coding loci are often characterised by the presence of antisense and bidirectional non-coding transcripts (Katayama et al., 2005), as well as sense transcripts with intronic and/or exonic overlaps (Mercer and Mattick, 2013; Kapranov et al., 2005). At the same time, numerous lncRNAs are transcribed at independent loci devoid of protein-coding genes and are named long intervening noncoding RNAs (lincRNAs) (Guttman et al., 2009).

The majority of lncRNAs are transcribed by RNA Polymerase II, therefore possessing a 5′ methylguanosine cap, a 3′ polyA tail as well as histone modifications typical of canonical RNA Polymerase II transcripts, such as trimethylation of the lysine 4 of histone H3 (H3K4me3) at the Transcriptional Start Site (TSS) and trimethylation of the lysine 36 of histone H3 (H3K36me3) along the transcript body (Guttman et al., 2009; Rinn and Chang, 2012). Most lncRNAs are multi-exonic and subject to alternative splicing, but they tend to have fewer exons than mRNAs (Derrien et al., 2012), although this might be an artefact due to their low expression and the consequent increased difficulty for full-length assembly (Bussotti et al., 2016). The genomic features described above reflect the characteristics of typical lncRNAs, however there are notable exceptions to most of these rules. For example, it was recently shown that certain lncRNAs are post-transcriptionally circularised, either by back-splicing the 3′ end to an upstream 5′ exon (circRNAs, Salzman et al., 2012; Memczak et al.,

2013) or by stabilisation of the lariat loop formed during canonical intron splicing (circular intronic RNAs, Zhang et al., 2013). These families of lncRNAs lack polyA tails and are therefore excluded in the majority RNA-Seq studies, which select – prior to sequencing – only polyadenylated RNAs. Similarly to circular RNAs, also ncRNAs transcribed by RNA Polymerase III lack a polyA tail and represent a class in which only a few members are characterised (Dieci et al., 2007). For example, the mitochondrial RNA processing endoribonuclease (*RMRP*) is a lncRNA transcribed by RNA polymerase III with important roles in stem cell and immune cell biology (Maida et al., 2009; Huang et al., 2015). Circular lncRNAs and lncRNAs transcribed by RNA Polymerase III are not the only examples of lncRNAs that lack a polyA tail. Other lncRNAs, in fact, undergo an alternative process of 3′ end maturation which is catalysed by Ribonuclease P (RNase P). RNase P is a ribonucleoprotein with ribonuclease activity responsible for the processing and maturation of precursor tRNAs (Guerrier-Takada et al., 1983). However, it was recently found that RNase P can also process the 3′ end of some lncRNAs, such as *MALAT1* and *NEAT1*. Both these lncRNAs possess at their 3′ end a triple helical structure as well as tRNA-like structures. The tRNA-like structures are recognised and cleaved by RNase P, while the triple helix efficiently stabilises the cleaved lncRNA despite the lack of a polyA tail (Wilusz et al., 2008; Brown et al., 2012; Wilusz et al., 2012).

1.1.3 *Sequence and structure conservation of lncRNAs*

Sequence conservation is the hallmark of purifying selection and is a major indicator of functionality. Numerous studies have found that lncRNAs display significantly higher conservation than neutrally evolving intergenic sequences, although to a lower extent than the exons of protein coding genes (Carninci et al., 2005; Guttman et al., 2009; Derrien et al., 2012; Kutter et al., 2012; Marques and Ponting, 2009; Iyer et al., 2015). Generally, the exons of lncRNAs have sequence conservation comparable to that of the Unstranslated Regions (UTRs) of coding genes (Carninci et al., 2005); on the other hand, their promoters tend to display higher (Carninci et al., 2005) or similar (Guttman et al., 2009; Derrien et al., 2012; Necsulea et al., 2014) conservation to promoters of protein-coding genes, and are often enriched in binding sites for transcription factors. Although the low exonic conservation of lncRNAs is a norm, it is by no means the rule: in fact, Iyer et al. (2015) recently reported the identification of 597 intergenic lncRNAs that harbour Ultraconserved Elements (UCEs), defined as regions of DNA longer than 200nt with almost perfect conservation across multiple species (Bejerano et al., 2004).

Recent phylogenetic analysis of lncRNA evolution in tetrapods revealed that over 80% of lncRNAs are primate specific (Necsulea et al., 2014). Interestingly, it was also found that evolutionarily young lncRNAs, despite having low levels of exonic conservation, display signs of purifying selection, suggesting that a fraction of them might have acquired a function in recent times. A recent work by Hezroni et al. (2015) further investigated the evolutionary history of lncRNAs, confirming that the majority of them (>70%) do not have sequence orthologues in species separated by more than 50 million years. Interestingly, despite the modest conservation of lncRNA exonic sequences, numerous works have found that many of them are localised in syntenic regions and display a conserved position relative to neighbouring orthologous coding genes (Carninci et al., 2005; Engstrom et al., 2006; Lipovich et al., 2006; Dinger et al., 2008b; Ulitsky et al., 2011; Necsulea et al., 2014; Hezroni et al., 2015). This finding suggests that many lncRNAs might have a function independent of their sequence; indeed, several lncRNAs with positional conservation have been functionally characterised and in some cases shown to regulate the neighbouring protein coding genes (Dallosso et al., 2007; Feng et al., 2006; Amaral et al., 2009; Wang et al., 2011; Bell et al., 2016).

Taken together, these findings highlight the fact that large scale sequence conservation might not be the best indicator of lncRNA function. To overcome this limitation various studies have explored the possibility that other features of lncRNAs might be the hallmarks of their functionality. For example, it was recently observed that syntenic lncRNAs possess short patches of conserved sequences and are significantly enriched in specific sequence motifs (Hezroni et al., 2015). Some of these motifs were shown to represent binding sites for splicing factors or splicing enhancers (Schüler et al., 2014; Haerty and Ponting, 2015), while others might represent binding motifs for diverse transcription factors and/or RNA binding proteins. These findings provide an evolutionary basis for earlier observations showing that short domains of lncRNAs might be sufficient for their function (Chureau et al., 2002; Ulitsky et al., 2011; Quinn et al., 2014).

Structure is an additional important factor to consider when assessing the functionality of lncRNAs, as they often fold into complex and thermodynamically stable secondary and tertiary structures that are important for their functions (Zhang et al., 2010; Kertesz et al., 2010; Mercer and Mattick, 2013). For this reason, mutations that alter the primary sequence of an RNA but preserve base pairing (an event known as covariation) decrease the sequence conservation without being negatively selected (Washietl et al., 2005). Smith et al. (2013) recently screened mammalian genomes for evolutionarily constrained RNA structures, finding millions of genomic loci that undergo strong puri-

fying selection at the structural level but are not constrained at the sequence level. Moreover, it was also shown that lncRNAs are enriched in Evolutionarily Conserved Structures (ECSs) compared to intergenic regions, although to a lesser extent than protein coding genes (Smith et al., 2013). For example, the aforementioned secondary structures that appear in the 3′ end of *MALAT1* and *NEAT1* – which mediate their processing by RNase P – are evolutionarily conserved at the structural level (Smith et al., 2013).

Various recent works shed further light on the relationship between RNA structure and function. Xue et al. (2016) showed that an internal G-rich structural motif in the murine *BRAVEHEART* lncRNA is required for cardiomyocyte differentiation, as it binds to and antagonises the zinc-finger protein CNBP. Similarly, it was also found that the *COOLAIR* lncRNA is structurally constrained in several species of the *Brassicaceae* family, and inter-species variation in the length of one of its structural domains is linked to trait variation (Hawkes et al., 2016). In another study, Quinn et al. (2016) showed that in *D. Melanogaster* an engineered transgene carrying one or more copies of the conserved roXbox stem–loop motif rescues in a dose-dependent manner the phenotype of roX-null males *in vivo*. Taken together, these works provide strong experimental evidence that, at least in some cases, the secondary structure of a lncRNA is at the basis of its function.

1.1.4 *Roles and functions of lncRNAs*

One of the major problems faced in the identification and functional analysis of lncRNAs is that they are typically expressed at low levels and in a very tissue specific manner (Cabili et al., 2011), although their expression profiles are usually conserved across species (Chodroff et al., 2010; Necsulea et al., 2014; Washietl et al., 2014). One of the proposed explanations for the low expression levels measured is that lncRNAs are expressed in a restricted and specific manner by only a small number of tissues and/or cell types within a tissue (Mercer et al., 2008; Dinger et al., 2009), thus resulting in a low apparent level of expression when measured in a tissue or a whole organism. Moreover, numerous lncRNAs appear to be only expressed in very specific conditions, such as precise developmental time points (Zhang et al., 2014; Amaral and Mattick, 2008) or in response to stress and other external cues (reviewed in Amaral et al., 2013). In a recent study Cabili et al. (2015) measured the expression and subcellular localisation of 61 lncRNAs by single cell RNA Fluorescence *in situ* hybridization (FISH), reporting very precise sub-cellular expression patterns within a cell and homogeneity in expression across different cells, therefore dismissing the hypothesis that lncRNA expression could be spatio-temporally restricted

to specific cells. However, this study only analysed a small number of lncRNAs and in only three immortalised human cell lines, therefore ignoring the subtle complexities of expression in more physiological contexts. In contrast, a recent study employed single cell RNA-Seq to analyse expression profiles in the human neocortex, and found that numerous lncRNAs are specifically expressed in distinct cell types and are abundantly expressed in individual cells (Liu et al., 2016).

These studies on the cellular and sub-cellular patterns of lncRNA expression provide important insights toward defining their functions. Numerous early efforts to attribute a function to individual lncRNAs used their coexpression with protein coding genes as an indicator of functional commonality (Dinger et al., 2008b; Guttman et al., 2009). This "guilt by association" approach is still widely used today, and it recently led to interesting functional insights for numerous lncRNAs. For example, a recent study of lncRNAs in tetrapods found that numerous co-expression clusters are evolutionarily conserved, and identified lncRNAs potentially involved in processes such as spermatogenesis, synaptic transmission and placental development (Necsulea et al., 2014).

These genome-wide approaches to unravel the function of lncRNAs proved to be very effective at suggesting their possible roles. However, the last few years have seen a dramatic increase in the number of studies that dissected the functions of individual lncRNAs, attributing them precise roles from a biochemical and mechanistic point of view. The following paragraphs will describe specific examples of well characterised lncRNAs in an attempt to systematically summarise their range of functions. For the sake of simplicity, I will separately address lncRNAs with a function *in cis* and those with a function *in trans*. This distinction, albeit practical for academic purposes, is purely artificial: the boundaries between cis and trans functions are very blurred, lacking precise definitions and making little sense when considering the genome in its three dimensional architecture.

1.1.4.1 *Cis-acting lncRNAs*

Long non-coding RNAs involved in dosage compensation and imprinting were among the first to be characterised, and we now have a reasonably thorough understanding of their mechanisms of action.

Cis-acting lncRNAs in dosage compensation Dosage compensation is a process that balances gene expression from the sexual chromosomes of female and male cells. The mechanism of dosage compensation in placental mammals was first proposed in 1961 by Mary Lyon, who suggested that one of the two X chromosomes in female cells is randomly silenced during devel-

opment (Lyon, 1961). Since this early discovery, numerous works have characterised the process of X Chromosome Inactivation (XCI) (reviewed in Cerase et al., 2015). The genomic locus responsible for starting and sustaining this process is named the X Inactivation Center (XIC), which includes the 17kb, nuclear, spliced and capped lncRNA *XIST* (Brockdorff et al., 1991; Brockdorff et al., 1992; Penny et al., 1996; Wutz and Jaenisch, 2000). *XIST* is transcribed from the X chromosome that will undergo inactivation and spreads *in cis* to initiate and maintain the cascade of events that will lead to XCI. Following its upregulation, *XIST* initially makes contact with a limited number of gene-rich loci which are in topological proximity with the XIC (Simon et al., 2013; Engreitz et al., 2013) and subsequently spreads to the rest of the Inactive X Chromosome (Xi). In order to mediate the inactivation of the X chromosome, *XIST* either directly interacts with polycomb repressive complex 2 (PRC2) through the structurally conserved domain RapA (Zhao et al., 2008; Kaneko et al., 2010; Maenner et al., 2010; Kanhere et al., 2010) or it recruits it indirectly (Okamoto et al., 2004; Mak et al., 2004; Kohlmaier et al., 2004; Chaumeil et al., 2006) by first silencing transcription (McHugh et al., 2015) and/or by binding other proteins that directly interact with *PRC2* (Chu et al., 2015). *PRC2* is a histone methyltransferase that mediates the trimethylation of histone H3 on lysine 27 (H3K27me3), a repressive chromatin mark that leads to the silencing of the Xi. Interestingly, the *XIST* locus harbours another cis-acting lncRNA with an important role in XCI. *TSIX* is the antisense transcript of *XIST*, and its transcription mediates the repression of *XIST in cis* on the active X chromosome. Its mechanism of action is still debated, and there is contrasting evidence suggesting that either the *TSIX* RNA itself mediates *XIST* silencing or that it is *TSIX* transcription that interferes with *XIST* expression (reviewed in Augui et al., 2011). Interestingly, in a recent work Chu et al. (2015) developed a novel method for the comprehensive identification of RNA binding proteins by mass spectrometry (ChIRP-MS). This work characterised at an unprecedented scale the interactome of *XIST*, proposing a model in which *XIST* acts as a scaffold to recruit and organise two chromatin modifying complexes: the polycomb repressive complex, which deposits silencing marks, and SPEN, which in complex with MBD3-NURD favours the removal of activating marks such as histone acetylation (Chu et al., 2015). In an independent work, Minajigi et al. (2015) developed a similar proteomics method called iDRiP (identification of direct RNA-interacting proteins), that allowed to identify the binding partners of *XIST*. This study revealed that the *XIST* interactome is composed of 80 to 200 proteins that fall into several functional categories, among which cohesins were among the highest confidence hits. Interestingly, further experiments demonstrated that *XIST* repels the cohesins SMC1A and RAD21 from

the Xi, thus promoting the acquisition of chromatin topology not favourable to transcription (Minajigi et al., 2015). Concomitantly, McHugh et al. (2015) developed a different quantitative proteomics technique that allowed the identification of proteins that interact with *XIST*. This work revealed that *XIST* directly interacts with the Class E basic helix-loop-helix protein 41 (SHARP), which in turn interacts with an activator of the histone deacetylase HDAC3 thus leading to the exclusion of RNA Polymerase II from the Xi (McHugh et al., 2015).

Cis-acting lncRNAs and imprinting The process of imprinting is a particularly suitable context to demonstrate the functions of lncRNAs acting *in cis*. Several lncRNAs have been shown to be involved in imprinting well before the recent surge in ncRNA research, and they are still among the best characterised examples. Imprinting was discovered in 1984 following the observation that pronuclear transplantation to produce a diploid genome from two male or two female cells fails to produce an embryo that undergoes normal development (McGrath and Solter, 1984; Surani et al., 1984). It was then observed that certain regions of the genome were differentially active between paternally and maternally derived chromosomes (Cattanach and Kirk, 1985), and that defects in this process were involved in human genetic disorders such as Prader-Willi syndrome (Nicholls et al., 1989). Numerous later studies have extensively characterised this process, finding that certain genomic loci are expressed mono-allelically and in a parent-of-origin specific way (reviewed in Adalsteinsson and Ferguson-Smith, 2014 and Peters, 2014). To date, 151 imprinted genes have been identified in the mouse genome (MouseBook database, Williamson et al., 2013) and they tend to be located in clusters. The expression of the genes in each cluster is controlled *in cis* by regions defined as Imprinting Control Regions (ICRs), which undergo allele-specific differential methylation according to their parent of origin. Each imprinted cluster contains at least one lncRNA which often has important roles in the establishment and maintenance of the allele-specific expression *in cis*. A clear example to illustrate both the fundamentals of imprinting and the role on lncRNAs in this process is provided by the *IGF2R/AIRN* locus. The Insulin-like growth factor 2 receptor (*IGF2R*) is a protein coding gene located in a maternally expressed imprinted cluster together with the genes *SLC22A2* and *SLC22A3* (Sleutels et al., 2002). This locus also encodes a lncRNA named *AIR*, which is an unspliced, nuclear RNA of 108 kb in length transcribed from an antisense promoter located in the second intron of *IGF2R* (Lyle et al., 2000). *AIR* is also imprinted, but unlike the rest of the cluster it is exclusively expressed from the paternal allele, whereas its maternal copy is silenced by promoter hypermethylation. It

was shown by DNA-RNA FISH that the *AIR* RNA localises in proximity of its own locus and is in contact with the distal protein coding genes of the cluster (Nagano et al., 2008), suggesting a cis-acting function. In fact, further studies have confirmed that the premature termination of *AIR* by the insertion of a polyadenylation signal or the deletion of its promoter result in the reactivation of the paternal allele (Sleutels et al., 2002). Moreover, it was shown that *AIR* RNA accumulates at the promoter of *SLC22A3* and recruits the Histone-lysine N-methyltransferase (EHMT2, also known as G9A), which in turn mediates the trimethylation of histone H3 on lysine 9 (H3K9me3) and mono-allelic silencing. *AIR* does not achieve silencing of *IGF2R* by the same mechanism of epigenetic regulation; instead, the act of transcription of *AIR* is sufficient to interfere with the transcription of the overlapping gene *IGF2R*, thus demonstrating a further function for the *AIR* locus independent of its RNA product (Latos et al., 2012).

Enhancer RNAs Unlike the lncRNAs involved in imprinting and dosage compensation, numerous lncRNAs mediate the activation of neighbouring genes or transcripts. Several studies have attempted to identify and categorise such ncRNAs and have defined often overlapping groups of lncRNAs based on their genomic characteristics and functions. One of these groups predominantly consists of bidirectional, unspliced and non-polyadenylated ncRNAs (Koch et al., 2011) transcribed from enhancer regions and marked by high levels of monomethylated histone H3 at lysine 4 (H3K4me1) and low levels of H3K4me3 (reviewed in Lam et al., 2014). These lncRNAs are termed enhancer RNAs (eRNAs) and their expression was shown to correlate with that of neighbouring genes (Kim et al., 2010; De Santa et al., 2010). Interestingly, some eRNAs were shown to directly regulate the expression of neighbouring genes *in cis*. For example, a recent work found that an enhancer downstream of the *DHRS4* locus produces and eRNA named AS1eRNA that promotes the physical interaction of the enhancer with the protein-coding locus, mediating its transcription in cooperation with RNA polymerase II and the P300/CBP transcriptional coactivators (Yang et al., 2016). Similarly, an earlier study showed that a bidirectional eRNA at the KLK3e locus promotes physical interaction of an enhancer upstream of the *KLK3* gene with the protein-coding gene *KLK2*, in an androgen receptor-mediated manner (Hsieh et al., 2014). Additionally, the oestrogen receptor α (ER-α) was shown to upregulate the production of eRNAs from enhancers adjacent to its target genes (Li et al., 2013b), and the authors proposed that these eRNAs potentiate the looping interaction between enhancers and target gene promoters. These examples might highlight a common theme where eRNAs regulate neighbouring genes through the formation

of higher order chromatin structures. However, there is little evidence confirming how widespread this mechanism might be, and a recent study based on single molecule fluorescence *in situ* hybridisation (smFISH) revealed that eRNA accumulation at enhancers is not required to promote the transcription of the target gene (Rahman et al., 2016). On the other hand, there is also evidence of other eRNAs functioning in different ways. For example, the eRNA produced by the enhancer of Activity-regulated cytoskeletal protein (*ARC*) was shown to act as decoy for the negative elongation factor (*NELF*) and thus promotes transition of RNA polymerase II to the elongation phase (Schaukowitch et al., 2014).

Other activating lncRNAs There are several other characterised lncRNAs with activating functions that do not fall into previous categories. One of the best known examples is provided by the mechanism of regulation of the *HOXA* locus, a cluster of homeotic genes collinearly expressed along the cranio-caudal axis during embryonic development. At the 5′ and 3′ end of the *HOXA* locus there are two lncRNAs, respectively named *HOTTIP* and *HOTAIRM1*. Initially identified in relation to myelopoiesis, *HOTAIRM1* regulates HOX genes *in cis* (Zhang et al., 2009); on the other hand, *HOTTIP* was found to be an activating lncRNA (Wang et al., 2011). *HOTTIP* is a spliced and polyadenylated transcript of 3.7kb with a TSS ~330bp upstream of *HOXA13*. The expression of *HOTTIP* mirrors that of genes at the 5′ of the *HOXA* locus, with higher expression levels found in caudal anatomical districts. Consistently, the *HOTTIP* locus was found to be enriched in the bivalent chromatin marks H3K4me3 and H3K27me3, typical of poised genes, in cranial regions, while enriched in only H3K4me3 in caudal regions (Wang et al., 2011). Chromosome conformation capture experiments have shown that the *HOTTIP* locus resides in spatial proximity to the genes at the 5′ of the *HOXA* cluster, and the *HOTTIP* lncRNA directly binds to WD repeat-containing protein 5 (WDR5), an anchor protein that in turn interacts with the MLL complex. The mixed-lineage leukemia (MLL) proteins are a family of SET-domain-containing lysine methyltransferases with important roles in the activation of the *HOX* genes (Wang et al., 2009); *HOTTIP*, through the anchor protein WD5, mediates the recruitment of MLL at the 5′ end of the *HOXA* locus, thus promoting the formation of a broad domain of H3K4me3 and transcriptional activation (Wang et al., 2011) of *HOXA* genes. This example offers an interesting paradigm where lncRNAs act as a scaffold recruiting chromatin remodelling complexes in specific target loci. Intriguingly, even in the caudal anatomical districts where *HOTTIP* is expressed at the highest level, its copy number is less than one per cell (Wang et al., 2011). This very low expression level supports the idea that

HOTTIP acts in a spatially restricted way, only influencing the proximal genes in its chromosomal neighbourhood.

A conceptually similar mechanism that was recently proposed by Sigova et al. (2015) suggests that certain lncRNAs transcribed from regulatory sites such as enhancers or promoters might bind specific transcription factors and, as a consequence, increase their local concentration. In particular, this work focused on the transcription factor Yin-Yang 1 (YY1), which is ubiquitously expressed and has the capacity to bind both DNA and RNA *in vitro*. It was found that YY1 tends to bind ncRNAs species transcribed from active promoters and enhancers where it is bound to DNA (Sigova et al., 2015). According to the model proposed, nascent lncRNAs can bind transcription factors via weak interactions and increase their local concentration in proximity to regulatory regions. As a consequence of the higher local concentration, the transcription factors are therefore more likely to engage in stronger interactions with their binding sites in the regulatory DNA regions. According to this model, some lncRNAs – in particular bidirectional transcripts – might act as an additional regulatory layer that controls the binding kinetics of transcription factors (Sigova et al., 2015).

A further group of lncRNAs with positive roles in the regulation of neighbouring genes was identified by Ørom et al. (2010). These lncRNAs, termed ncRNA-activating (ncRNA-a), are typically ~800nt in length, are spliced and marked by H3K4me3 at their promoters and H3K36me3 in the transcript body (Ørom et al., 2010). Although there are no precise criteria that distinguish this sub group from other annotated lncRNAs, it was shown that ncRNA-a recruit the Mediator complex, a 30 subunit co-activator complex important for the regulation of RNA polymerase II (reviewed in Malik and Roeder, 2010). The Mediator complex, in turn, promotes phosphorylation of serine 10 of histone H3, a histone modification deposited by its subunit CDK8 and important for transcriptional activation (Knuesel et al., 2009).

Repressor lncRNAs In addition to lncRNAs involved in dosage compensation and imprinting, there are several other cis-acting lncRNAs that repress other genes. A common class of ncRNAs that repress genes *in cis* is antisense lncRNAs, also known as Natural Antisense Transcripts (NATs). Antisense transcription is widespread in mammalian genomes, with 50–70% of annotated transcripts having antisense partners (Carninci et al., 2005; Galante et al., 2007). Antisense transcription can affect the expression of other transcripts in several ways, the best described of which are promoter competition, transcriptional interference and epigenetic silencing (Guil and Esteller, 2012). A clear example of the latter mechanism is provided by the lncRNA *AN-*

RIL. *ANRIL* is transcribed in the antisense orientation from the coding locus INK4b/ARF/INK4a, which encodes the tumour suppressor proteins P14, P15 and P16. It was found that *ANRIL* directly interacts with the chromodomain of the protein chromobox 7 (CBX7). CBX7 is a subunit of PRC1 which also interacts with H3K27me3, and its recruitment to the INK4b/ARF/INK4a locus maintains transcriptional suppression (Yap et al., 2010).

An alternative mechanism of transcriptional repression by NATs depends on transcriptional interference. It was shown in yeast that the head-to-head collision of elongating RNA polymerases stops transcription (Hobson et al., 2012). Similarly, it was shown that an elongating RNA polymerase II can displace the assembly of the transcription preinitiation complex from other promoters (Shearwin et al., 2005), a phenomenon referred to as "sitting duck". Such mechanisms of transcriptional interference are responsible for the silencing of the imprinted gene *IGF2R* by the lncRNA *AIRN* as described in the previous paragraphs.

1.1.4.2 *Trans-acting lncRNAs*

Dozens of well characterised lncRNAs have been shown to exert a function *in trans*, i.e. a function which is independent of the genomic locus where they are transcribed. Rinn and Chang (2012) recently proposed a division of trans-acting lncRNAs into three subclasses based on commonalities of their modes of action. I will use here the same classifications to present selected examples of the mechanisms by which lncRNAs exert their regulatory functions *in trans*.

Guides A broad group of lncRNAs have been shown to be able to bind proteins or protein complexes and mediate, or guide, their localisation to specific genomic loci. One of the best characterised examples of guide lncRNAs is *HOTAIR* (HOX transcript antisense intergenic RNA). *HOTAIR* is a spliced and polyadenylated transcript produced from an antisense locus between the genes *HOXC11* and *HOXC12* in the *HOXC* cluster (Rinn et al., 2007). Interestingly, it was found that *HOTAIR* possesses a 5′ domain that binds the PRC2 complex and a 3′ domain that binds the histone demethylase LSD1 (Tsai et al., 2010) and guides them to hundreds of genomic loci resulting in their silencing via methylation of H3K27 (mediated by PRC2) and demethylation of H3K4 (via LSD1) (Tsai et al., 2010). A recent proteomics study revealed that the inhibition of *HOTAIR* caused the differential expression of 170 proteins in HeLa cells, and caused mitochondrial dysfunctions and ultrastructural damage (Zheng et al., 2015). Similarly, several other studies have reported that the expression of *HOTAIR* is altered in primary breast tumours and high expression levels are correlated with lower survival rates (Gupta et al., 2010). Over-

all, these data highlight the widespread roles of *HOTAIR* in diverse biological contexts and suggest that lncRNAs acting *in trans* might be key players in the regulation of complex epigenetic programs. An interesting aspect of this mode of regulation is the mechanism by which lncRNAs are able to target specific genes. The molecular basis of *HOTAIR* targeting is still unclear but Mondal et al. (2015) recently described the mechanisms by which the lncRNA *MEG3* recognises its targets. *MEG3* (Maternally Expressed 3, also known as Gene trap locus 2 in the mouse) is an imprinted lncRNA that acts as a tumour suppressor (Benetatos et al., 2011; Zhou et al., 2012). Using chromatin oligonucleotide-affinity precipitation (ChOP, a technique based on hybridisation and precipitation that allows the detection of DNA-RNA interactions; Mariner et al., 2008) it was shown that *MEG3* directly interacts with over 5000 genes, many of which are components of the TGF-β pathway (Mondal et al., 2015). Furthermore, it was also found that *MEG3* possesses two distinct domains, one interacting with the Polycomb complex PRC2 and the other directly interacting with GA-rich chromatin regions through the formation of DNA-RNA triplex structures (Mondal et al., 2015). Overall, this work proposes an interesting mechanism that might provide an explanation for how lncRNAs guide protein complexes, such as epigenetic regulators, to thousands of genomic loci *in trans*.

Decoys Certain lncRNAs have been shown to provide binding sites for DNA binding proteins. Thus, these ncRNAs act as decoys for proteins or other regulatory molecules, preventing them from binding their targets. Examples of such a mechanism are provided by the lncRNAs *GAS5* - which has an hairpin motif that mimics the binding site of the glucocorticoid receptor (Kino et al., 2010) - and *PANDA*. *PANDA* is a lncRNA transcriptionally regulated by Cellular tumor antigen p53 (P53) that binds the transcription factor NF-YA, preventing it from occupying target gene promoters. NF-YA typically promotes the transcription of pro-apoptotic genes, therefore the expression of *PANDA* promotes cell survival (Hung et al., 2011). Another recently described lncRNA that acts as a decoy is *NORAD*. *NORAD* contains 17 binding sites for the two Pumilio RNA binding proteins PUM1 and PUM2. The sequestration of Pumilio by *NORAD* prevents them from binding to their target mRNAs, effectively modulating their abundance (Tichon et al., 2016; Lee et al., 2016). A further example of decoy ncRNA is provided by *CIRS-7*, a circRNA which contains 70 binding sites for the microRNA miR-7. *CIRS-7* efficiently binds miR-7 in complex with the Argonaute protein (AGO) and greatly decreases the activity of miR-7 (Hansen et al., 2013; Memczak et al., 2013). It is still unclear how widespread this phenomenon might be, however these data suggest that some circRNAs might act as decoys to finely regulate the activity of microRNAs.

Scaffolds A number of lncRNAs act as molecular scaffolds, as they mediate the formation of large ribonucleoprotein complexes. As mentioned above, RNAs possess numerous binding domains that mediate their interaction with other molecules. Firstly, they can bind other RNAs through base pairing, therefore acting as sensors for mRNAs, miRNAs or other lncRNAs (Mercer and Mattick, 2013). Secondly, they can also bind one or more proteins through separate domains, as in the case of *HOTAIR* and *ANRIL*. The Gene Ontology (GO; Gene Ontology Consortium, 2015) currently annotates 3517[1] human proteins as *Interacting selectively and non-covalently with an RNA molecule or a portion thereof*, highlighting the preponderance of RNA-protein interactions. The specificity of the interactions between RNA Binding Proteins (RBPs) and their RNA substrates are only starting to be explored, but recent works suggest that short highly conserved RNA motifs represent a largely unexplored RNA-binding recognition code (Ray et al., 2013). Lastly, lncRNAs can also bind DNA, likely in a sequence specific way. These interactions can be mediated by direct RNA-DNA base pairing with the formation of duplexes or triplexes, as in the case of the lncRNA *GAS5*, or potentially they might also be mediated by the presence of specific RNA structural domains that form binding pockets for specific DNA sequences (Mercer and Mattick, 2013). A recent example of a DNA-binding, scaffold lncRNA is provided by *FIRRE*. *FIRRE* is a nuclear retained and chromatin associated lncRNA that localizes to a 5Mb domain around its genomic locus on the X chromosome. It was recently shown that *FIRRE* binds five distinct trans-chromosomal loci via a 156nt sequence and physically interacts with the Heterogeneous Nuclear Ribonucleoprotein U (hnRNPU), bringing them in close spatial proximity (Hacisuleyman et al., 2014).

The features of lncRNAs reported above highlight their flexibility as regulatory molecules. Britten and Davidson proposed in 1969 the existence of *activator and integrator genes* that are able to sense, integrate and respond to external signals and finely control the activity of the genome. They went on postulating that the functional product of these activator genes would be RNA molecules with the following characteristics:

> "(i) [Activator RNAs] will, in the main, be confined to the nucleus, that is they are not precursors of cytoplasmic polysomes. (ii) When observed in their functional role, they would be found in chromatin, bound to DNA in a sequence-specific manner. (iii) They are often the product of the redundant fraction of the genome. (iv) They include sequences not present in the polysomes

[1] Data from geneontology.org, database accessed 10/8/2016

> carrying producer-gene[1] templates, that is, most or all cytoplasmic polysomes." (from Britten and Davidson, 1969, page 354)

It is hard not to notice *a posteriori* a striking resemblance between Britten and Davidson's activator RNAs and the molecules now known as lncRNAs. It is clear that the development of complex organisms requires robust regulation of gene expression programs. The properties of lncRNAs presented in the previous paragraphs make them the ideal molecules to integrate diverse signals and mediate multiple effector functions in an orchestrated way (Amaral et al., 2008; Mattick, 2009).

[1] The term "producer gene" is used by Britten and Davidson to refer to protein coding genes

1.2 microRNAs

MicroRNAs (miRNAs) are small non-coding RNAs of ~20nt in size present in the majority of eukaryotes. Their existence was discovered in 1993 by the groups of Victor Ambros and Gary Ruvkun in the nematode *Caenorhabditis elegans* with the identification of the lin-4 miRNA (Lee et al., 1993; Wightman et al., 1993). The beginning of the 2000s saw numerous other studies that identified miRNAs in virtually all other animals and plants (Pasquinelli et al., 2000; Elbashir et al., 2001; Lau et al., 2001; Lagos-Quintana et al., 2001), identified the mechanisms of their biogenesis (Grishok et al., 2001; Hutvagner et al., 2001), elucidated their modes of action (Enright et al., 2003; Doench and Sharp, 2004; Lewis et al., 2005) and characterised their biological roles (Giraldez et al., 2006; Rodriguez et al., 2007).

We now know that miRNAs are transcribed from the genome in the form of long primary transcripts (pri-miRNAs) which then undergo a sequential multi-step processing to produce mature, single stranded RNA molecules of about 20-23nt. The main function of mature miRNAs is to regulate the expression of other genes by multiple mechanisms. The following paragraphs will summarise the current literature on miRNA biogenesis, evolution and function.

1.2.1 *Genomic organisation of microRNAs*

The human genome encodes 1881 miRNA primary transcripts and 2588 distinct mature sequences (data from obtained from miRBase, version 21; Griffiths-Jones et al., 2008). Typically, miRNAs are transcribed as long primary transcripts and then processed into their mature form by a complex enzymatic machinery that will be described in detail in the following sections. The majority of pri-miRNAs are transcribed by RNA Polymerase II, capped and polyadenylated (Lee et al., 2004) and often reside in the introns of coding or non-coding transcripts (Cai et al., 2004), although some miRNAs overlap exonic sequences. Interestingly, it was shown that a smaller group of miRNAs that reside within repetitive elements, such as Alu elements, are instead transcribed by RNA Polymerase III (Borchert et al., 2006). A large number of miRNAs are organised in clusters and transcribed in polycistronic units (Lee et al., 2002; Saini et al., 2007), while other are transcribed independently.

The aforementioned principles of genomic organisation of miRNAs represent by no means precise rules: in fact, there is evidence that intronic miRNAs can be transcribed independently of their host coding transcripts (Ramalingam et al., 2014), and also that miRNAs residing in clusters can be transcribed in-

dependently from each other and regulated by alternative splicing (Monteys et al., 2010; Ramalingam et al., 2014). These data highlight the fact that the transcriptional landscape of miRNAs is not as simple as initially thought. For this reason, various groups have dedicated their efforts at defining the structure of pre-miRNAs and identifying their promoters. The majority of these works leveraged RNA Polymerase occupancy, histone modifications, CAGE and RNA-Seq data to identify the putative TSSs of miRNAs, finding that the majority of intronic miRNAs are transcribed from their own promoters which are independent from those of the host transcripts (Saini et al., 2007; Ozsolak et al., 2008; Marson et al., 2008; Monteys et al., 2010; Marsico et al., 2013). In a recent work, Marsico et al. (2013) subdivided miRNA promoters in three classes based on their genomic context: intergenic promoters, for miRNA that do not reside within the introns of other genes, intronic promoters, for miRNAs independently transcribed from a promoter that resides in an intron of a host gene, and host gene promoters, for miRNAs that are transcribed from the same promoter as the host gene. The analysis of these three classes of promoters revealed that intronic promoters tend to be shorter, less conserved and with a smaller CpG content than the host gene promoters; on the other hand, they were also found to be enriched in TATA box elements and to contain binding sites for a set of transcription factors distinct from that of intergenic and host gene promoters (Marsico et al., 2013). Furthermore, it was also found that miRNAs with an independent intronic promoter tend to be evolutionarily older than those transcribed from the host gene promoter. These data led to the interesting speculation that the introns of protein coding genes might be a favourable substrate for the evolution of new functional miRNAs, which would evolve an independent promoter at a later time (Marsico et al., 2013).

1.2.2 *The biogenesis of microRNAs*

1.2.2.1 *Processing of pri-miRNAs*

Typically, pri-miRNAs are several kilobases long (Marsico et al., 2013) and have a stem-loop structure that encompasses the miRNA. The stem loop is composed of an apical loop of variable size, a double stranded stem of ~35bp and two single stranded basal regions at the 5′ and 3′ of the stem (Han et al., 2006) (**Figure 1.3A**). The first step in the maturation process is catalysed by an enzymatic complex called the Microprocessor, which contains the nuclear RNase III DROSHA and its cofactor DGCR8 (known as Pasha in *D. melanogaster*). The Microprocessor catalyses the endonucleolitic cleavage of the pri-miRNA at the base of the stem-loop structure, releasing the hairpin (called the

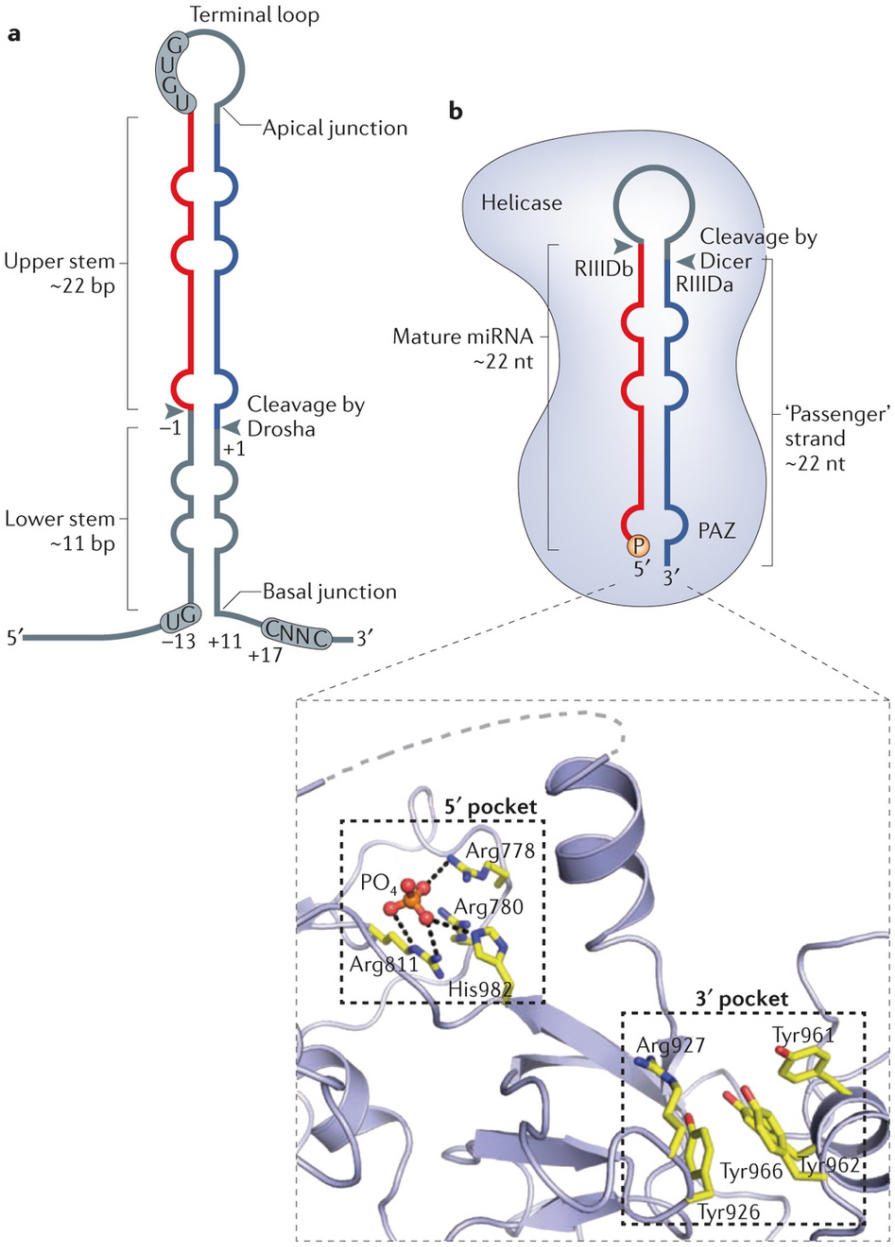


Figure 1.3 *A: Schematic representation of the stem-loop structure of a pri-miRNA.* ***B****: Diagram showing the recognition of pre-miRNAs by Dicer. Illustration adapted from Ha and Kim, 2014*

pre-miRNA) containing the mature miRNA sequence (Lee et al., 2003; Denli et al., 2004; Gregory et al., 2004).

DROSHA is a highly conserved protein of ~159 kDa. Its amino-terminus (N-terminus) is not directly involved in pri-miRNA processing (Han et al., 2004), but it contains a Nuclear Localisation Signal (NLS), which, upon phosphorylation of one of two Serine residues (S300 and S302), mediates localisation to the nucleus (Tang et al., 2010). In line with these findings, it was recently observed that DROSHA splice variants devoid of N-terminal regions have cytoplasmic localisation (Link et al., 2016), and there is evidence suggesting that DROSHA possesses pri-miRNA processing ability in the cytoplasm (Dai et al., 2016). On the other hand, the carboxy-teminus (C-terminus) of DROSHA contains two RNase III domains and one double strand RNA Binding Domain (dsRBD), which are directly involved in pri-miRNA processing. The dsRBD binds the pri-miRNA in concert with the two dsRBD of DGCR8, which are necessary for efficient binding (Nguyen et al., 2015), while the RNase III domains form an intramolecular dimer and respectively cleave the 3′ and 5′ strands of the pri-miRNA stem. DGCR8 interacts with DROSHA via its C-terminal domain, and also binds ferric heme [Fe(III)], which is a required cofactor for its activity (Barr et al., 2012; Weitz et al., 2014) and enhances Microprocessor's accuracy and efficiency (Roth et al., 2013; Quick-Cleveland et al., 2014; Nguyen et al., 2015). Both DGCR8 and DROSHA are associated with the chromatin at the sites of miRNA transcription, and their enzymatic activity occurs co-transcriptionally (Pawlicki and Steitz, 2008; Morlando et al., 2008). For intronic miRNAs, DROSHA cleavage happens before splicing and the 5′ and 3′ exposed ends of the intron are quickly degraded by exonucleolitic cleavage (Morlando et al., 2008). Interestingly, it was also postulated that clearance of introns might facilitate exon splicing, thus promoting mRNA maturation.

Two recent studies by the groups of Narry Kim and Jae-Sung Woo shed light on the biochemical and structural properties of the Microprocessor, revealing the assembly dynamics of the complex and the molecular basis for its activity. These studies found that the Microprocessor is a heterotrimer consisting of two DGCR8 subunits and one DROSHA subunit (Nguyen et al., 2015) which form an elongated complex that spans the whole length of the pri-miRNA (Kwon et al., 2016). In this complex, DROSHA can recognise the basal junction between the stem and the single-stranded RNA through a helical structure named "Bump". "Bump" is located ~28 Å from the catalytic site of the RNase III domain responsible for cleaving the 3′ end of the pri-miRNA stem. This spatial conformation leads to the precise cleavage of the 3′ strand at ~11bp from the basal junction, in accordance with previous biochemical observations (Zeng et al., 2005; Han et al., 2006; Nguyen et al., 2015; Kwon

et al., 2016). These data highlight the importance of the pri-miRNA secondary structure for proper recognition and processing by the Microprocessor. Nevertheless, there is also substantial evidence demonstrating that primary sequence features of the pri-miRNA also play an important role in conferring specificity to DROSHA. In fact, it was found that the majority of human pri-miRNAs possess specific motifs in the basal region and/or in the terminal loop (Auyeung et al., 2013). A UG motif in the basal region is directly recognised by DROSHA, while DGCR8 recognises a UGU motif in the loop. Concurrently, these two interactions were shown to confer specificity to Microprocessor's binding and to increase its efficiency in cleaving pri-miRNAs (Nguyen et al., 2015). An additional CNNC motif in the basal region also seems to play a role in regulating Microprocessor's activity, as it provides a binding site for the splicing factor SRp20, which enhances processing of the pri-miRNA (Auyeung et al., 2013). In a recent study, Fang and Bartel (2015) analysed the cleavage efficiency of 50 000 artificially generated pri-miRNA variants and proposed a unifying model of pri-miRNA processing that incorporates novel and previously identified features. They found that a mismatched GHG motif at the base of the stem, a stem length of 34 nt to 36 nt, and complementarity throughout the stem, augment Microprocessor's capacity to process pri-miRNAs. Incorporating these observations with the previously identified features of pri-miRNAs, they proposed that the ideal Microprocessor's substrate is a 35bp hairpin with a flanking single-stranded region, a mismatched GHG motif in position 7 of the stem, and a basal UG, apical UGU and flanking CNNC motifs. These features allowed the design and generation of artificial pri-miRNAs that are efficiently processed by Microprocessor, strongly supporting the validity of such model (Fang and Bartel, 2015).

Post-transcriptional RNA editing of the pri-miRNAs provides a recently discovered layer of further regulation. It was observed that the enzyme Double-stranded RNA-specific adenosine deaminase (ADAR), which converts adenosine residues to inosines, can modify miRNA primary transcripts altering their efficiency of processing by DROSHA (Chawla and Sokol, 2014). This process might be responsible - at least in part - for the differential expression of miRNAs that reside in polycistronic clusters (Chawla and Sokol, 2014).

The Microprocessor-dependent maturation of pri-miRNAs described above represents the canonical pathway responsible for the production of the majority of pre-miRNAs. However, there is evidence of alternative biogenesis mechanisms that are DROSHA independent. For example, Mirtrons are a class of miRNAs located inside small introns of other genes that, upon splicing, can directly fold into a pre-miRNA, thus completely bypassing DROSHA-mediated cleavage (Ruby et al., 2007; Okamura et al., 2007). Similarly, another group

of pre-miRNAs, known as 5′-capped miRNA precursors, are directly transcribed by RNA Polymerase II as short pre-miRNAs and therefore possess a 5′ methylguanosine cap (Xie et al., 2013b). Like Mirtrons, 5′-capped miRNAs also bypass DROSHA processing and are directly exported to the cytoplasm for DICER processing.

1.2.2.2 *Processing of pre-miRNAs*

The pre-miRNAs generated by DROSHA in the nucleus are exported to the cytoplasm, where they exert their regulatory functions on mRNAs. The nuclear export is mediated by the protein Exportin-5 (EXP-5), which binds pre-miRNAs and shuttles them through the nuclear pore complex in a manner regulated by the small GTPase RAN (Yi et al., 2003; Lund et al., 2004). The high resolution structure of a pre-miRNA bound to the export complex revealed that EXP-5 recognises the 3′ overhang as well as the double stranded stem of pre-miRNAs in a glove-like structure, and in addition to mediating the nuclear export protects them from exonucleolytic degradation (Yi et al., 2003; Okada et al., 2009). A recent study assessed the cytoplasmic expression of mature miRNAs after knocking out *XPO5* (the gene that encodes EXP-5) revealing that a considerable number of miRNAs are not affected by the knock out (Kim et al., 2016). These results suggest that certain miRNAs do not require EXP-5, and are instead exported to the cytoplasm by other factors.

In the cytoplasm, pre-miRNAs are further processed by the enzyme Endoribonuclease Dicer (DICER1 or DICER), which mediates the cleavage of the apical loop releasing a double-stranded stem of ~20bp (**Figure 1.3B**). This processing step is fundamental for the generation of mature miRNAs, as highlighted by the fact that the knock-out of DICER is embryonically lethal in the mouse during gastrulation (around day 7.5, Bernstein et al., 2003) and causes abnormal gastrulation, brain formation, somitogenesis and heart development in zebrafish (Giraldez et al., 2005). Human DICER is a protein of 219 kDa composed of an RNA helicase domain, a Domain of Unknown Function (DUF283), a PIWI-AGO-ZWILLE domain (PAZ), two RNAse III domains and a C-terminal double-strand RNA binding domain. In a mechanism similar to that described for DROSHA, the two RNAse III domains form the catalytic centre in an intramolecular homodimer (Zhang et al., 2004), while the PAZ domain and the N-terminal helicase domain respectively recognise the 3′ overhang and the apical loop of the pre-miRNA (Tian et al., 2014; Tsutsumi et al., 2011). The crystal structure of DICER revealed that the PAZ domain and the catalytic centre are located ~65 Å apart, thus acting as a molecular ruler that leads to the cleavage of the stem at ~25 nucleotides from the 3′ end of the pre-miRNA (Macrae et al., 2006).

It was found that RNA editing can also modulate DICER processing. For example, the conversion of adenosine to inosine in pri-miR-151 significantly abolishes the processing capacity of DICER, leading to the accumulation of pre-miR-151 (Kawahara et al., 2007). Furthermore, pre-miRNAs maturation can also be influenced by RNA binding proteins that modulate DICER activity. For example, it was shown *in vitro* that binding of LIN28 to pre-let-7 recruits the 3′ terminal uridylyl transferase ZCCHC11 (zinc finger, CCHC domain containing 11), which mediates pre–let-7 uridylation and its subsequent degradation (Heo et al., 2008; Hagan et al., 2009). This regulatory mechanism of LIN28-mediated regulation of pre-let-7 was also demonstrated *in vivo* in the nematode *Caenorhabditis elegans*, where LIN28 recruits the Poly(U) Polymerase PUP-2 (Lehrbach et al., 2009). Taken together, these results indicate that pre-miRNA sequestration and degradation represent an additional conserved regulatory layer in the pre-miRNA processing pathway.

1.2.3 *RISC loading*

After DICER processing, the resulting double strand RNA molecule is loaded into the RNA-induced Silencing Complex (RISC), a multi-protein complex that contains a member of the Argonaute family (AGO). In humans there are four members of the AGO family, named AGO1-4, and all of them contribute to miRNA mediated regulation and share similar functions (Su et al., 2009), albeit displaying different characteristics in their preference to bind certain miRNAs as well as in the range of effects that they mediate on target mRNAs (reviewed in Ha and Kim, 2014). In particular, AGO2 plays a prominent role, as highlighted by the observation in knock-down mice that maternal *Ago2* is required for maternal-zygotic transition and embryonic development after the two-cells stage (Lykke-Andersen et al., 2008).

A key property of AGO loading is the relative orientation of binding between the miRNA duplex and AGO, because it dictates which strand will be kept for target recognition (guide strand) and which will be discarded and degraded (passenger strand) (Khvorova et al., 2003). The dynamics of miRNA loading into AGO are still not fully clear, but several works suggest that the process takes place through the direct interaction of DICER with AGO and the double-strand RNA binding proteins TRBP and PACT (Sasaki and Shimizu, 2007; Noland and Doudna, 2013). The loading of the miRNA duplex on AGO is an ATP-dependent process (Kawamata et al., 2009), which is followed by the unwinding of the passenger strand. Early works suggested that in *Drosophila Melanogaster* the rules for strand selection rely on the relative thermodynamic stability of the 3′ and 5′ end of the duplex, with the guide strand having lower

stability at its 5′ end (Khvorova et al., 2003; Schwarz et al., 2003; Tomari et al., 2004). However, emerging evidence suggests that in mammals this process is more complicated and depends on multiple factors, such as duplex thermodynamics, 5′ sequence and structure as well as the specific protein partners interacting with the complex (Noland and Doudna, 2013). A recent structural study revealed that the double-stranded RNA binding proteins TRBP and PACT bind DICER in a mutually exclusive way with profound effects on strand selection and product length determination (i.e. isomiR formation) (Wilson et al., 2015). In particular, it was observed that the binding of TRBP or PACT induces a shift in the position cut by DICER, promoting the formation of miRNAs of 22nt. In some instances, for example miR-30e in the mouse, an extra nucleotide at the 5′ of the miRNA duplex influences the binding preference of AGO, thus changing the choice of guide strand (Wilson et al., 2015).

1.2.4 *RISC-target recognition*

The miRNA-RISC complex (miRISC), which results from AGO-loading and guide strand selection, is the final effector of the miRNA pathway and is able to target mRNAs to reduce the abundance of the proteins that they encode, by translational repression and/or by mRNA degradation through decapping and deadenylation. Important insights into the modes of target recognition by RISC came from solving the crystal structure of a bacterial AGO protein in complex with a 21nt DNA guide strand. It was found that the guide strand is tethered to AGO at both the 5′ and 3′ end, while the nucleotides in positions 2 to 6 are protruding, with their Watson-Crick edges exposed and positioned to interact with target mRNAs (Wang et al., 2008). These data are in line with the previous observation that target recognition is mediated, to a large extent, by perfect complementarity of 7 or 8 nucleotides at the 5′ end, a region referred to as the seed region (Lewis et al., 2003; Ma et al., 2005; Parker et al., 2005). In addition to exposing the Watson-Crick edges of the guide RNA, the binding between the PIWI/MID domain of AGO and the miRNA induces a 300-fold increase in the affinity of interaction between the seed and a complementary target by reducing the entropy cost of the base pairing (Parker et al., 2009). This mechanism confers high stability and specificity to the RISC target recognition, which is thus primarily dictated by the sequence of the seed. Despite these data, there is also conflicting evidence suggesting that base-pairing at the 3′ end could either compensate mismatches in the seed, or have limited effects on target recognition (reviewed in Hausser and Zavolan, 2014). In addition to the complementarity between seed and target site, the sequence context of the target site is another factor that influences the effect of the miRNA on the

target mRNA. The majority of miRNA target sites appear in 3′ UTRs (Bartel, 2009), whereas binding to 5′ UTRs or Coding DNA Sequences (CDSs) is less frequent (Easow et al., 2007; Forman et al., 2008). A recent study used PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) on AGO2 to shed light on the factors that contribute to miRNA efficacy (Hafner et al., 2010). It was found that the length of the target site as well as the number of target sites play an important role, as multiple and longer sites have the greatest effect in reducing mRNA stability. Furthermore, this study also found that target sites in the CDS have limited effects on mRNA stability disregarding of the extent of miRNA-mRNA pairing, likely due to steric hindrance between miRISC and the translating ribosome (Hausser and Zavolan, 2014). Finally, an additional important factor is the accessibility of the target site, as it was observed that the majority of target sites are found in regions that require low free energy to solve local secondary structures (Hafner et al., 2010).

The short length of the miRNA seed implicates a high probability of occurrence of target sites throughout 3′ UTRs. In fact, the majority of miRNAs bind multiple transcripts (Lim et al., 2005) and more than 60 % of the transcriptome displays signs of selective pressure to maintain miRNA target sites (Friedman et al., 2009). Hence, miRNAs usually act in a combinatorial manner regulating the expression of multiple mRNAs, which in turn are regulated by multiple miRNAs. This model is further complicated by the fact that multiple miRNAs share the same seed sequence, constituting large miRNA families that share the same or similar sets of targets.

1.2.5 *Effects of RISC-target binding*

In animal cells, miRNAs have two predominant effects on target mRNAs: the degradation of the target mRNA or the inhibition of translation. In both cases, the functional result is a decrease in the expression of the protein encoded by the target. The following paragraphs will discuss the current literature on these two modes of action.

Translational repression Several independent works have shown that miRNAs can inhibit the translation of target mRNAs by acting on various levels. In 2004, Pillai et al. (2004) found that the miRNA-independent targeting of the human AGO proteins to the ’3 UTR of an mRNA induces translational repression, thus demonstrating that miRNAs are required for AGO targeting but are dispensable for silencing. Further work by the same group and others, applied sucrose gradient fractionation of the polysomes to demonstrate that

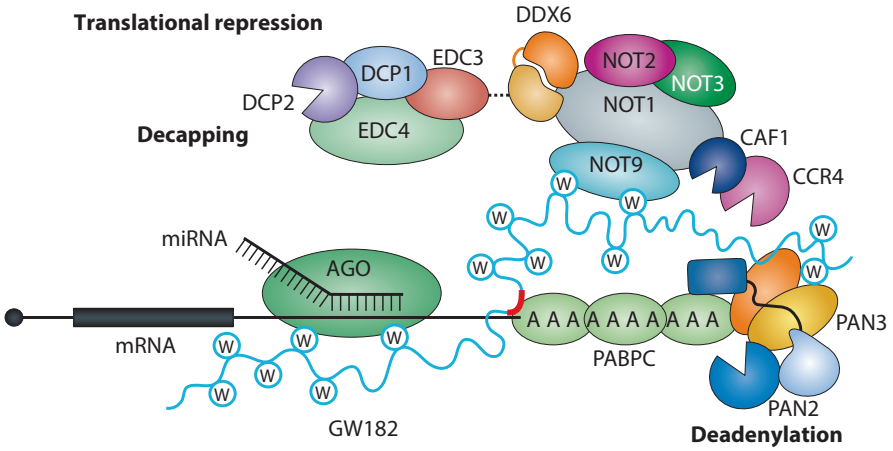


Figure 1.4 Schematic representation of the effects of miRNA-mediated gene silencing. Illustration from Jonas and Izaurralde, 2015.

miRNA targeting causes the dissociation of the ribosomes from the mRNA (Pillai et al., 2005; Humphreys et al., 2005). In addition, it was also noticed that miRNAs only induced a repression of translation from the 5′ methylguanosine cap, while they did not have effects on translation initiated from an Internal Ribosome Entry Site (IRES) (Pillai et al., 2005; Humphreys et al., 2005). These results suggest that the translation inhibition mediated by miRNAs acts at the initiation step, and likely involves cap recognition and/or assembly of the translation initiation factors. In fact, subsequent works further supported this hypothesis, showing that vectors with a non-physiological A(5′)ppp(5′)G cap are not translationally repressed by miRNAs (Mathonnet et al., 2007; Wakiyama et al., 2007), while increasing concentrations of the initiation complex eIF4F rescued the miRNA-mediated translational inhibition (Mathonnet et al., 2007).

These works strongly suggested that translational inhibition mediated by miRNAs acts at the level of initiation. However, it is worth noting that there are also contrasting results that show some inhibitory effects at the post-initiation stage (Nottrott et al., 2006; Maroney et al., 2006; Petersen et al., 2006). According to these studies, miRNAs might cause the early dissociation of the ribosome and - in contrast with the results of Pillai et al. (2005) and Humphreys et al. (2005) - they are also able to repress translation initiated from an IRES (Petersen et al., 2006).

Target degradation There is a vast body of research demonstrating that miRNAs, in addition to inhibiting translation, also mediate the deadenylation and degradation of the target mRNAs (**Figure 1.4**). In fact, studies that combined transcriptomics, proteomics and ribosome footprinting revealed that for the majority of mammalian miRNAs, the decrease in target mRNA

levels is the cause of reduced protein expression (Lim et al., 2005; Giraldez et al., 2006; Selbach et al., 2008; Baek et al., 2008; Hendrickson et al., 2009; Guo et al., 2010). The degradation of the target is a sequential, multi-step process that terminates with 5′-3′ exonucleolytic degradation. The recruitment of miRISC to the target mRNA is the triggering event in this process, which is followed by the target's deadenylation, then decapping and finally degradation by a cytoplasmic exonuclease (reviewed in Jonas and Izaurralde, 2015 and in Filipowicz and Sonenberg, 2015). This process is mediated by a very large number of proteins and multi-protein complexes that catalyse the reactions involved in each of these steps, but the key player that orchestrates and starts the degradation pathway is the AGO interacting partner GW182. GW182 is a conserved family of proteins that in humans is composed of three Trinucleotide repeat-containing (TNRCs) proteins, named TNRC6A, TNRC6B and TNRC6C. The GW182 proteins owe the name to their functional unstructured domains characterised by the presence of repetitive tryptophan motifs often followed or preceded by glycine residues. Recent structural studies revealed that the W-containing motifs of GW182 mediate the interaction with AGO, by inserting into consecutive hydrophobic tryptophan binding pockets in the PIWI domain (Schirle and MacRae, 2012). At the same time, GW182 also binds in a similar fashion the deadenylase complex PAN2-PAN3, via the tryptophan binding pockets of PAN3 (Christie et al., 2013). The PAN2-PAN3 complex interacts with the target's polyA tail via the direct binding of PAN3 to the cytoplasmic poly(A)-binding protein (PABPC) (Siddiqui et al., 2007), while PAN2 is the active subunit of the complex that catalyses deadenylation (Uchida et al., 2004). The PAN2-PAN3 complex mediates the deadenylation of long polyA tails without completely degrading them, and leaves a stub of~25 residues before dissociating from PABPC (Baer and Kornberg, 1983). Subsequently, the deadenylation process is completed by the concerted action of the CCR4-NOT complex (Yamashita et al., 2005). This complex also associates with the GW182 proteins through the tryptophan binding pockets of the NOT9 subunit (Chen et al., 2014) and mediates translational repression (via the interacting partner DDX6, Mathys et al., 2014), deadenylation (Fabian et al., 2011) and target mRNA decapping and degradation. The CCR4-NOT interacting partner DDX6 is a decapping factor that, in addition to mediating translational inhibition, provides a physical and functional link between the CCR4-NOT complex and the decapping complex, which consists of multiple activator proteins such as the enhancers of mRNA decapping (EDC3 and EDC4) and the DCP1:DCP2 enzymatic complex that catalyses the decapping reaction (She et al., 2008; Fromm et al., 2012; Chen et al., 2014; Mathys et al., 2014; Mugridge et al., 2016). Upon decapping, the mRNA is degraded by the

activity of the 5′-3′ exonuclease 1 (XRN1), which recognises as a substrate 5′ monophosphorylated mRNAs (Braun et al., 2012).

Taken together, these results highlight the fact that the effects of miRISC on target mRNAs are mediated by a complex series of protein factors and tightly coupled molecular events, leading to the concerted translational repression and destabilisation of target mRNAs. Despite the great progress achieved in this field in the last decades, the thorough understanding of miRNA-mediated repression is still a "long unfinished march" (Filipowicz and Sonenberg, 2015) which will require further work before completion.

2

Extracellular secretion of RNAs

2.1 Exosomes and extracellular vesicles

Cell-to-cell communication is typically mediated by mechanisms that involve the interaction of a receptor on the surface of a cell with a specific ligand. This mode of communication is usually mediated be either membrane proteins (such as in direct cell-to-cell interaction) or secreted soluble molecules, such as cytokines, chemokines, hormones and growth factors. However, an additional mode of cell-to-cell communication that has gained wide interest in recent years is the secretion of extracellular membrane vesicles.

Extracellular Vesicles (EVs) are a heterogeneous class of small vesicles surrounded by a double-layer lipid membrane and secreted into the extracellular space by the majority of cell types. The first observations of EVs date back to the 1960s, and 1970s, when it was observed that platelets release a lipidic particulate, named "platelet dust", that could be sedimented by centrifugation (Wolf, 1967). Shortly after, electron microscopy studies allowed the identification of small membrane vesicles within the cartilage matrix (Anderson, 1969) and in the extracellular space between the microvilli of rabbit taste bud cells (Fujimoto, 1973). Almost two decades later, Johnstone et al. (1987) described for the first time that during their maturation, reticulocytes remove unwanted proteins by secreting a sub-class of EVs in a process akin to endocytosis but in reverse; for this reason, these vesicles were named Exosomes, using the same term previously coined by Trams et al. (1981). Since then, numerous works have further characterised the population of EVs, uncovering a broad class of vesicles with different biophysical characteristics, biological functions and routes of biogenesis. The field still lacks a clear consensus on the nomenclature of these vesicles (Gould and Raposo, 2013), but they are typically distinguished based on their biogenesis. Ectosomes, microparticles, microvesicles and shedding vesicles usually refer to the EVs that bud directly from the plasma membrane, and are in a size range of 150 nm–1000 nm (Colombo et al., 2014). On the other hand, the term "exosomes" refers to smaller EVs of 30 nm–100 nm that are formed and released to the extracellular space by the inward budding of the membrane of the late endosomal compartment, which gives rise to the formation of Multi Vesicular Bodies (MVBs), and subsequent fusion of the MVBs with the plasma membrane (Colombo et al., 2014). Although the mode

of biogenesis provides a robust theoretical definition for the various classes of vesicles, from a practical point of view EVs and exosomes are most often distinguished based on the methods by which they are purified. The most widely used technique for vesicles purification is differential centrifugation, and the community typically refers to the material recovered at 10 000 g as EVs, whereas the pellet recovered at 70 000 g to 100 000 g is considered to be enriched in exosomes (Gould and Raposo, 2013). This operational definition and the lack of standardised purification procedures introduce potential biases (due to both the technical variability of the purification as well as to differences in the protocols between different labs) in the real nature of what individual research groups call EVs and exosomes.

In the last decade, numerous works have explored the range of possible functions of EVs, and it is now apparent that they represent a robust and conserved mechanism of cell-to-cell communication that allows a cell to spread signals to the micro-environment. EVs have been identified in the majority of organisms, ranging from virtually all eukaryotes to bacteria and archaea (Deatherage and Cookson, 2012). For example, gram-negative bacteria secrete outer membrane vesicles containing toxins and other virulence factors that spread to the micro-environment, thus facilitating the colonisation of the host (Kuehn and Kesty, 2005). In animals, EVs have been detected in the majority of biological fluids, such as blood, saliva, urine, synovial fluid, bronchoalveolar lavage fluid, amniotic fluid, sperm and breast milk (Raposo and Stoorvogel, 2013). Several in depth studies have analysed their roles, revealing that they are involved in a heterogeneous spectrum of physiological functions as well as in many pathological processes such as cancer and neurodegenerative disorders (Smith et al., 2014).

In addition to their roles in physiological or pathological processes, EVs and exosomes are also being studied for their role as circulating disease biomarkers that can be easily and readily obtained from biological fluids (An et al., 2015; Smith et al., 2014). EVs and exosomes have also gained interest due to their potential therapeutic application as vectors for drug delivery (Lener et al., 2015). Several studies have investigated the feasibility of using engineered EVs for delivering therapeutic cargoes, finding that they are particularly suitable due to their biophysical characteristics (e.g. capacity to cross the blood brain barrier), immune-compatibility, and capacity to target specific cell types (Fuster-Matanzo et al., 2015). Despite the relative youth of this field, EV-based therapeutics are under rapid development and are currently being tested in cancer and type I diabetes patients in a handful of phase I and phase II clinical trials (Lener et al., 2015).

The following sections will describe the current literature, specifically addressing the biogenesis of EVs and exosomes, their content and roles in cell-to-cell communication. Despite the clear differences between EVs and exosomes, the literature often lacks precise distinctions between them, and the scientific community has only recently developed suitable methods for their differential centrifugation and characterisation. In light of this, in this introduction I will described EVs in general and focus on exosomes when appropriate.

2.2 The content of EVs and exosomes

Numerous groups have dedicated their efforts to the identification and systematic cataloguing of the content of exosomes and EVs. These efforts have led to the generation of databases such as EVpedia and Vesiclepedia, that annotate the RNAs, proteins and lipids identified inside EVs and exosomes (Kim et al., 2015a; Kalra et al., 2012). At present, Vesiclepedia reports 92 897 proteins, 27 642 mRNAs, 4934 miRNAs and 584 lipids from 538 studies in 33 different species (database accessed on 12th September 2015), but these numbers are bound to increase with the addition of new studies to the database, the development of new detection methods and the optimisation of the vesicle purification protocols. The following paragraphs will present an overview of the EV and exosome proteome and transcriptome described so far.

2.2.1 *Proteins*

In recent years, the proteome of EVs and exosomes has been the subject of a number of studies. These works utilised high throughput untargeted methods, such as mass spectrometry, as well as low throughput antibody-based methods, such as western blot and cytofluorometry, for the detection of specific proteins.

These studies revealed some important features of the protein composition of vesicles. First, it is now clear that the proteome of EVs reflects, at least in part, the proteome of the parent cell (Tauro et al., 2013). However, they do not merely sample the parent cell's protein composition, but they display an enrichment for specific proteins and a depletion of others. Typically, proteins expressed in the cytosol, plasma membrane and endosomes are enriched in EVs, while those expressed in the nucleus, golgi or endoplasmic reticulum tend to be depleted (Théry et al., 2002a; Colombo et al., 2014). This characteristic composition of EVs likely reflects their route of biogenesis, as they often display an enrichment for components of the endocytic pathway, such as the members of the Endosomal Sorting Complex Required for Transport (ESCRT) (e.g. TSG101 and ALIX) as well as tetraspanins that typically localise

to the late endosomes (e.g. CD9, CD63, CD81; Zöller, 2009). However, there is also much evidence supporting the presence of soluble cytosolic proteins, such as cytoskeletal proteins, metabolic enzymes and adaptor proteins of various signalling pathways.

Recent studies indicate that the protein composition of exosomes is very heterogeneous even within a single preparation, suggesting that the exosome population is more complex than initially thought (Tauro et al., 2013). In accordance with this, it was shown that there is a great degree of overlap in the protein content of exosomes and EVs, suggesting that the majority of exosome-specific proteins still remain to be identified (Turiák et al., 2011). Recently, Smith et al. (2015) have developed a method based on Raman spectroscopy to analyse the composition of individual exosomes. This study revealed the presence of four distinct groups of exosomes independent of parent cell type. The major difference between the four groups related to their lipid composition and surface protein expression, confirming that the exosomes population is heterogeneous and likely encompasses multiple types of vesicles with distinct characteristics and composition (Smith et al., 2015).

2.2.2 *RNAs*

One of the first works to describe the RNA content of EVs was published by Ratajczak and colleagues in 2006. They showed that EVs purified from embryonic stem cells are enriched in specific mRNA species (such as the pluripotency factors *OCT4*, *REX1*, *NANOG* and *GATA2*) compared to the parental cell, and these are translated into proteins when transferred to recipient cells (Ratajczak et al., 2006). These data led to the hypothesis that cells might possess a sorting mechanism capable of increasing the secretion of specific mRNA species (Ratajczak et al., 2006). The observation that EVs contain mRNAs was quickly expanded to exosomes, when Valadi et al. (2007) showed that human and mouse mast-cell derived exosomes contain approximately 1300 RNA species, some of which could not be detected in the donor cell. These exosomal RNAs were shown to be functional by *in vitro* translation assays and to be translated in recipient cells in cross-species experiments (Valadi et al., 2007). In addition, this work also described the presence of exosomal miRNAs, and further studies confirmed that they can be transferred to recipient cells to exert their regulatory functions on target mRNAs (Montecalvo et al., 2012; Pegtel et al., 2010; Ismail et al., 2013).

Following these early observations that EVs and exosomes can traffic functional RNA molecules, various groups have dedicated efforts to characterising their functions. It was observed that the cell-to-cell transfer of mRNAs has

important roles in stem-cell maintenance and reprogramming (Ratajczak et al., 2006), cancer progression (Skog et al., 2008; Balaj et al., 2011), stress response (Eldh et al., 2010), angiogenesis (Deregibus et al., 2007) and immune response (Robbins and Morelli, 2014). The function of exosomal miRNAs has been extensively studied, revealing their involvement in a similar set of functions. For example, Mittelbrunn et al. (2011) showed that, in the formation of the immune synapse, miRNAs are unidirectionally transferred from T cells to dendritic cells, whereas Pegtel et al. (2010) demonstrated that viral miRNAs are transferred to recipient cells in the context of viral infections. In addition, other groups have implicated the transfer of miRNAs in a variety of other processes, such as the induction of inflammatory responses in cancer (Fabbri et al., 2012), endothelial cell migration and angiogenesis (Zhuang et al., 2012) and suppression of pathogenic T lymphocytes by T-regulatory cells (Okoye et al., 2014).

In contrast with these data demonstrating roles for exosomal miRNAs, a recent work by Chevillet et al. (2014) stoichiometrically quantified the number of exosomes and miRNAs found in biological fluids and cell culture supernatants, finding that even for the most abundant miRNAs there is less than one copy per 100 exosomes (Chevillet et al., 2014). These results suggest that either the biological effects of exosomal miRNAs depend on a great number of exosomes, or that only a small fraction of exosomes contain miRNAs in biologically relevant quantities. The apparent heterogeneity in exosome preparations observed by the proteomics studies reported before (Tauro et al., 2013) seem to support the latter hypothesis, and prompt for new studies aimed at better defining the individual contribution of exosome subfamilies in the process of cell-to-cell miRNA transfer.

2.3 The biogenesis of exosomes

Exosome formation and release is a regulated process that occurs during the maturation of organelles of the endocytic pathway (**Figure 2.1**). Endosomes are intracellular organelles surrounded by a lipidic membrane responsible for the internalisation of extracellular substances and/or membrane proteins and their subsequent degradation or recycling (Klumperman and Raposo, 2014). The early endosome is a dynamic network of interconnected organelles located at the cell's periphery. Its main function is to internalise the clathrin-coated vesicles formed by endocytosis at the plasma membrane. The early endosomes gradually mature into morphologically and biochemically distinct structures termed late endosomes or MVBs (Stoorvogel et al., 1991), and in this process they acquire increasing numbers of Intraluminal vesicles (ILVs). The forma-

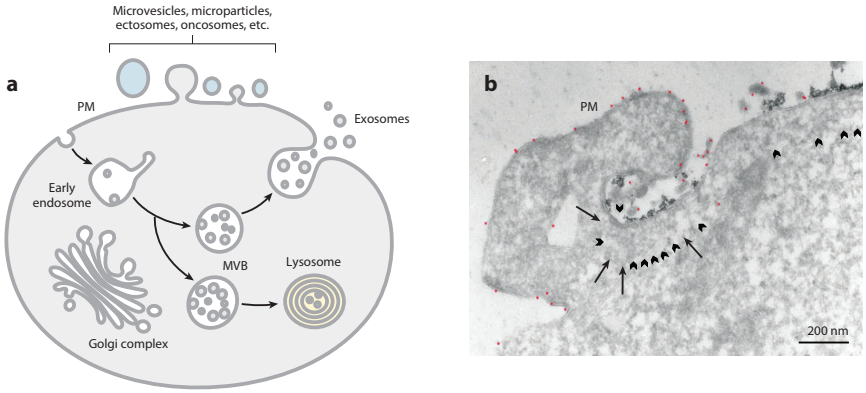


Figure 2.1 A: *Schematic representation of exosomes biogenesis (PM: plasma membrane. MVB: Multi Vesicular Bodies).* ***B:*** *Transmission electron microscopy photograph showing the fusion of MVBs with the plasma membrane in B cells transformed with the Epstein-Barr virus. Adapted from Colombo et al., 2014.*

tion of ILVs is a tightly regulated process that in mammals is controlled by multiple, independent mechanisms, the best characterised one being mediated by the ESCRT. The main function of the ESCRT is to sort ubiquitinated proteins into ILVs for their subsequent sorting to the lysosomes for degradation (Klumperman and Raposo, 2014). However, a subpopulation of MVBs can escape this process and instead fuse with the plasma membrane, thereby releasing the ILVs, at this point called exosomes, into the extracellular space (**Figure 2.1B**; Raposo et al., 1996).

The ESCRT complex is composed of four multi-protein subunits named ESCRT-0, -I, -II and -III. ESCRT-0 associates with the endosomes through the recognition of phosphatidylinositol-3-phosphate, a phospholipid enriched on endosomal membranes (Schmidt and Teis, 2012), and then mediates the recruitment of ubiquitinated membrane proteins as well as ESCRT-I. In turn, ESCRT-I and ESCRT-II, in addition to interacting with ubiquitinated cargos, promote the inward budding of the endosomal membrane and stabilise the base of the budding ILV. Finally, ESCRT-III is responsible for confining the cargo proteins to the forming ILV, mediating their de-ubiquitination (for ubiquitin recycling) as well as favouring membrane budding/scission through mechanisms that are not completely understood (Wollert and Hurley, 2010; Hurley and Hanson, 2010). The final step in ILV maturation is mediated by the mechanoenzyme VPS4, which disassembles the ESCRT-III complex in an ATP-dependent way in order to recycle its subunits to the cytoplasm (Schmidt and Teis, 2012).

A recent work by Baietti et al. (2012), further investigated the process that leads to the sorting of cargo proteins to the ESCRT complex, discovering an unexpected role for the Syndecan–Syntenin–Alix axis. Syndecans are transmembrane proteoglycans carrying heparan sulphate chains which act as co-

receptors for numerous intracellular signalling pathways (Couchman, 2010). It was found that the soluble cytosolic protein SYNTENIN acts as an adaptor molecule for recruiting Syndecans to ALIX, which is a component of the ESCRT machinery (Baietti et al., 2012). This mechanism allows the specific sorting of Syndecan cargoes to exosomes, therefore modulating their content. In accordance with these results, it was recently found that heparanase, the enzyme that cleaves heparan sulphate, is able to modulate the Syndecan-Syntenin-Alix pathway and promote the biogenesis of exosomes by cleaving the heparan sulphate chains on syndecans and, therefore, inducing their internalisation (Roucourt et al., 2015).

Interestingly, there is evidence in support of exosome-biogenesis mechanisms that are independent of ESCRT. In fact, knock-down by siRNAs of multiple ESCRT subunits greatly impairs the maturation of the endosome compartments but does not completely abolish the formation of ILVs (Stuffers et al., 2009). The modification of lipids of the endosomal membrane was one of the first mechanisms proposed for the ESCRT-independent formation of exosomes, and it was shown that the enzyme neutral sphingomyelinase catalyses the hydrolyses of sphingomyelin to ceramide on the endosomal membranes, thus favouring the inward budding of the membrane (Trajkovic et al., 2008). Other ESCRT-independent routes of exosomes biogenesis and protein sorting were shown to be dependent on other proteins - in particular tetraspanins. For example, the protein LMP1 of the Epstein-Barr virus is released into exosomes because of its association with the exosomal tetraspanin CD63 (Verweij et al., 2011). Similarly, other exosomal proteins such as CD9 and CD82 can be "hijacked" by other proteins to facilitate their sorting to exosomes, and the overexpression of CD9 or CD82 was shown to increase the abundance of β-catenin inside exosomes (Chairoungdua et al., 2010). More recently, it was found that the binding partners of CD81 compose a large fraction of the exosomal proteome, leading to the speculation that CD81 might act as a binding platform for sorting proteins to exosomes (Perez-Hernandez et al., 2013).

2.4 Exosomes and EVs in physiology and pathology

The last decade has seen a substantial increase in the number of studies addressing the roles and functions of exosomes and EVs. Although the majority of the information on their physiological role derives from *in vitro* observations, there are strong indications that EVs and exosomes are involved in virtually all biological process.

The first indications on the functions of EVs date back to the 1970s, when it was proposed that they play a role in bone calcification (Anderson, 1969),

and the 1980s, when it was observed that vesicles derived from the plasma membrane of multiple cell types possess 5'-nucleotidase activity (Trams et al., 1981). Shortly after, Johnstone et al. (1987) showed that the release of exosomes is an important process during reticulocyte maturation, and suggested that it serves the function of shedding, or clearing, unwanted proteins. In more recent times the range of EV functions both in physiological and pathological processes has been greatly extended, and it is now clear that in addition to protein clearance (Johnstone et al., 1987) they are also involved in the immune response (Raposo et al., 1996; Bobrie et al., 2011; Mittelbrunn et al., 2011), spreading of viral infections (Wiley and Gummuluru, 2006; Bukong et al., 2014; Longatti et al., 2014a), cell-to-cell signalling (Eder, 2009; Li et al., 2013a), prion disease (Fevrier et al., 2004), cancer (Al-Nedawi et al., 2009) as well as several neurodegenerative disorders (Smith et al., 2014). In particular, it has emerged that EVs play a particularly prominent role in modulating the immune response. For example, Raposo et al. (1996) have shown that exosomes released by human and murine B lymphocytes carry on their surface the major histocompatibility complex (MHC) class II bound to antigens; this complex was shown to be able to induce an antigen specific response in T cells (Raposo et al., 1996). Following this early observation that exosomes contribute to antigen presentation, it was also found that Dendritic Cells (DCs) release antigen-bearing exosomes which induce an antigen-specific response *in vivo* in $CD4^+$ T lymphocytes (Théry et al., 2002b). The observation that vesicles transfer the MHC-antigen complex between DCs and from DCs to T lymphocytes, thereby amplifying the primary immune response, has been further confirmed by several studies in human and mouse both *in vivo* and *in vitro* (Arnold and Mannie, 1999; Bedford, 1999; Patel et al., 1999; Nolte-'t Hoen et al., 2009). It was also observed that DCs exchange exosomes containing different populations of miRNAs depending on their maturation stage, which are functional in the recipient DC and predicted to affect pathways such as differentiation, cytokine synthesis and TGF-β signalling (Montecalvo et al., 2012). It was reported that a similar exosome mediated exchange of miRNAs takes place in a unidirectional way between T cells and antigen presenting cells in the context of the immune synapse (Mittelbrunn et al., 2011), also mediating modulation of the immune response. Other functions of exosomes related to immune stimulation include the induction of cytokine pathways (Bhatnagar and Schorey, 2007) and the direct transport of cytokines to recipient cells (Qu et al., 2007). Despite this evidence suggesting a role for EVs and exosomes in promoting the immune response, there are also numerous reports showing that they have immunosuppressive functions (Andreola, 2002; Clayton et al., 2007; Szajnik et al., 2010). Taken together, these data support contradictory

roles for EVs and exosomes in modulating the immune response, suggesting that their function is dependent on the cell type of origin, the target cell type and likely the microenvironment where the interaction takes place.

In addition to their physiological functions in the regulation of the immune response, exosomes and EVs have well established roles in the context of viral infections. For example, it was found that human B lymphocytes infected by the ɣ-herpesvirus Epstein-Barr virus (EBV) release exosomes containing several miRNAs of viral origin. These exosomes are then internalised by uninfected monocyte-derived DCs, where the viral miRNAs downregulate the expression of known EBV targets (Pegtel et al., 2010). More recently, it was also shown that exosomes secreted from cells infected by the human hepatitis C virus are able to spread the infection *in vitro* in a virion-independent way through the transfer of viral genomic RNA (Longatti et al., 2014b). In the context of Human Immunodeficiency Virus (HIV)-1 infection it was shown that exosomes released by peripheral blood mononuclear cells are able to transfer the viral co-receptor CCR5 to other cells that normally do no express it, therefore making them susceptible to viral infection (Mack et al., 2000). In addition, HIV-1 infected DCs are able to spread the infection by releasing viral particles associated with exosomes (Wiley and Gummuluru, 2006). Interestingly, these particles are ten times more infectious than cell-free HIV particles, therefore representing an efficient mechanism for viral trans infection. More recently, it was suggested that $CD4^+$ exosomes released by T lymphocytes act as decoys for HIV-1 infection, whereas the viral protein Nef is able to reduce the exosomal levels of CD4 and therefore reduce the inhibitory effects of these exosomes (Carvalho et al., 2014). In addition to viruses, other pathogens appear to have evolved mechanisms that exploit exosomes to modulate the host's immune response. For example, it was recently found that the gastrointestinal nematode *Heligmosomoides polygyrus* releases exosomes that suppress the innate immune response of infected mice (Buck et al., 2014).

Exosomes play further important roles in the context of cell-to-cell signalling in the central nervous system. EVs and exosomes are released by the majority of the cell types in the brain, and they are involved in processes such as synaptic function, microglia activation, endothelium activation and communication between neurons and glial cells (Cossetti et al., 2012; Smith et al., 2014; Iraci et al., 2016). However, in the brain EVs and exosomes are also relevant in pathological contexts. They contribute to the spread of prions (Fevrier et al., 2004), to the cell-to-cell transmission of α-synuclein in Parkinson's disease (Emmanouilidou et al., 2010), to the pathogenesis of Alzheimer's disease (Rajendran et al., 2006) as well as to the spread of inflammation in neuroinflammatory disorders such as multiple sclerosis (Sáenz-Cuesta et al., 2014).

Exosomes and EVs have been implicated in several other pathological conditions, most prominently in cancer, where numerous works have shown that EVs and/or exosomes released by cancer cells export cargos that promote tumour cell growth, proliferation and metastasis (Zhang et al., 2015).

2.5 Mechanisms of RNA secretion

The picture that emerged from the numerous works that have characterised the transcriptome of exosomes suggests the existence of dedicated secretion mechanisms that specifically load certain RNA species into exosomes while retaining others inside the cell. In fact, Valadi et al. (2007) have reported that certain mRNAs found within exosomes could not be detected in the donor cells. Similarly, it was also observed that certain miRNA species are significantly more abundant or exclusively present in exosomes compared to the parental cells, while others display the opposite trend and are depleted in exosomes (Mittelbrunn et al., 2011). It is also clear that cells can modulate the content of exosomes and EVs in response to external stimuli. For example, several works have shown that the RNA and protein repertoire of exosomes and EVs changes in response to perturbations and stress conditions, such as heat shock, hypothermia, hypoxia, oxidative stress or viral infections (Lancaster and Febbraio, 2005; Gastpar, 2005; Clayton, 2005; Taylor et al., 2007; Gupta and Knowlton, 2007; Zhan et al., 2009; Eldh et al., 2010; Jong et al., 2012; Beninson and Fleshner, 2014; Pegtel et al., 2010; Kalamvoki et al., 2014). These works indicate that the exosomal sorting machinery has the intrinsic capacity to sense external perturbations and react accordingly. For example, it is well established that the activation of intracellular signalling pathways induces robust and specific changes in the exosome content, and in some cases it also modulates their functions (Li et al., 2013a; Cossetti et al., 2014a; Kore and Abraham, 2014; Kato et al., 2014; Ekström et al., 2014; Jong et al., 2012; Squadrito et al., 2014).

Following these observations, various groups have tried to identify the molecular machinery responsible for the secretion of mRNAs and miRNAs, leading to remarkable advances in our understanding of this process. The following two paragraphs will describe the mechanisms so far described for the secretion of mRNAs and miRNAs respectively.

2.5.1 *Secretion of mRNAs*

Following the observation in 2007 that certain mRNAs are enriched in exosomes secreted by human and mouse mast cells (Valadi et al., 2007), Batagov et al. (2011) investigated for the first time the characteristics of secreted mRNAs.

This work reports that inside exosomes there is a significant enrichment for long non-coding RNAs, whereas RNA species abundant inside the cell are enriched for mRNAs (Batagov et al., 2011). Interestingly, it was also found that exosomal RNAs tend to have a shorter half-life than their cellular counterparts, and are enriched in 145 short (8nt) linear motifs. However, none of these motifs was detected in more than a small fraction of exosomal RNAs, the most frequent one being found in 24 % of the RNAs, suggesting that multiple motif combinations are important to drive RNA secretion. The structural analysis of the three most enriched motifs revealed that they are often embedded in structured RNA regions that form hairpins with an internal loop (Batagov et al., 2011). The hypothesis that structural features might be important for mRNA secretion was further supported in an independent work, which identified a 25nt structured motif in the 3' UTR of exosomal mRNAs (Bolukbasi et al., 2012). This motif was shown to fold into a stem loop that contained a core `CTGCC` motif in the loop as well as a binding site for miR-1289. By cloning the 25nt motif in the 3' UTR of a reporter vector in was shown that its presence is sufficient to drive the exosomal secretion of an RNA. Furthermore, it was also found that the binding of miR-1289 to its target site in the loop of the 25nt motif further increased RNA secretion (Bolukbasi et al., 2012).

These data are of particular interest because they suggest a functional connection between the miRISC complex and the process of extracellular RNA secretion. This idea is further supported by the observation that the miRISC-target complex accumulates into discrete cytoplasmic foci known as GW-bodies, which in turn have been shown to be associated with multivesicular bodies (Lee et al., 2009). Concurrent with the publication of these data, an independent work by Gibbings et al. (2009) confirmed that GW-bodies are enriched in GW182 and AGO2 and associated with multivesicular bodies. They also found that miRNAs and their targets are enriched at multivesicular bodies, and GW182 is enriched in exosomes (Gibbings et al., 2009). However, this work also reported that miRNA targets are under-represented in exosomes compared to whole-cell RNAs, questioning whether the miR-1289 mechanism described by Bolukbasi et al. (2012) is a general phenomenon.

More recently, Szostak et al. (2014) reported that none of the previously identified motifs was enriched in exosomes secreted by a mouse liver progenitor cell line, while they could find a new 12nt motif folded in a stem-loop structure enriched in the 3' UTR of secreted mRNAs. Additionally, they also found that cloning the motif in the 3' UTR of a luciferase reporter vector was sufficient to increase the secretion of luciferase mRNA (Szostak et al., 2014).

Taken together, these data clearly suggest the existence of a dedicated mechanism that drives the secretion of mRNAs, and this process seems to be physic-

ally and functionally coupled with the activity of miRISC. However, the studies that investigated this process are still limited in their number and discordant in their findings, suggesting that the full complexity of this mechanism is still to be discovered.

2.5.2 *Secretion of miRNAs*

Some of the earliest results describing the mechanisms that drive the secretion of miRNAs have been obtained by the laboratory of Francisco Sanchez-Madrid. Following their observation that certain miRNAs are enriched in exosomes (Mittelbrunn et al., 2011), they sought to discover the molecular machinery responsible for the process. Using multiple alignments and motif enrichment analysis of the miRNAs enriched in the exosomes secreted by human primary T cells, it was found that the majority of secreted miRNAs enriched in exosomes possess a short `GGAG` motif in the 3' of their mature sequence. Similarly, 66.6 % of miRNAs enriched in the whole cell (i.e. depleted in exosomes) carry a `UGCA` motif in the 3' end of their mature sequence (Villarroya-Beltri et al., 2013). Interestingly, the conversion of the secretion motif into a retention motif by site directed mutagenesis induced the cellular retention of a secreted miRNA. Conversely, mutating a retention motif into a secretion motif induced the secretion of a cellular miRNA. Additional experiments demonstrated that the secretion motif is recognised by the Heterogeneous nuclear ribonucleoprotein A2B1 (hnRNPA2B1), and that its binding to miRNAs is induced by its sumoylation (Villarroya-Beltri et al., 2013). The RNA binding protein hnRNPA2B1 has well established roles in mRNA localisation in neurons (Munro et al., 1999) as well as in viral genomic RNA localisation (Lévesque et al., 2006; Gordon et al., 2014). The data by Villarroya-Beltri et al. expand on its range of functions, implicating that it also acts as a carrier protein responsible for shuttling a specific subset of miRNAs toward exosomes.

Other groups have independently focused on the characterisation of miRNA secretion in different biological contexts. A recent study reported that in bone marrow derived macrophages, genetic depletion of DICER induced a decrease of miRNAs which is significantly more pronounced in exosomes than in cells (Squadrito et al., 2014). It was also found that the overexpression of miRNA targets decreases the abundance of the miRNA in exosomes, and shifts its subcellular localisation from the multivesicular bodies to P-bodies and/or GW-bodies (Squadrito et al., 2014). These data suggest that physiological changes in the expression of miRNA targets have the potential to modulate the subcellular localisation of miRNAs, and consequently influence their secretion in exosomes or retention in the cell. This hypothesis was confirmed on a tran-

scriptome wide level by showing that changes in the expression of mRNAs in response to IL-4 treatment of bone marrow derived macrophages explain the observed changes in the secretion of miRNAs (Squadrito et al., 2014). These data are in line with the observation by Gibbings et al. (2009) that miRNA targets are under-represented in exosomes compared to whole-cells.

A further and independent mechanism of miRNA secretion was suggested in 2014 by a separate study, which showed that non-templated nucleotide additions to the 3' end of miRNAs correlate with their secretion (Koppers-Lalic et al., 2014). In particular, it was found that in Epstein-Barr virus transformed lymphoblastoid B cells, miRNAs with non-templated adenylation at the 3' end were enriched in the cells compared to the exosomes, whereas miRNAs with 3' urydilations were found enriched in exosomes (Koppers-Lalic et al., 2014). These data suggest that nucleotidyl transferases, such as the uridyltransferase TUT4, might be able to modulate miRNA localisation and/or secretion. This work, however, does not prove a causal link between non-templated nucleotide additions and miRNA secretion; therefore, the correlative data reported might be the consequence of the differential distribution of nucleotidyl transferases between cells and exosomes rather than the cause of the differential distribution of miRNAs.

A study by Melo et al. (2014) further increased the apparent complexity of this process, showing that exosomes purified from the MCF-7 breast cancer line contain pre-miRNAs that undergo maturation after being secreted. This work also shows that exosomes contain components of the RISC loading complex, among which AGO2, TRBP as well as DICER, which is necessary for pre-miRNA processing inside exosomes (Melo et al., 2014). DICER was shown to interact with CD43, a plasma membrane anchor implicated in protein targeting to exosomes (Shen et al., 2011), and its silencing reduced the amount of DICER secreted in exosomes. These data show a role for CD43 in loading RISC-associated pre-miRNAs into exosomes, highlighting the complexity of this process.

In conclusion, the last three years have witnessed remarkable advances in our understanding of the mechanisms that drive miRNA secretion. However, these results do not point in a unique direction, suggesting that the selective secretion of miRNAs is a phenomenon that likely results from the interplay of multiple factors, such as presence of motifs, localisation of targets and interacting proteins. Our understanding of how these processes lead to the secretion of miRNAs is still very limited, and further studies will be required to reach a comprehensive understanding.

Part III

Results

3

Positional conservation identifies topological anchor point (tap)RNAs linked to developmental loci

The work presented in this chapter has been deposited on bioRxiv[1]. All the bioinformatic analysis are the result of my own work, except for the transcription factor binding analysis in pcRNA promoters, the pcRNA loop coverage plots, the motif enrichment and conservation analysis and the microarray meta-analysis in cancers. The experimental work has been conducted in the laboratory of Prof Kouzarides under the supervision of Dr Amaral. When I show results obtained by others the figure legends specify the author names.

3.1 Introduction

In recent years several independent lines of research have shown that the mammalian genome is pervasively transcribed to produce large numbers of long noncoding RNAs. Despite their abundance, lncRNAs are still largely uncharacterised from a functional point of view. More recently, numerous studies discovered, on an *ad hoc* basis, the function of a number of lncRNAs, revealing that they have important roles in physiological and pathological processes and act through a variety of molecular mechanisms, which for nuclear lncRNAs typically involve epigenetic regulation (Rinn and Chang, 2012). However, there are still no precise criteria that allow to infer the function of a lncRNA from its sequence or its genomic context. Interestingly, lncRNAs tend to be less conserved than typical protein coding genes (Carninci et al., 2005; Guttman et al., 2009; Derrien et al., 2012; Kutter et al., 2012; Marques and Ponting, 2009; Iyer et al., 2015), but their promoters display high conservation (Carninci et al., 2005; Guttman et al., 2009; Derrien et al., 2012; Necsulea et al., 2014) and their genomic location appears to be syntenically conserved across species (Carninci et al., 2005; Engstrom et al., 2006; Lipovich et al., 2006; Dinger et al., 2008b; Ulitsky et al., 2011; Necsulea et al., 2014; Hezroni et al., 2015). For example, the lncRNA *SOX2OT*, which overlaps the transcription factor *SOX2*, has a conserved promoter and displays the same expression patterns across all vertebrates (Amaral et al., 2009). Several recent studies have revealed that these syntenic lncRNAs often have the ability to regulate the expression of the neighbouring protein coding genes, as in the case of *WT1-AS*

[1] `http://dx.doi.org/10.1101/051052`

(Dallosso et al., 2007), *EVF2* (Feng et al., 2006), *SOX2OT* (Amaral et al., 2009), *HOTTIP* (Wang et al., 2011), *EVX1AS* (Bell et al., 2016), and many others. Several hypothesis can be laid out to explain the syntenic conservation of lncRNAs. First, it is possible that in certain cases the positional conservation between lncRNAs and neighbouring coding genes is simply the consequence of sequence conservation in a region that produces non functional transcripts. Second, the syntenic conservation could be the consequence of a process of convergent evolution that led to the independent formation of transcripts that over time evolved similar regulatory functions. Third, it could be the consequence of a common evolutionary origin and shared functionality between the human and mouse lncRNAs.

In this work we used the positional conservation of lncRNAs across mammalian genomes as an indicator of functional relatedness across species. We identify 665 positionally conserved lncRNA (pcRNAs) promoters in the mouse and human genomes that are preserved in genomic position relative to orthologous protein coding genes. We find that pcRNAs are genomically associated with developmental transcription factors, with which they are co-expressed in a tissue-specific manner. Interestingly, we observed that the majority of pcRNAs are linked to chromatin organisation structures, overlapping binding sites for CTCF and residing at the anchor points of chromatin loops. We named this group of RNAs topological anchor point (tap)RNAs and we show that they possess short stretches of highly conserved sequence that are enriched in binding motifs for Zinc Finger proteins. Knock down experiments revealed that tapRNAs and their neighbouring protein-coding genes are functionally connected, regulating each other's expression and having similar influences on the metastatic phenotype of cancer cells *in vitro*. This work identifies positional conservation as a functional indicator for lncRNAs and introduces the idea of an "extended gene" model, in which conserved developmental genes are genomically and functionally linked to regulatory lncRNA loci across mammalian evolution.

3.2 Identification of positionally conserved RNAs in human and mouse

Despite the large number of annotated lncRNAs in mammalian genomes, our ability to assign them to specific functional categories is still extremely limited, mostly due to the lack of common features that predict functionality. Moreover, the sequence of lncRNAs often shows very little conservation across species, further complicating the task of identifying common functions.

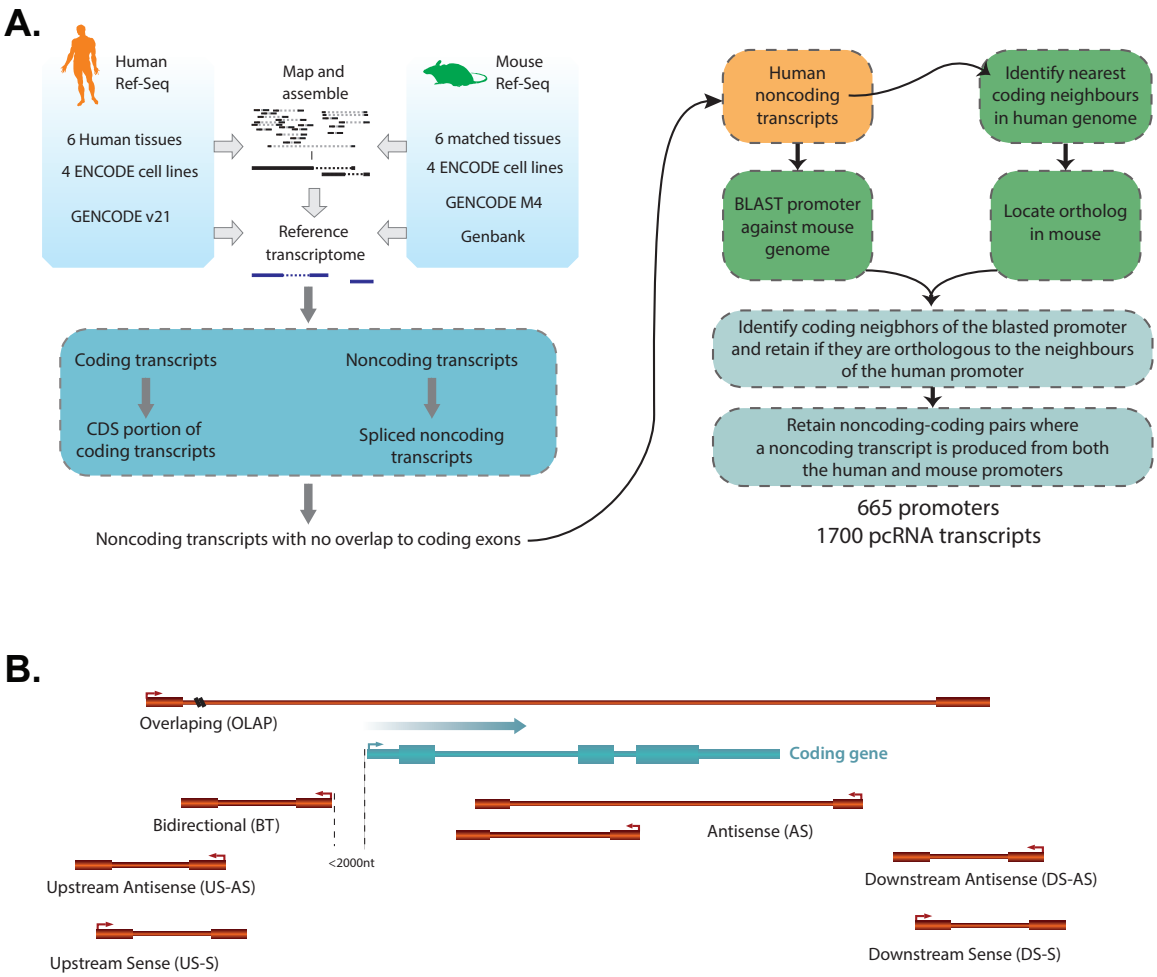


Figure 3.1 *A: Workflow used for the identification of pcRNAs.* ***B****: Schematic diagram showing the possible orientations of a pcRNA (red) relative to a coding gene (blue).*

In this work we considered the conserved position of a lncRNA relative to a neighbouring syntenic protein coding gene as an indicator of common, conserved function between species. We therefore implemented an analysis pipeline to identify spliced lncRNAs that are positionally conserved in the human and mouse genomes. We compiled a comprehensive catalogue of human and mouse transcripts based on 1) Gencode annotation (Harrow et al., 2012); 2) human and mouse RNA sequencing (RNA-Seq) from six matched tissues (brain, cerebellum, heart, kidney, liver and testis; Brawand et al., 2011); and 3) RNA-Seq data from four similar human and mouse cell lines (embryonic stem (ES), leukemia, lymphoblast and muscle cells) produced by the ENCODE project (Djebali et al., 2012; ENCODE Project Consortium et al., 2012). The list of all datasets used is available in Table 3.1, page 91. In total, we processed 80 RNA-Seq datasets and mapped 2.6 billion reads. We then implemented a pipeline that identifies human and mouse transcripts from both the Gencode annotation and the RNA-Seq data, and selects those with evidence of splicing, no overlap with coding exons in the same transcriptional orientation and no significant coding potential (**Figure 3.1A**). The sequence of the promoters of human lncRNAs were then aligned against the mouse genome in order to identify syntenic lncRNAs (see Methods, section 9.5). This approach was based on the observation that the promoters of lncRNAs tend to be highly conserved, even though the RNA sequence often shows little or no conservation (Carninci et al., 2005). We then annotated syntenic lncRNAs as positionally conserved if their promoters in both the human and mouse genomes were associated with orthologous protein-coding genes and produced spliced lncRNAs in the same relative transcriptional orientation (either sense or antisense relative to the coding gene) in both mouse and human (see Methods, section 9.5).

This approach led to the identification of 1700 positionally conserved lncRNAs (pcRNAs) transcribed from 665 distinct conserved promoters and associated with a total of 626 orthologous coding genes. The differences in these numbers reflect the fact that multiple pcRNA isoforms can be transcribed from the same promoter and multiple promoters can be associated with the same protein coding gene. The majority of pcRNAs (82 %, 1401/1700 transcripts, transcribed from 595 independent promoters) were already annotated in the Gencode human transcriptome, while 299 (18 %) represented novel transcripts assembled from the RNA-Seq data. A smaller fraction of pcRNAs (138 transcripts, transcribed from 32 independent promoters) overlapped syntenic miRNA loci, likely representing primary miRNA transcripts.

We then set to assign to each pcRNA a class based on its genomic orientation relative to the associated protein coding gene (**Figure 3.1B**): Antisense (AS,

direct overlap with the coding gene but transcribed from the opposite strand), Bidirectional (BT, transcribed in the opposite orientation with the transcription start site (TSS) within 2kb upstream the TSS of the coding gene), Overlapping (OLAP, partially overlapping the coding locus in the same orientation, with no overlap to coding exons and less than 50 % overlap with untranslated regions), and transcripts Upstream (US) or Downstream (DS) of the coding genes in either sense (S) or antisense (AS) orientation.

The majority of pcRNAs are Bidirectional transcripts (42 % of all pcRNAs), followed by Antisense (18 %), while all other categories are similarly represented between 5 % and 9 % (**Figure 3.2A**). The average length of pcRNA is 1.3kb and they are typically composed of 3-4 exons (mean 3.6 exons per pcRNA), with most having only 2 exons (**Figure 3.2B,C**). By definition, pcRNAs are in proximity of protein coding genes, but the distance between the pcRNA promoters and the TSS of the associated coding genes varies according to the positional category of the pcRNAs (**Figure 3.2D**). Approximately 70 % of pcRNAs are within 12kb of their associated genes. BT, AS and OLAP promoter positions are, as expected, closest (median TSS to TSS distances of 215bp, 520bp and 35bp, respectively), whereas promoters of upstream and downstream transcripts tend to be more distal (median TSS to TSS distances of 154kb (DS-S), 100kb (DS-AS), 44kb (US-AS), and 49kb (US-S)). In line with previous observations (Carninci et al., 2005; Guttman et al., 2009; Derrien et al., 2012; Necsulea et al., 2014), we found that pcRNAs tend to be less conserved than their associated coding genes. However, on average, human pcRNAs have 31 % sequence identity with their mouse counterparts (**Figure 3.2E**).

3.3 Positionally conserved RNA genes are associated with genes encoding developmental transcription factors

Our analysis identified 626 protein-coding genes associated with pcRNAs. To obtain a better idea of their function we performed a GO enrichment analysis and found a very strong enrichment for genes with roles in *Regulation of transcription from RNA polymerase II promoter* (GO:0045944 and GO:0000122, adjusted p-values = 1.2×10^{-15} and 8.5×10^{-10} respectively, **Figure 3.3** and Supplementary Table 3.2, page 93). In particular, we found significant enrichment for processes such as *Cell fate determination* (GO:0001709, adjusted p-value = 2.65×10^{-3}) and *Developmental induction* (GO:0031128, adjusted p-value = 2.65×10^{-3}), and in general these genes were part of a variety of developmental pathways, such as gastrulation, stem cell maintenance and organ morphogenesis (Supplementary Table 3.2, page 93). Notably, many of these genes belong to well-known gene families containing regulators of lineage spe-

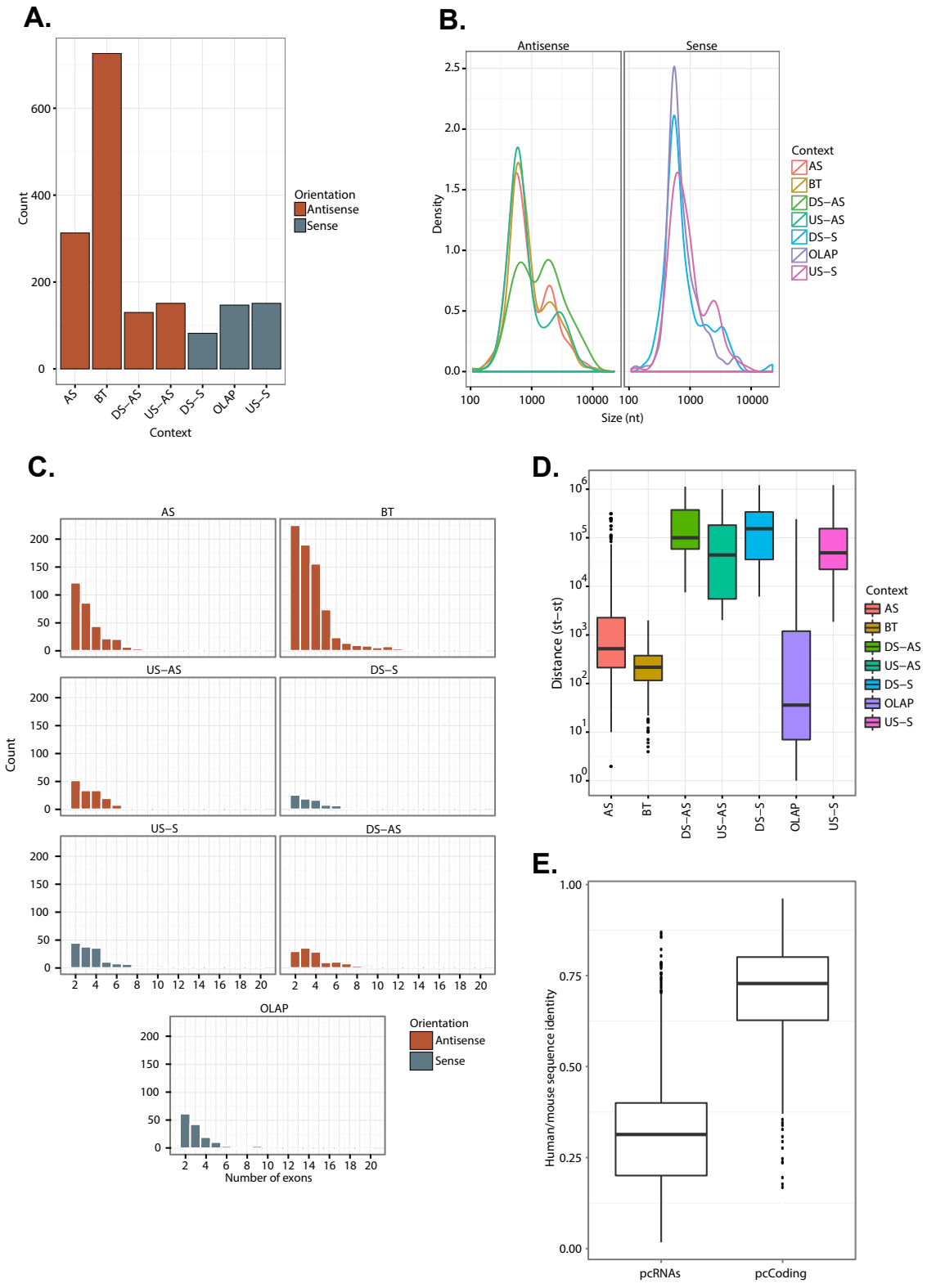


Figure 3.2 legend on next page

Figure 3.2 *(previous page)* ***A****: Bar chart showing the number of pcRNAs in each orientation.* ***B****: Density distribution of the distance between pcRNAs and respective coding genes, color-coded by positional orientation. The left plot shows pcRNA in antisense orientations, while the right plot shows pcRNAs in sense orientations.* ***C:*** *Bar chart showing exon-number distribution for each pcRNA.* ***D****: Boxplot showing the distribution of the distances between the TSS of pcRNAs and the TSS of their corresponding coding gene.* ***E:*** *Boxplot showing the fraction of sequence identity between human and mouse pcRNAs and human and mouse pcRNA-associated protein coding genes. Sequence identity was calculated with the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970).*

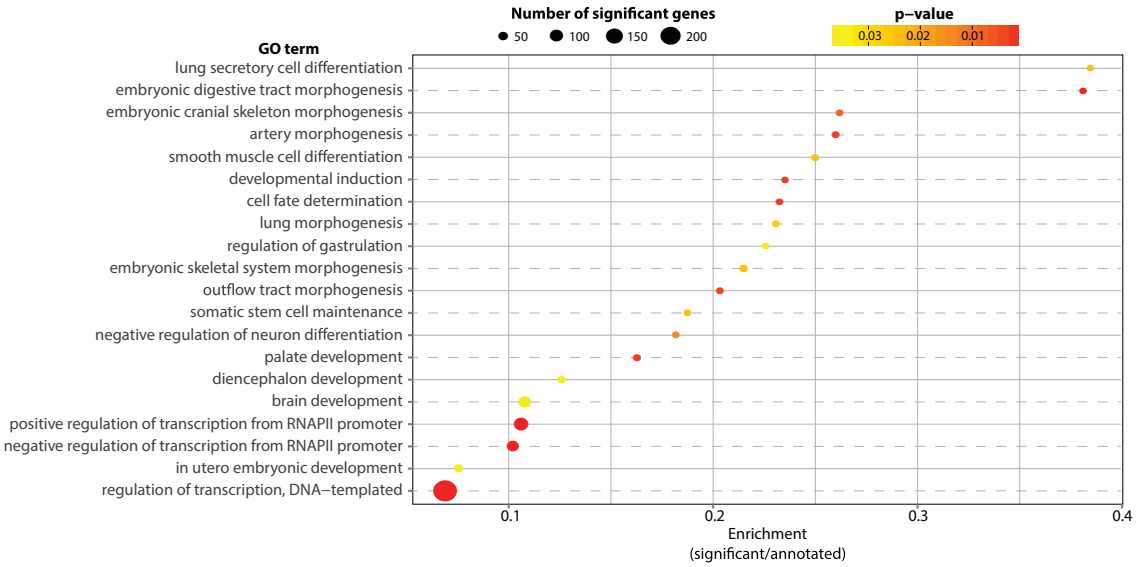


Figure 3.3 *GO enrichment analysis of pcRNA-associated coding genes. The x-axis shows the enrichment score, calculated as the number of pcRNA-associated genes in a given GO category divided by the total number of genes in the category. The size of the points indicates the absolute number of pcRNA-associated genes in the given GO category. The color-coding indicates the adjusted p-value. For full methods, see Methods, section 9.6*

cification, such as *SOX* genes (including *SOX1*, *2, 4, 9* and *21*); *FOX* genes (*FOXA2*, *D3*, *E3*, *F1*, *I* and *P4*); *HOX* genes (e.g. *HOXA1*, *A2*, *A3*, *A11*, *A13, B3, C5* and *D8*) and other homeodomain genes, as well as several nuclear receptors, such as *NR2E1*, *NR2F1* and *NR2F2*.

We found that pcRNA-associated coding genes were enriched in the aforementioned pathways disregarding the relative orientation of the pcRNA-coding gene pair. However, we also found that certain positional categories were enriched in specific pathways relative to all pcRNA-associated coding genes. For example, genes associated with bidirectional and overlapping pcRNAs were enriched for signal transduction and signalling pathways (such as *IGF2*, *TGFB2* and *PIK3R5*).

3.4 Positionally conserved RNAs have similar tissue specific expression patterns in mouse and human

We then used RNA-Seq data from a panel of human and mouse somatic tissues and cell lines (Table 3.1, page 91) to characterise the expression profiles of pcRNAs and associated coding genes. Consistent with previous observations (Cabili et al., 2011; Derrien et al., 2012; Dinger et al., 2008b; Ravasi et al., 2006), pcRNAs tend to be modestly expressed (**Figure 3.4A**), and their expression is usually restricted to one or a few tissues (**Figure 3.4B-D**).

We also found that the expression profiles of pcRNAs are similar in human and mouse tissues and cell lines, with a mean Spearman correlation of 0.26 (p-value $< 1 \times 10^{-6}$, **Figure 3.4E**). By calculating the correlation of expression between all pairs of pcRNAs we could identify numerous clusters of co-expressed pcRNAs, whose expression often peaked in the same tissue (**Figure 3.4B,C**), suggesting that pcRNAs may have conserved roles in tissue identity and cell type specification in mouse and human. For the majority of pcRNAs the expression was highest in testis, followed by total brain, ES cells and cerebellum (**Figure 3.4F**).

The protein coding genes associated with tissue-specific pcRNAs expressed in any given tissue tended to be enriched for GO terms involved in developmental and differentiation processes relevant to the particular tissue, such as neural differentiation genes in brain and endoderm developmental genes in liver (**Figure 3.5A-D**).

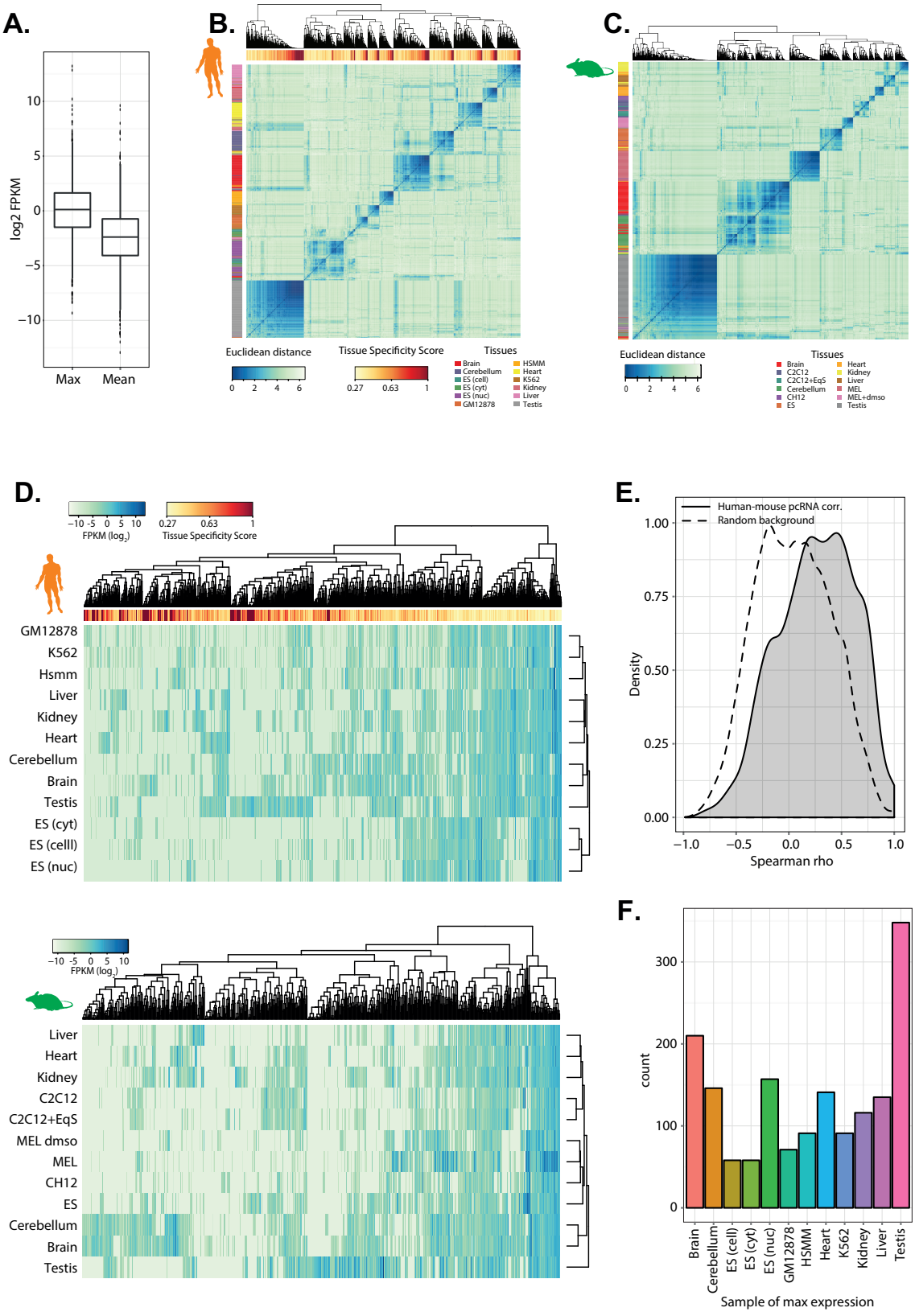


Figure 3.4 legend on next page

Figure 3.4 *(previous page)* ***A:*** *Boxplot showing the distribution of the highest FPKM measured across all samples for each pcRNA (left) and the mean FPKM across all samples for each pcRNA (right).* ***B,C:*** *Heatmap showing the Euclidean distance between the expression profiles of human (B) and mouse (C) pcRNAs. The horizontal side bar reports the tissue specificity score of pcRNAs. The vertical sidebar reports the tissues in which each pcRNA has maximal expression.* ***D:*** *Heatmap showing the expression profiles of human (top) and mouse (bottom) pcRNAs across tissues and cell lines. The horizontal sidebar reports the tissue specificity score of pcRNAs, ranging from 0.27 (white) to 1 (red).* ***E:*** *Density distribution of the Spearman's correlation coefficients between human and mouse pcRNA pairs. Mean Spearman's rho between human and mouse 0.26, permutation test p-value $<1 \times 10^{-6}$. The dotted line shows the background distribution of all pairwise Spearman's correlations between human and mouse pcRNAs.* ***F:*** *Bar chart showing the number of pcRNAs (y-axis) detected to have the highest expression in each given tissue (x-axis).*

3.5 Positionally conserved RNAs and genomically associated coding genes are co-expressed and co-induced

Positionally conserved RNAs showed a significantly positive correlation with their associated coding genes (mean Spearman rho 0.25, p-value $< 1 \times 10^{-6}$ **Figure 3.6A**); in the majority of cases this result held true disregarding of the orientation of the pcRNA relative to the coding gene (**Figure 3.6A**, inset) as well as the distance between their Transcriptional Start Sites (TSSs) (**Figure 3.6B**). We then calculated for each pcRNA a tissue specificity score using an entropy-based metric based on the Jensen-Shannon divergence (Cabili et al., 2011), which ranges from 0 for pcRNAs equally expressed across all tissues to 1 for pcRNAs restricted to a single tissue (see Methods, section 9.6). We found that pcRNAs are significantly more tissue specific than their associated coding genes (mean tissue specificity scores of 0.55 and 0.37 for pcRNAs and associated coding genes respectively; p-value = 4.25×10^{-220}, **Figure 3.7A**). To take into account the lower expression level of pcRNAs, and hence their higher probability of being identified in a single tissue due to the stochasticity of RNA-Seq, we partitioned pcRNAs and coding genes into five subclasses with matched expression levels. This approach allowed us to confirm that pcRNAs have higher tissue specificity than their associated coding genes disregarding of their expression level (**Figure 3.7B,C**).

Taken together, these data indicate that the expression of pcRNAs and their corresponding coding genes might be subject to similar regulatory mechanisms. In fact, when we measured by qRT-PCR the expression of six pcRNAs and their associated coding genes (*HOXB5/6-AS*, *NR2F1-BT*, *TBX2-BT*, *SOX2-OT*, *HOXA5-7-AS* and *EVX1-AS*, selected based on expression level, literature and visual inspection of their genomic loci) in human NT2 (NTERA/D1) teratocarcinoma cells upon differentiation with all-trans retinoic acid (a widely-used system for regulation of Hox genes and ncRNAs Sessa et al., 2007) we

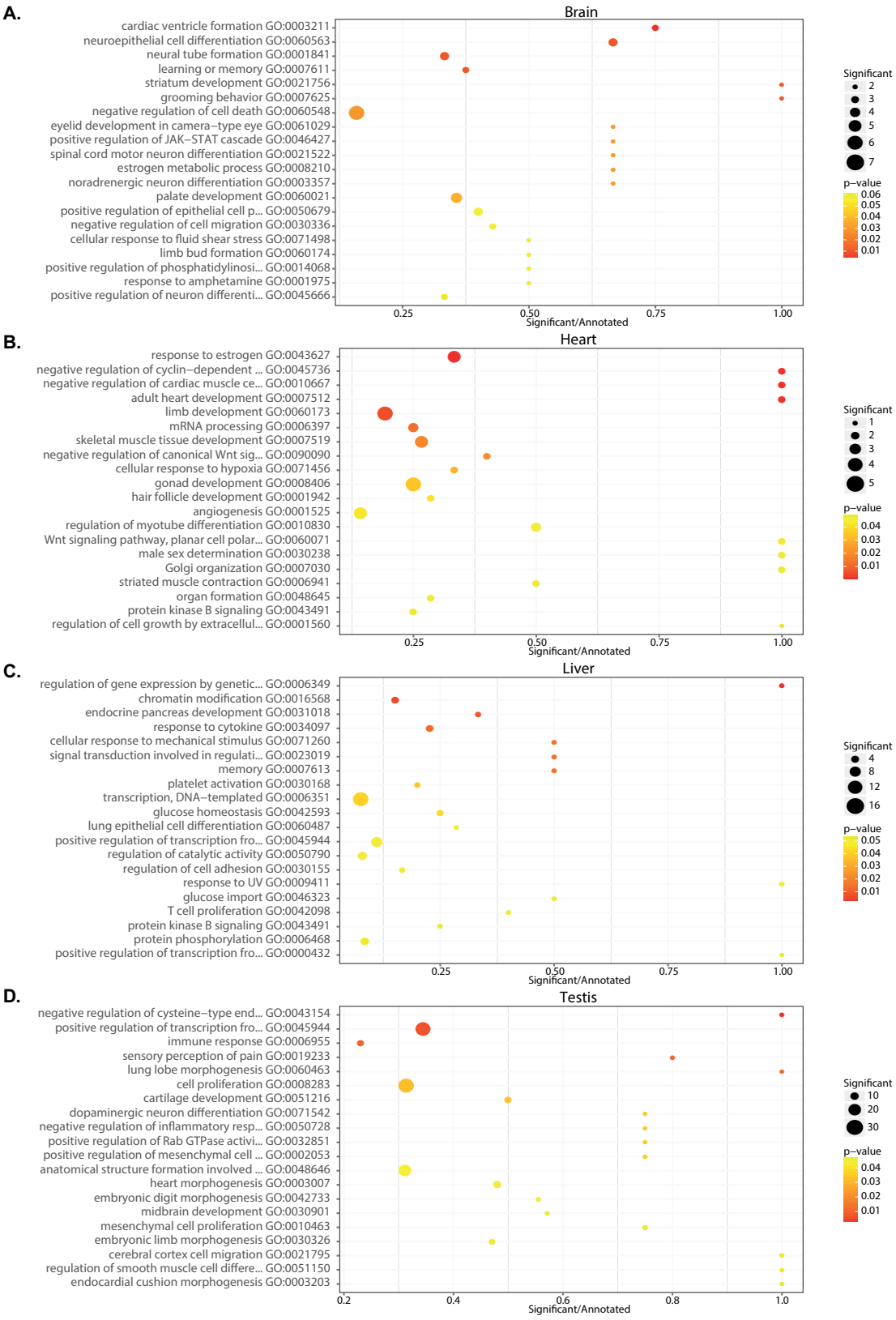


Figure 3.5 A-D: *GO enrichment analysis of coding genes associated to pcRNAs with expression specific for Brain (A), Heart (B), Liver (C), Testis (D). The x-axis shows the enrichment score, calculated as the number of pcRNA-associated genes in a given GO category divided by the total number of genes in the category. The size of the points indicates the absolute number of pcRNA-associated genes in the given GO category. The color-coding indicates the p-values.*

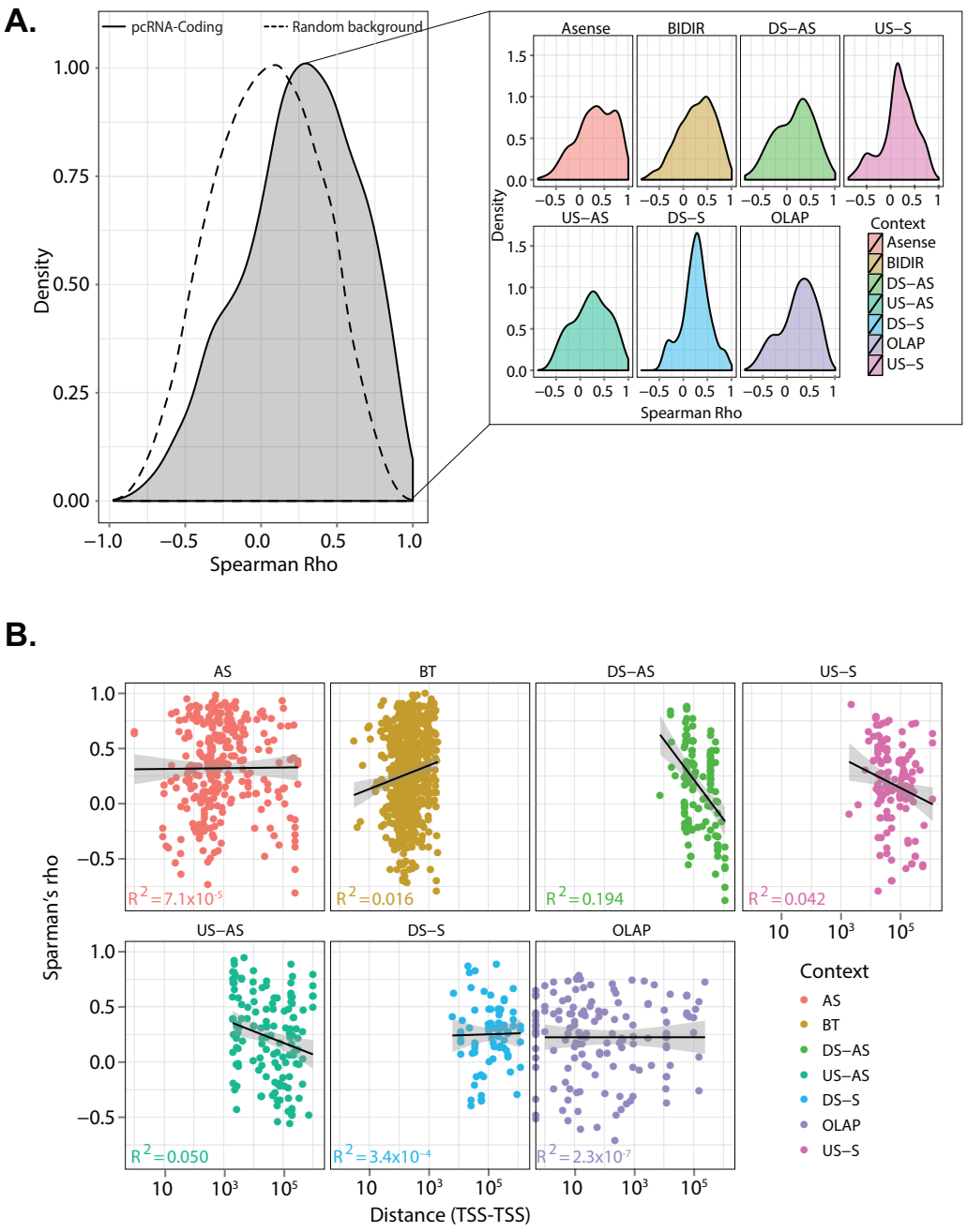


Figure 3.6 A: *Density distribution of the Spearman's correlation coefficients between pcRNAs and corresponding coding genes in human tissues and cell lines (mean Spearman's rho 0.25, permutation test p-value $<1 \times 10^{-6}$). The dotted line shows the background distribution of all pairwise Spearman's correlations between pcRNAs and pcRNA-associated coding genes. Inset: Distributions of the Spearman's correlation coefficients dived by the positional category of the pcRNA.* ***B:*** *Plot showing the Spearman correlation coefficient between the expression of pcRNAs and their corresponding coding genes as a function of their distance (TSS to TSS), indicating independence of TSS to TSS distance (R^2=0.008, p-value = 3.23×10^{-4}). The black lines represent the linear fit.*

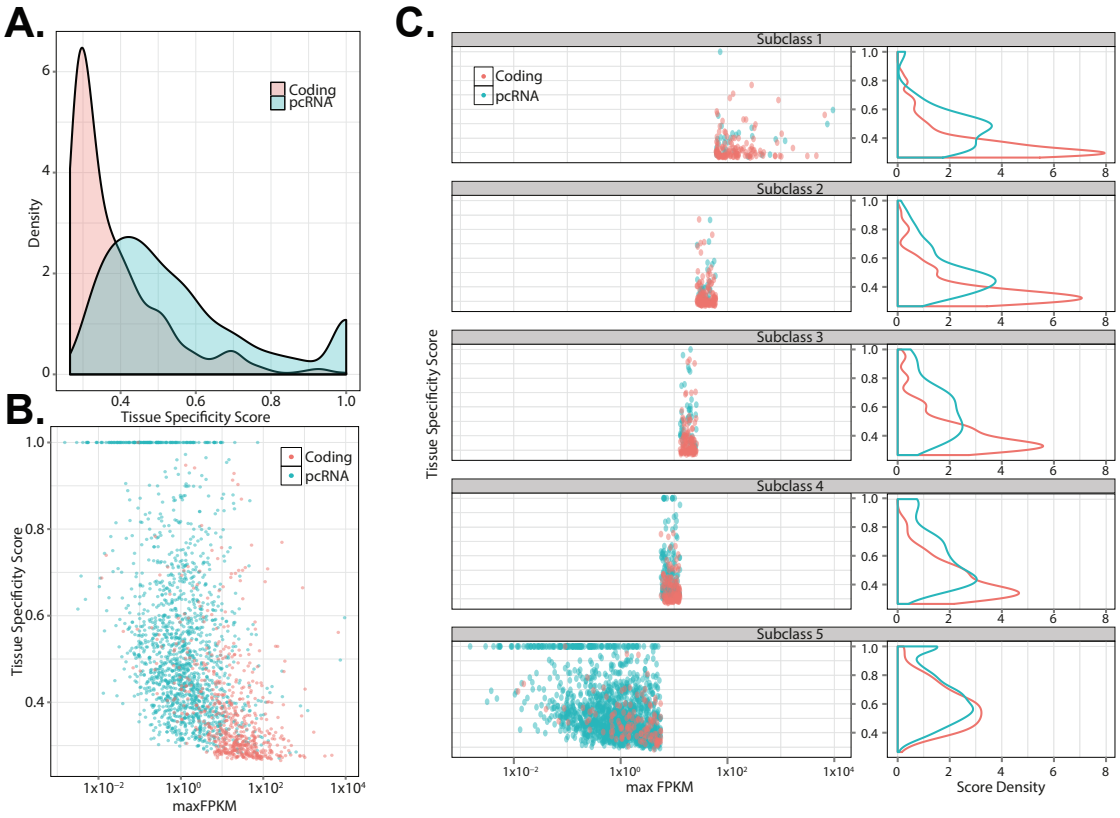


Figure 3.7 *A: Density distribution of the Tissue Specificity Score (see Methods, section 9.6) for pcRNAs (blue) and pcRNA-associated coding genes (red) showing significant higher specificity for pcRNAs (mean pcRNA tissue specificity score 0.55, mean associated coding gene tissue specificity score 0.37, p-value = 4.25×10^{-220}, Wilcoxon test).* ***B:*** *Scatterplot showing the highest FPKM observed across tissues (x-axis) for pcRNAs (blue) and pcRNA-associated coding genes (red) plotted against their tissue specificity score (y-axis).* ***C:*** *Scatterplot and density distribution of Tissue Specificity Scores for pcRNAs (blue) and pcRNA-associated coding genes (red) divided into 5 expression sub-groups. Each of the five sub-plots only displays pcRNAs and coding genes with similar expression levels (see Methods, section 9.6) and shows the highest FPKM observed across tissues (x-axis) plotted against their tissue specificity score (y-axis). The right part of the plot shows the distribution of tissue specificity scores for each sub-group, showing that pcRNAs have higher tissue specificity score than pcRNA-associated coding genes independently of their expression level.*

found that they are co-induced in a similar manner (**Figure 3.8A-D**). For example *HOXB6* and the associated pcRNA *HOXB5/6-AS* are both expressed at a high level in kidneys only (**Figure 3.8A**) and are both induced at day four of differentiation upon treatment of NT2 cells with retinoic acid (**Figure 3.8B,C**).

To validate these results in a broader panel of human and mouse tissues and cell lines we used the NanoString® expression assay, which provides an orthogonal method for single molecule detection of targeted RNAs with high sensitivity and specificity (Geiss et al., 2008). We designed a custom codeset to probe fifty human and mouse manually selected pcRNAs and associated orthologous protein-coding genes, across an RNA panel of six matched human and mouse tissues (see Methods, section 9.7). We also sampled additional tissues and cells such as mouse eye, spinal cord, 4 embryonic developmental stages and pluripotent cell lines from both species at various time-points of differentiation with retinoic acid, and a panel of 18 human cancer cell lines.

This method allowed us to confirm the RNA-Seq data and to generalise to a broader panel of conditions the finding that pcRNAs and associated coding genes are often co-expressed (median Spearman's Rho between pcRNAs and corresponding coding genes 0.43 and 0.57 for human and mouse respectively, **Figure 3.9A,B**). Additionally, we also confirmed that the expression profiles of pcRNAs are conserved between human and mouse (median Spearman's Rho 0.5, **Figure 3.9C**). For example, we observed that the pcRNA *FOXA2-DS-S* and its associated coding gene *FOXA2* are expressed at very similar levels in all tissues tested (Spearman's Rho 0.54, **Figure 3.9D,G**), with the expression peaking in liver and lung. Very similar profiles were also observed in mouse for *Foxa2-DS-S* (Spearman's Rho 0.73 between human and mouse, **Figure 3.9F,G**) and *Foxa2* (Spearman's Rho 0.52 between human and mouse, **Figure 3.9F,G**). Similar results were also observed for *HNF1* and its pcRNA (**Figure 3.9H-K**).

We then calculated the correlation between all pairs of human pcRNAs and coding genes (**Figure 3.10A**) and then clustered the resulting correlation matrix with the Markov Cluster Algorithm (van Dongen, 2008) in order to identify clusters of co-expressed genes (see Methods, section 9.7). Interestingly, we found that pcRNAs often form clusters with functionally related tissue-specific genes (**Figure 3.10B**). For example, several of the master regulator transcription factors of endoderm differentiation (*HNF1A*, *FOXA2* and *HNF4A*, Odom et al., 2006) appear, together with their pcRNAs, in two connected clusters. Taken together, these data prompted us to further investigate whether pcRNAs are involved in the regulation of expression of the neighbouring protein coding genes.

To further characterise the mechanisms regulating pcRNA expression we analysed the chromatin modification profiles at their TSSs. To this end we used

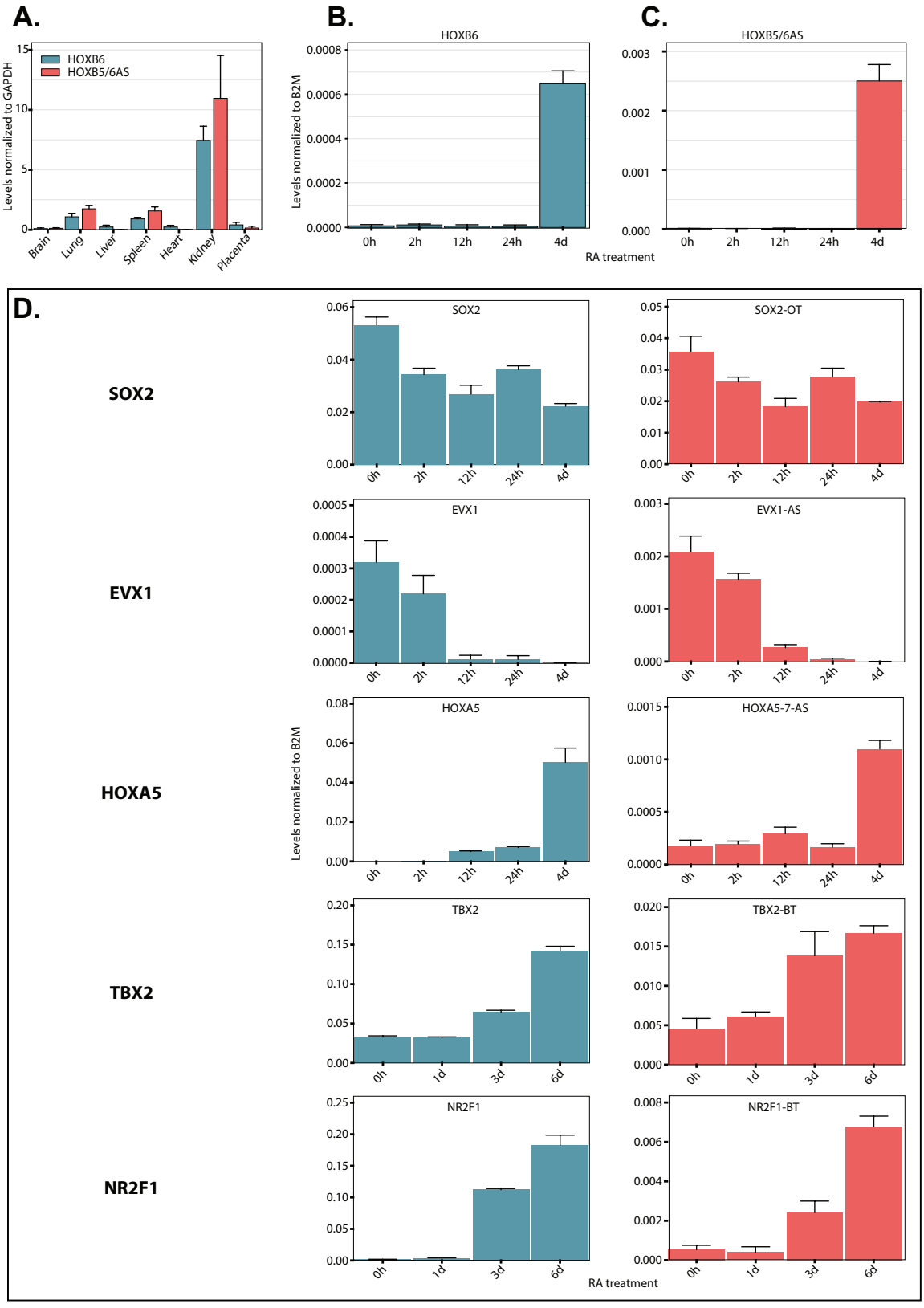


Figure 3.8 legend on next page

Figure 3.8 *(previous page)* ***A:*** *Real time PCR data showing the expression of HOXB6 (blue) and HOXB5/6AS (red) in a panel of 7 human somatic tissues. The data is expressed relative to the expression of GAPDH; the error bars indicate the standard error of the mean (SEM) across 3 technical replicate experiments.* ***B:*** *Real time PCR data showing the expression of HOXB6 (left) and HOXB5/6AS (right) over 5 time-points of NT2 cells differentiation with retinoic acid (RA). The data is expressed relative to the expression of B2M; the error bars indicate the standard error of the mean (SEM) across 3 replicate experiments.* ***C:*** *Real time PCR data showing the expression of SOX2, EVX1, HOXA5, TBX2, NR2F1 (left) and associated pcRNAs (right) over 5 time-points of NT2 cells differentiation with retinoic acid (RA). The data is expressed relative to the expression of B2M; the error bars indicate the standard error of the mean (SEM) across two replicate experiments. The data in this figure were obtained by Dr Amaral, Dr Viré and Ms Büscher in the laboratory of Prof Kouzarides.*

the ENCODE Chip-Seq data sets on the four tier 1 human cell lines (GM12878, H1-hESC, HSMM and K562, ENCODE Project Consortium et al., 2012) and we identified a clear enrichment in tri-methylation and di-methylation of Histone 3 lysine 4 (H3K4me3 and H3K4me2) as well as lysine 9 and 27 acetylation (H3K9ac, H3K27ac, **Figure 3.11A-D**). Interestingly, we also identified two embryonic stem cell specific signatures of pcRNA promoters displaying either high levels of both tri-methylation of H3K27 (H3K27me3) and H3K4me3 (bivalent promoters) or high levels of H3K27me3 and intermediate levels of H3K4me3 (**Figure 3.12A,B**). The pcRNAs clustered in these two groups show intermediate or no expression in ES cells respectively (**Figure 3.12C,D**). Both clusters are associated with developmental genes, but the bivalent cluster is particularly enriched in central nervous system development (**Figure 3.12E-H**). These results suggest that pcRNAs in this group are targets of Polycomb and silenced or transcriptionally poised in undifferentiated pluripotent cells (Bernstein et al., 2006). This observation is consistent with their roles in differentiation and development. On the other hand, we could not detect a peak of H3K4me1 in any of the cell types analysed.

Overall, these histone modification profiles are very similar to those observed for generic GENCODE annotated protein-coding genes, suggesting that pcRNAs are typical RNA Polymerase II transcripts and are not produced from enhancer regions (eRNAs). In fact, interrogation of the FANTOM5 Consortium database (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014) that annotates over 40,000 enhancer regions, identified the promoters of only three pcRNAs as enhancers (associated with *GATA2*, *HES1* and *KLF4*)[1].

[1] The analysis of FANTOM5 enhancer regions was done by Dr Zhang in the laboratory of Professor Shiekhattar

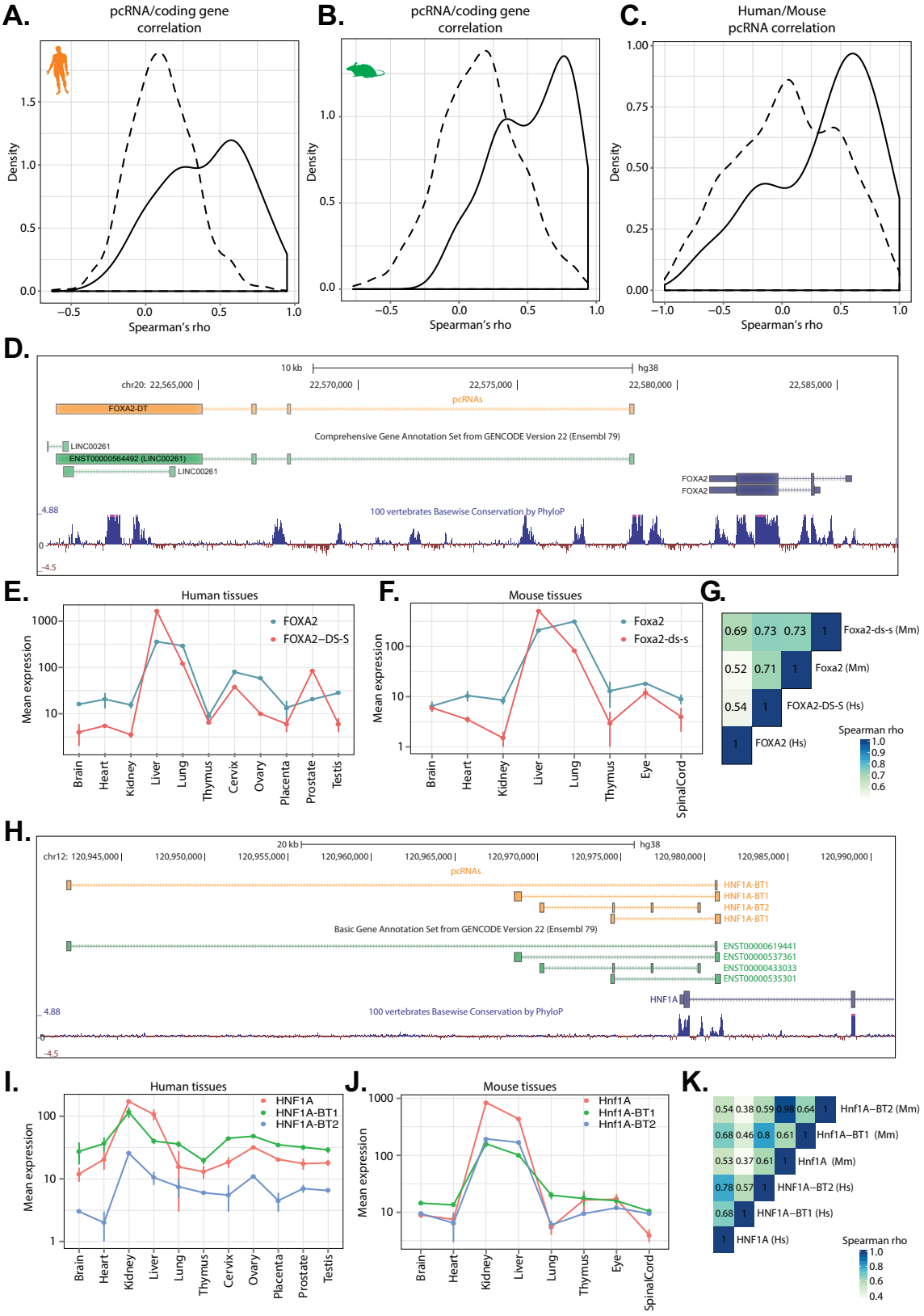


Figure 3.9 legend on next page

Figure 3.9 *(previous page)* ***A-B:*** *Density distribution of the Spearman's correlation coefficients calculated from NanoString® data for human (**A**) and mouse (**B**) pcRNAs and corresponding coding genes showing significant positive correlation (mean Spearman's rho 0.40 and 0.53 for human and mouse respectively, permutation test p-values $<1 \times 10^{-6}$). The dotted line shows the background distribution of all pairwise Spearman's correlations between pcRNAs and pcRNA-associated coding genes.* ***C:*** *Density distribution of the Spearman's correlation coefficients on NanoString® data between human and mouse pcRNAs pairs, showing conserved expression profiles across species (mean Spearman's rho 0.33, permutation test p-value $<1 \times 10^{-6}$).* ***D:*** *Illustration of the FOXA2 locus modified from a screenshot of the UCSC genome browser. For clarity, only one representative isoform of the coding gene is displayed.* ***E,F:*** *NanoString® expression profiles of FOXA2 and FOXA2-associated pcRNAs across human (**E**) and mouse (**F**) tissues. The plots report the mean value of two technical replicates, while the error bars report the value of each replicate.* ***G:*** *Heatmap showing Spearman's correlation coefficients between human and mouse FOXA2 and FOXA2-DS-S.* ***H-K:*** *as in D-G but for HNF1A and its pcRNAs HNF1A-BT1 and HNF1A-BT2. The samples for the NanoString® assay were prepared by Dr Amaral and the assay realised by NanoString® Inc.*

3.6 Identification of topological anchor point (tap)RNAs

Our expression analysis showed that pcRNAs and associated coding genes are co-expressed and co-regulated. To further explore the molecular determinants responsible for this co-regulation we analysed the promoters of pcRNA and associated genes for Transcription Factor (TF) binding profiles. To this end we made use of ChIP-Seq data for TFs generated by the ENCODE project (ENCODE Project Consortium et al., 2012) and found that the TF binding profiles for pcRNAs and corresponding coding genes are remarkably similar (Pearson correlation coefficient 0.67, p-value $< 10 \times 10^{-3}$, **Figure 3.13A**). This result was also confirmed when we used the presence of a known TF binding motif rather than a ChIP-Seq peak as an indicator of TF binding (Pearson correlation coefficient 0.63, p-value $< 10 \times 10^{-3}$, **Figure 3.13B**). These data show that the promoters of pcRNAs and associated coding genes are bound by highly similar groups of transcription factors, providing a likely explanation for their co-expression.

Interestingly, the characterisation of TF binding profiles in the promoters of pcRNAs revealed that the vast majority of them are bound by the CCCTC-binding factor (CTCF). In fact, we found that 72 % of pcRNA promoters contain a CTCF peak (**Figure 3.14A**), a significantly higher fraction than what we found for other spliced lncRNAs (p-value = 7.62×10^{-66}).

CTCF plays an important role in the topological organisation of the genome (Rao et al., 2014; Tang et al., 2015) and is one of the factors responsible for the formation of chromatin loops and Topologically Associating Domains (TADs). The three dimensional organisation of the genome is an important layer of regulation that controls, and is controlled by, gene expression (Cavalli and Misteli, 2013).

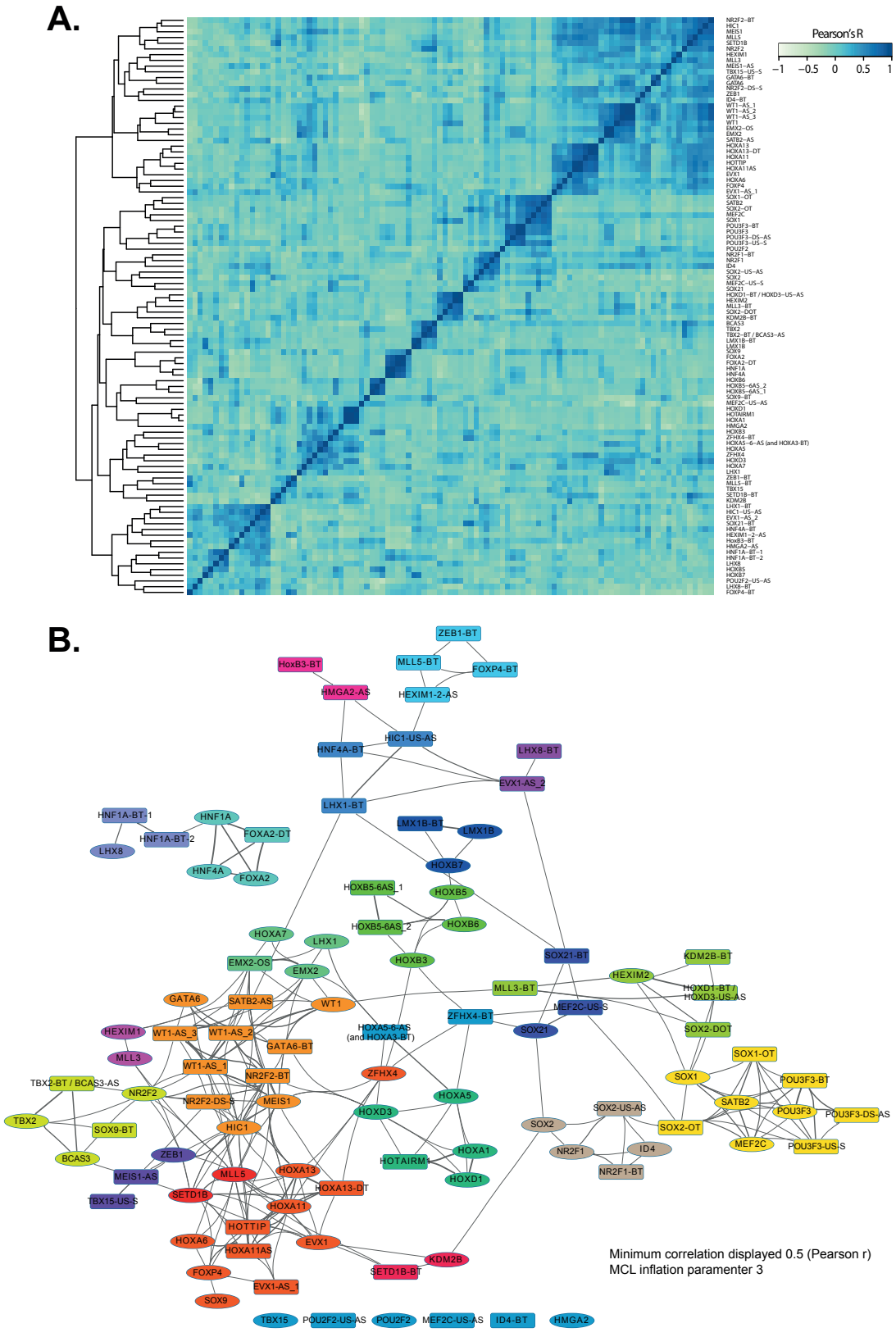


Figure 3.10 *A: Heatmap showing the pairwise Pearson correlation coefficients between all human transcripts included in the NanoString® experiment (both pcRNAs and pcRNA-associated coding genes).* ***B:*** *Network displaying all human transcripts included in the NanoString® experiment (nodes) and the Pearson correlation coefficient between their expression profiles (edges). Only edges with correlation coefficient higher than 0.5 are shown. The color coding of the nodes indicates the result of applying the Markov Clustering Algorithm to the matrix of correlation coefficients (see Methods, section 9.7). The samples for the NanoString® assay were prepared by Dr Amaral and the assay realised by NanoString® Inc.*

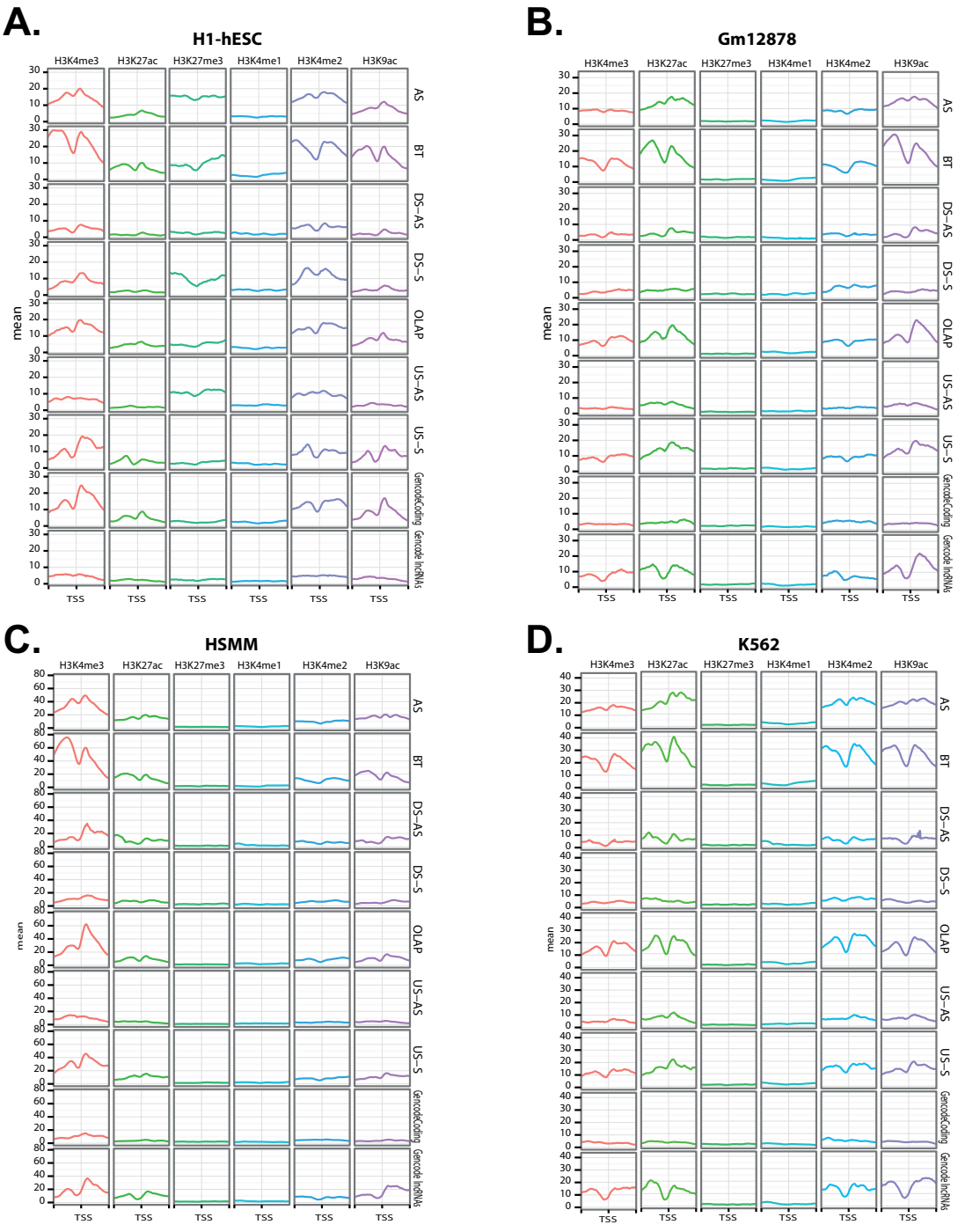


***Figure 3.11 A-D**: Histone modification profiles of pcRNA promoters (split by their relative orientation), promoters of 1000 random Gencode lncRNAs and promoters of 1000 random Gencode coding genes based on ChIP-Seq data by the ENCODE project on H1-hESCs (**A**), GM12878 (**B**), HSMM (**C**) and K562 (**D**). The lines represent the mean ChIP-Seq coverage and the shaded area around the line represents the standard deviation of the mean.*

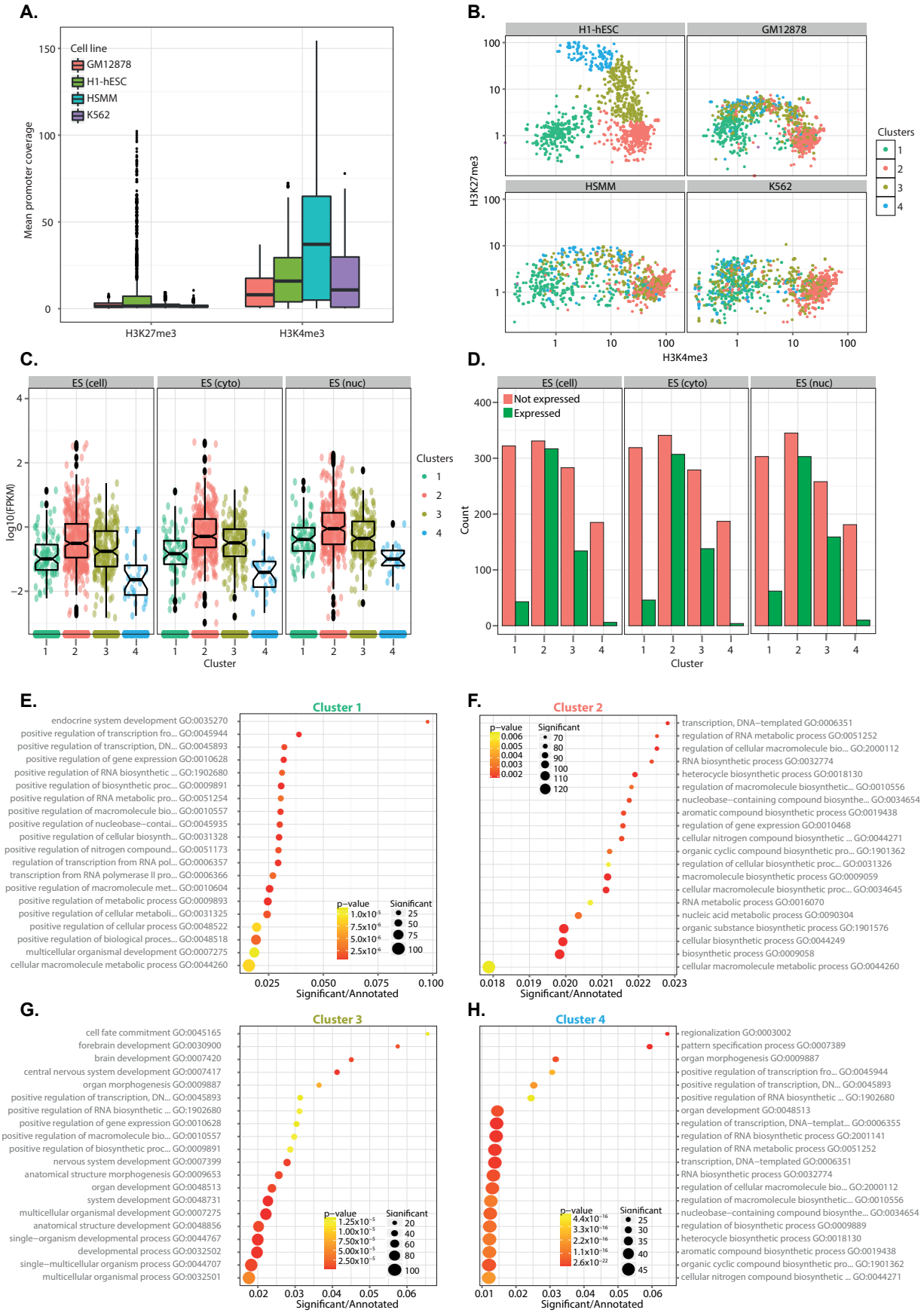


Figure 3.12 legend on next page

Figure 3.12 *(previous page)* ***A****: Boxplot showing the mean coverage of pcRNA promoters based on ChIP-Seq signal for H3K27me3 and H3K4me3 in GM12878, H1-hESCs, HSMM and K562.* ***B****: Scatter plot reporting the signal intensities of H3K4me3 (x-axis) and H3K27me3 (y-axis) in the promoters of pcRNAs. The four subplots represent data from H1-hESCs, GM12878, HSMM and K562. The colour coding reports the hierarchical clustering results. A single pcRNA had 0 H3K4me3 signal in H1hESCs and fell alone in a fifth cluster (not shown).* ***C****: Boxplot showing the expression (log_{10}FPKM) of pcRNAs based on RNA-Seq data on ES cells (left total cells; middle, cytoplasm; right, nucleus) and split by the cluster determined by applying hierarchical clustering to the H3K27me3 and H3K4me3 ChIP-Seq data (see Methods, section 9.11).* ***D****: Histograms showing the number of expressed pcRNAs based on RNA-Seq data on ES cells (left total cells; middle, cytoplasm; right, nucleus) and split by the cluster determined by applying hierarchical clustering to the H3K27me3 and H3K4me3 ChIP-Seq data (see Methods, section 9.11). pcRNAs with FPKM higher than 0.1 were considered expressed.* ***E-H:*** *GO enrichment analysis of coding genes associated to pcRNAs in each of the clusters determined by applying hierarchical clustering to the H3K27me3 and H3K4me3 ChIP-Seq data (see Methods, section 9.11). The x-axis shows the enrichment score, calculated as the number of pcRNA-associated genes in a given GO category divided by the total number of genes in the category. The size of the points indicates the absolute number of pcRNA-associated genes in the given GO category. The color-coding indicates the adjusted p-value.*

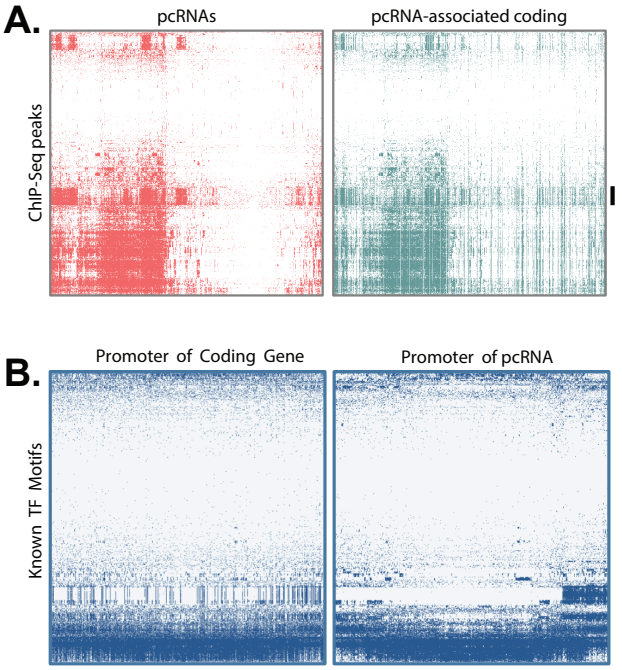


Figure 3.13 A,B: *Transcription factor binding ChIP-Seq peaks (A) or transcription factor binding motifs (B) in promoters of pcRNAs (left) and their associated coding genes (right). The heatmaps in (A) present the distribution of experimentally validated TF-binding sites from ENCODE 2,216 ChIP-seq experiments (y-axis), showing strong co-relations between the promoters of pcRNAs (x-axis) and their corresponding coding genes. The black bar indicates the binding pattern of the CCCTC-binding factor (CTCF). The heatmaps in (B) present known motifs from JASPAR (freeze 2014-12-10, 263 motifs) Kheradpour and Kellis, 2014 (2,065 motifs) and Jolma et al., 2013 (843 motifs). These analysis have been realised by Dr Han in the laboratory of Professor Kouzarides.*

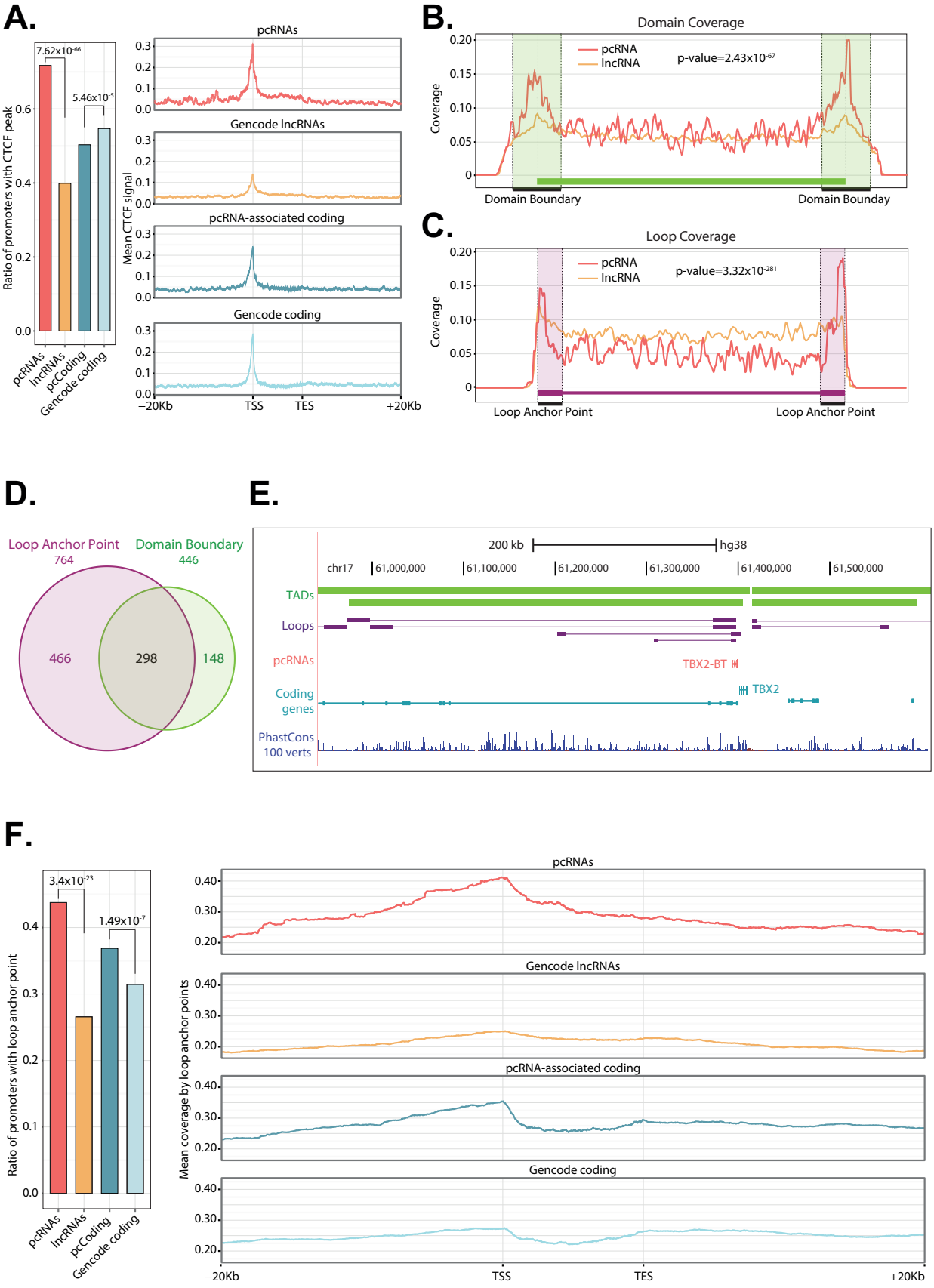


Figure 3.14 legend on next page

Figure 3.14 *(previous page)* ***A:*** *Bar chart showing the proportion of pcRNAs, pcRNA-associated coding genes, Gencode lncRNAs and Gencode coding genes with a CTCF peak (based on Encode ChIP-Seq data) overlapping their promoter. The p-values reported were calculated with hypergeometric tests. Right: CTCF peaks coverage of loci of pcRNAs, pcRNA-associated coding genes, Gencode lncRNAs and Gencode coding genes. The plots report the loci from 20kb upstream of the TSS to 20kb downstream of the transcription end site (TES). For visualization purposes these profiles show the coverage of a random sample of 5000 Gencode lncRNAs and 5000 random Gencode coding genes.* ***B,C:*** *Aggregation density plots showing the distribution of the TSS of pcRNAs (red) and lncRNAs (orange) relative to chromatin topological domains (**B**) and chromatin loop anchor points (**C**). Domains and loop anchor points were defined based on HiC data.* ***D****: Venn diagram showing the number of pcRNAs whose promoter overlap a Loop Anchor Point (purple) or a Domain Boundary (green)* ***E:*** *Schematic representation of the TBX2 locus showing the pcRNA TBX2-BT and chromatin loops defined by HiC data (Rao et al., 2014). Modified from a screenshot of the UCSC genome browser.* ***F:*** *Bar chart showing the proportion of pcRNAs, pcRNA-associated coding genes, Gencode lncRNAs and Gencode coding genes with a HiC loop overlapping their promoter. The p-values reported were calculated with hypergeometric tests. Right: HiC loops coverage of loci of pcRNAs, pcRNA-associated coding genes, Gencode lncRNAs and Gencode coding genes. The plotted genomic regions encompass the loci from 20kb upstream of the TSS to 20kb downstream of the transcription end site (TES). For visualization purposes these profiles show the coverage of a random sample of 5000 Gencode lncRNAs and 5000 random Gencode coding genes. The analysis in panels B and C have been realised by Dr Han in the laboratory of Prof Kouzarides.*

By analysing high resolution HiC data (Rao et al., 2014), that precisely maps the location of distal interactions across the human genome, we found that pcRNAs are preferentially located at the boundaries of TADs and chromatin loop contact points (or "loop anchor points") (**Figure 3.14B,C**). In particular, we noticed that 54 % of pcRNAs (912 out of 1700 pcRNA isoforms) have a promoter that overlaps a TAD boundary (446 pcRNAs) and/or directly intersects a loop anchor point (764 pcRNAs, **Figure 3.14D**). For example, the pcRNA *TBX2-BT* and other pcRNAs associated with important developmental genes lie at TAD boundaries and overlap multiple loop anchor points (**Figure 3.14E**).

Strikingly, we found that the promoters of pcRNAs are significantly more likely to overlap a TAD boundary or a loop anchor point compared to the promoters of Gencode spliced lncRNAs (p-value = 3.4×10^{-23}, **Figure 3.14F** and **Figure 3.15A**). Similarly, also the promoters of pcRNA-associated protein coding genes seem to be enriched in loop anchor points compared to the promoters of other Gencode protein-coding genes (p-value = 1.49×10^{-7}), although this enrichment is smaller than that observed for pcRNAs.

One of the most interesting features of the overlap between pcRNA promoters and loop anchor points is that they display a clear peak of enrichment in precise correspondence with the TSS of the pcRNA (**Figure 3.14F** and **Figure 3.15B**); this higher contact probability at the TSS is mirrored by a corresponding increase in CTCF binding (**Figure 3.14A**), suggesting that the association of pcRNAs with topological loops is non-coincidental.

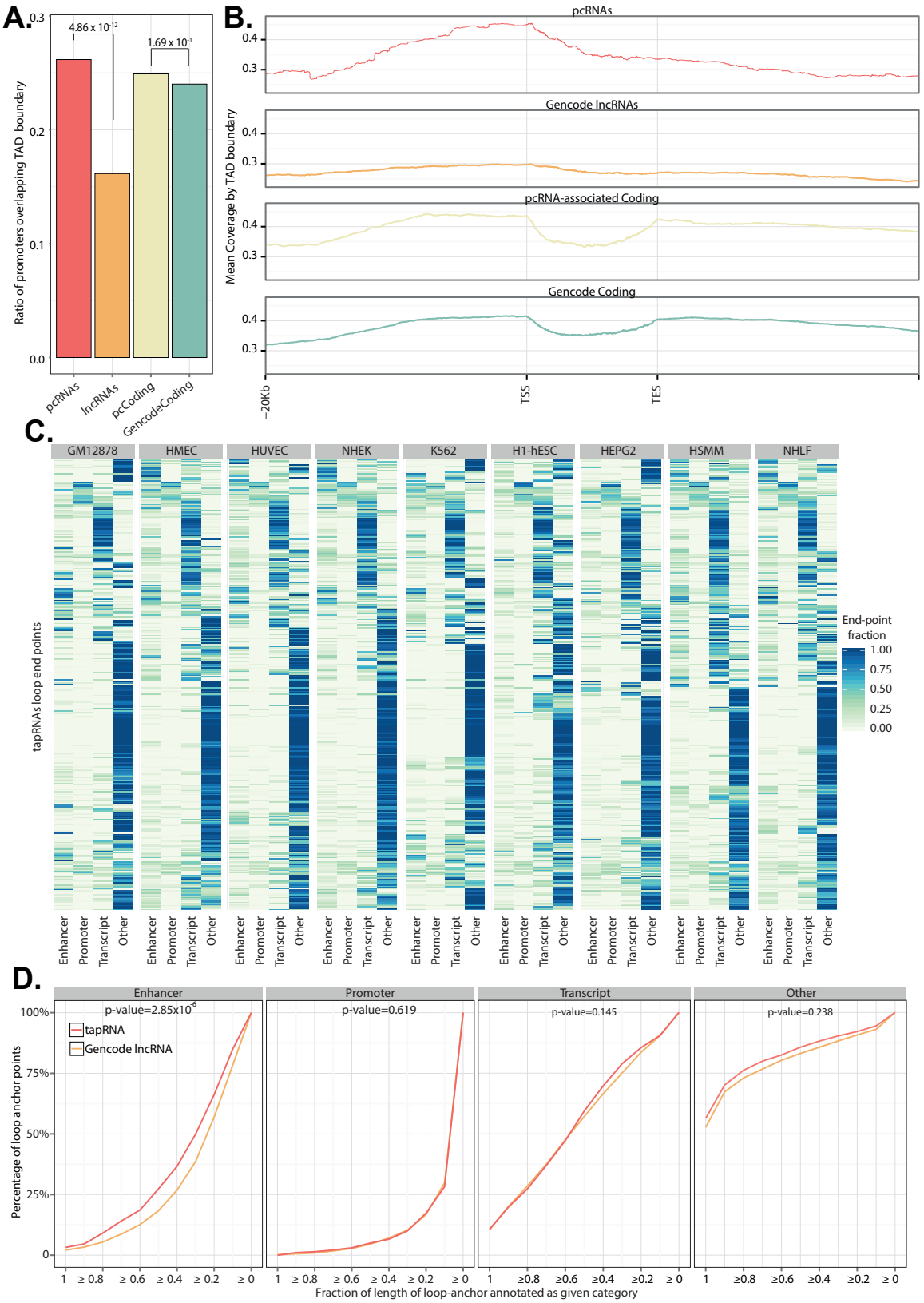


Figure 3.15 legend on next page

Figure 3.15 *(previous page)* ***A****: Bar chart showing the proportion of pcRNAs, pcRNA-associated coding genes, Gencode lncRNAs and Gencode coding genes with a TAD boundary overlapping their promoter. The p-values reported were calculated with hypergeometric tests.* ***B****: TAD boundary coverage of loci of pcRNAs, pcRNA-associated coding genes, Gencode lncRNAs and Gencode coding genes. The plots report the loci from 20kb upstream of the transcription start site (TSS) to 20kb downstream of the transcription end site (TES). For visualization purposes these profiles show the coverage of a random sample of 5000 Gencode lncRNAs and 5000 random Gencode coding genes* ***C****: Heatmap showing the proportion of each distal genomic region in contact with pcRNA promoter annotated in each genomic category derived from the ENCODE chromatin segmentation data (see Methods, section 9.15).* ***D****: Cumulative distribution plot showing the percentage of distal genomic regions in contact with pcRNA promoters (y-axis) as a function of the fraction of length of loop-end annotated as Enhancer, Promoter, Transcript or Other (see Methods, section 9.15). For example, the ">0.4" point (x-axis) of the red line in the first plot indicates that ~37 % (y-axis) of the distal genomic regions in contact with pcRNA promoters is annotated as Enhancer for 40 % or more of their length. Promoters of pcRNAs are significantly more often in contact through loops with enhancer elements compared to generic Gencode lncRNAs (p-value = 2.85×10^{-6}). The indicated p-values were calculated using the Kolmogorov-Smirnov test.*

In light of these observations, we defined the sub-group of 764 pcRNAs whose promoters overlap a loop anchor point "topological anchor point RNAs" (tapRNAs).

To obtain some indications of the potential roles of these tapRNAs and get insights into the possible functions of the promoter loops, we studied the characteristics of the distal regions that are brought into contact with the tapRNA promoters through the looping. To this end, we used a dataset of chromatin state segmentation generated by Hidden Markov Modelling for 9 cell lines (Ernst and Kellis, 2010). Due to their size of ~5kb, these loop anchor points are usually annotated in multiple chromatin states, but a significant proportion of them is usually marked as actively transcribed and/or enhancer (**Figure 3.15C**). Interestingly, we found that the distal regions interacting with tapRNA promoters have a significant enrichment for the *enhancer* state, compared to Gencode lncRNAs (p-value = 2.85×10^{-6}, **Figure 3.15D**). On the other hand, tapRNAs and Gencode lncRNAs are equally likely to interact with promoters, transcribed regions or other HMM defined genomic regions (**Figure 3.15D**).

Taken together these data show that tapRNAs are engaged in long-range interactions that bring their promoters in contact with distal enhancer regions.

3.7 Conserved domains and motifs in tapRNAs

When we examined the sequence of tapRNAs we found that they tend to be more conserved across vertebrates compared to generic Gencode lncRNAs, although this conservation is lower than that observed for protein coding genes (**Figure 3.16A**). This result prompted us to investigate whether there was

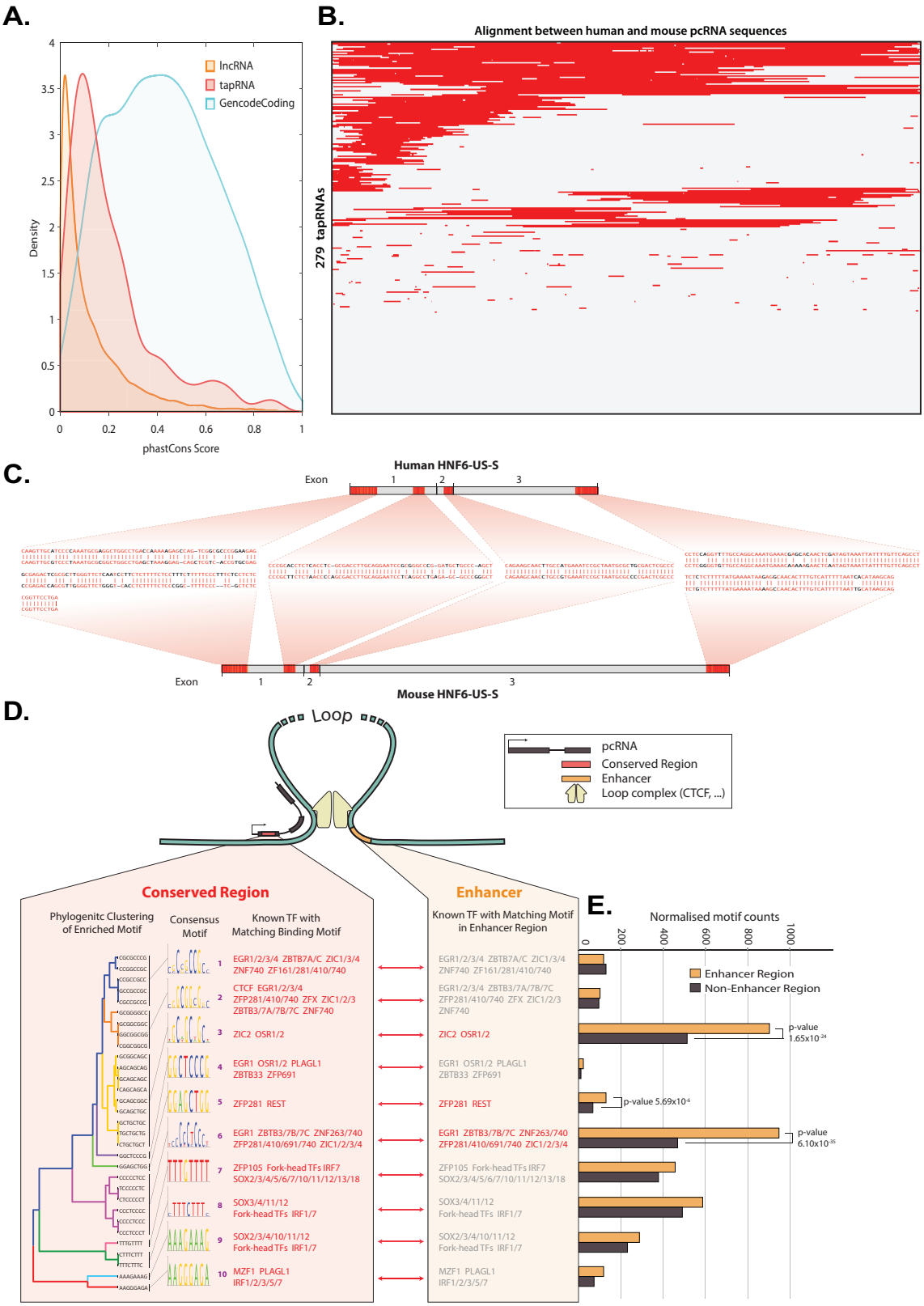


Figure 3.16 legend on next page

Figure 3.16 *(previous page)* ***A:*** *Comparison of conservation between tapRNAs, lncRNAs and protein coding genes. The curves are Kernel Density Estimation (KDE) of conservation scores calculated from the phastCons multiple alignments of 100 vertebrate species.* ***B:*** *Clustered heatmap of conserved domains in transcribed tapRNAs (see Methods, section 9.18). The red tracts indicate regions of human tapRNAs with an alignment in the mouse orthologs* ***C:*** *Example of conserved domains in a pcRNA. Identical sequence alignments (conserved domains) between human and mouse HNF6-US-S tapRNAs are represented in red, with RNA sequence alignments shown.* ***D:*** *Enriched TF-binding motif in both conserved domain of tapRNAs and enhancer region of loop anchor point. 32 significantly enriched 8-mer motifs (see* ***Figure 3.17;*** *p-value* $= 1 \times 10^{-4}$*) in conserved domains in tapRNAs are identified and clustered into 10 consensus motifs. De novo motif analysis discovers known TFs with matching binding consensus motifs. Seven out of ten consensus motifs are part of binding motifs of Zinc Finger proteins. The other three consensus motifs are part of binding motifs of developmental regulatory proteins.* ***E:*** *Extended motif search in enhancer regions of the other end of loop anchor points found significant enrichments of Zinc Finger protein motifs. These analysis have been realised by Dr Han in the laboratory of Prof Kouzarides.*

any conserved sequence feature important for the function of tapRNAs. To this end, we applied a sliding-window alignment approach to identify short patches of sequence conservation between human and mouse tapRNAs (see Methods, section 9.18). This analysis revealed that 73 % of tapRNAs show some extent of conservation, while the remaining 27 % appears to lack any recognisable conservation (**Figure 3.16B,C**).

The identification of short conserved stretches in the sequence of the majority of tapRNAs suggested the presence of conserved motifs. In fact, motif enrichment analysis revealed the presence of 32 motifs of 8nt that were significantly more frequent in the conserved regions of tapRNAs relative to the non-conserved ones (**Figure 3.16D** and **Figure 3.17**).

The alignment and clustering of these 32 motifs revealed that they could be organised as belonging to 10 consensus motifs that matched the known binding motifs of several transcription factors (**Figure 3.16D**). Interestingly, the predominant categories of transcription factors known to bind the 10 identified motifs are Zinc Finger (ZF) domain TFs (**Figure 3.16D**). Furthermore, when we inspected the ChIP-Seq data for these factors we did not detect binding peaks in the DNA underlying the motifs, suggesting the hypothesis that they might represent RNA-binding regions for the zinc finger transcription factors.

We next inspected whether the identified motifs also appeared enriched in the enhancer regions that are in contact with tapRNA promoters through looping. Interestingly, we found that three of the ten motifs identified are also enriched in the distal enhancer regions compared to the distal regions that are not annotated as enhancers (**Figure 3.16D,E**). These three motifs all match the known binding motifs of ZF proteins, among which the Zinc Finger Protein 143 (ZNF143) and the Zinc Finger Protein ZIC 2 (ZIC2), which are both

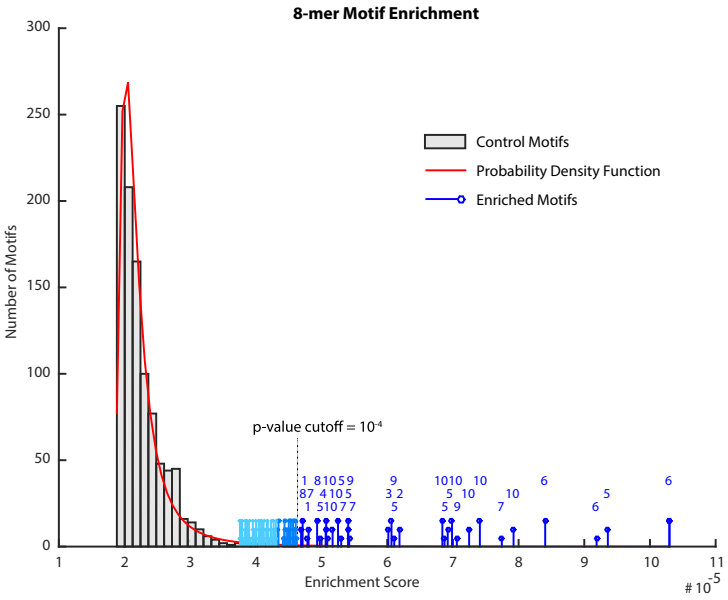


Figure 3.17 *Significantly enriched 8-mer motifs in conserved domains. Probability density function of Monte Carlo simulation results are shown in bar graph. The motifs that have p-value* $\leq 10 \times 10^{-4}$ *are considered as enriched motifs (shown in blue). The numbers on the enriched 8-mer motif stems are the consensus motif numbers as in* ***Figure 3.16D****. This analysis was performed by Dr Han in the laboratory of Prof Kouzarides.*

known to be involved in chromatin looping and enhancer function (Heidari et al., 2014; Luo et al., 2015; Xie et al., 2013a).

3.8 Functional analysis of positionally conserved RNAs

Co-regulation of gene expression The tissue-specific co-expression of pcRNAs and associated coding genes prompted us to investigate whether they could regulate each other's expression. The analysis of ChIP-Seq data (Ballester et al., 2014; Down et al., 2011) revealed that the promoter of the liver transcription factor *FOXA2* and the associated pcRNA *FOXA2-DS-S* share a highly concordant profile of transcription factor binding, both displaying binding peaks for key regulators of liver differentiation such as (*FOXA1*, *FOXA2*, *HNF4A* and *HNF6*, **Figure 3.18** and **Figure 3.19A**) This coordinated regulation of pcRNAs and associated coding genes by the same transcription factors might provide a molecular explanation for their observed co-expression and co-induction.

To further dissect the mechanisms regulating the expression of pcRNAs and coding genes we used RNAi to perform knock-down experiments of *FOXA2-DS-S*, finding that its down regulation causes a ~2.5 fold reduction in the expression level of *FOXA2* in Huh7 hepatocarcinoma cells (**Figure 3.19B**) and A549 lung adenocarcinoma cells (**Figure 3.20A**). Similarly, we also found that the knock-down of *FOXA2* also causes a decrease of expression of *FOXA2-DS-S* (**Figure 3.19B** and **Figure 3.20A**), consistent with what shown by the ChIP-Seq binding profiles (**Figure 3.19A**). These data were further supported by a microarray analysis of the genome-wide effects of FOXA2-DS-S or FOXA2 knock-down, which revealed a striking overlap (Jaccard similarity coefficient 0.61) in the set of genes differentially expressed in the two conditions (**Figure 3.19C,D**). In line with our results, a recent independent work showed that the lncRNA *FOXA2-DS-S* regulates the expression of *FOXA2* in differentiating definitive endoderm cells (Jiang et al., 2015).

In summary, these results indicate that the transcription factor *FOXA2*, in addition to regulating its own expression, also induces the expression of its neighbouring pcRNA *FOXA2-DS-S*. At the same time, *FOXA2-DS-S* is necessary to sustain the expression of *FOXA2* in different cell lines, thus establishing a local positive feed-back loop controlling *FOXA2*. Taken together, these data raise the possibility that this pcRNA acts *in cis* in order to mediate its regulatory effect on the neighbouring coding gene.

To expand an generalise this mode of *cis* regulation to other pcRNAs we repeated a similar set of experiments on *POU3F3-BT*, *NR2F1-BT*, *HNF1A-BT1* and their associated coding genes. In all these cases we found that the knock-down of the pcRNA significantly reduces the expression of the associated cod-

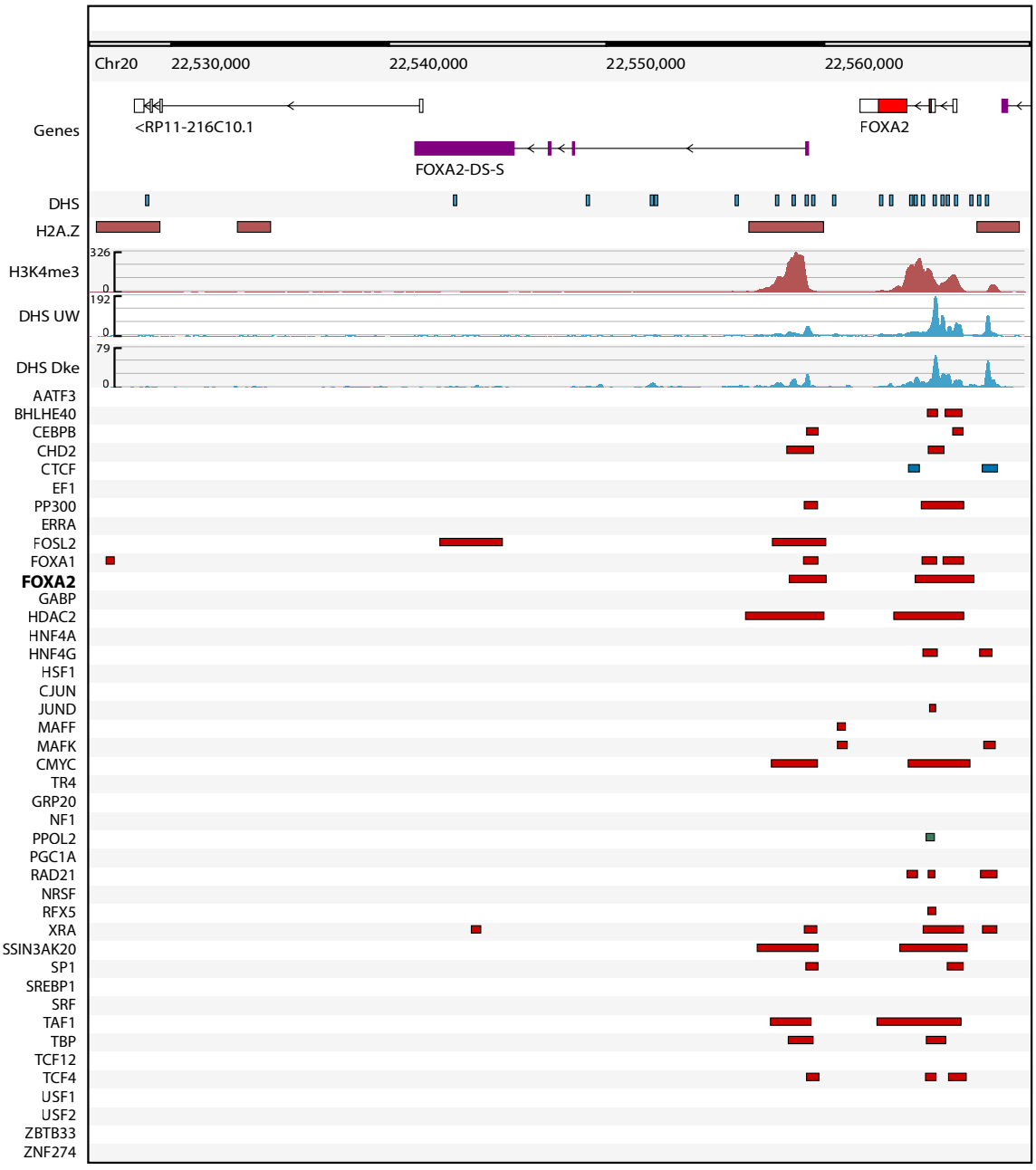


Figure 3.18 *Screenshot from the Dalliance genome browser (Down et al., 2011) showing the FOXA2 locus with tracks displaying coverage data for several ChIP-Seq experiments performed by the ENCODE project on HepG2 cells.*

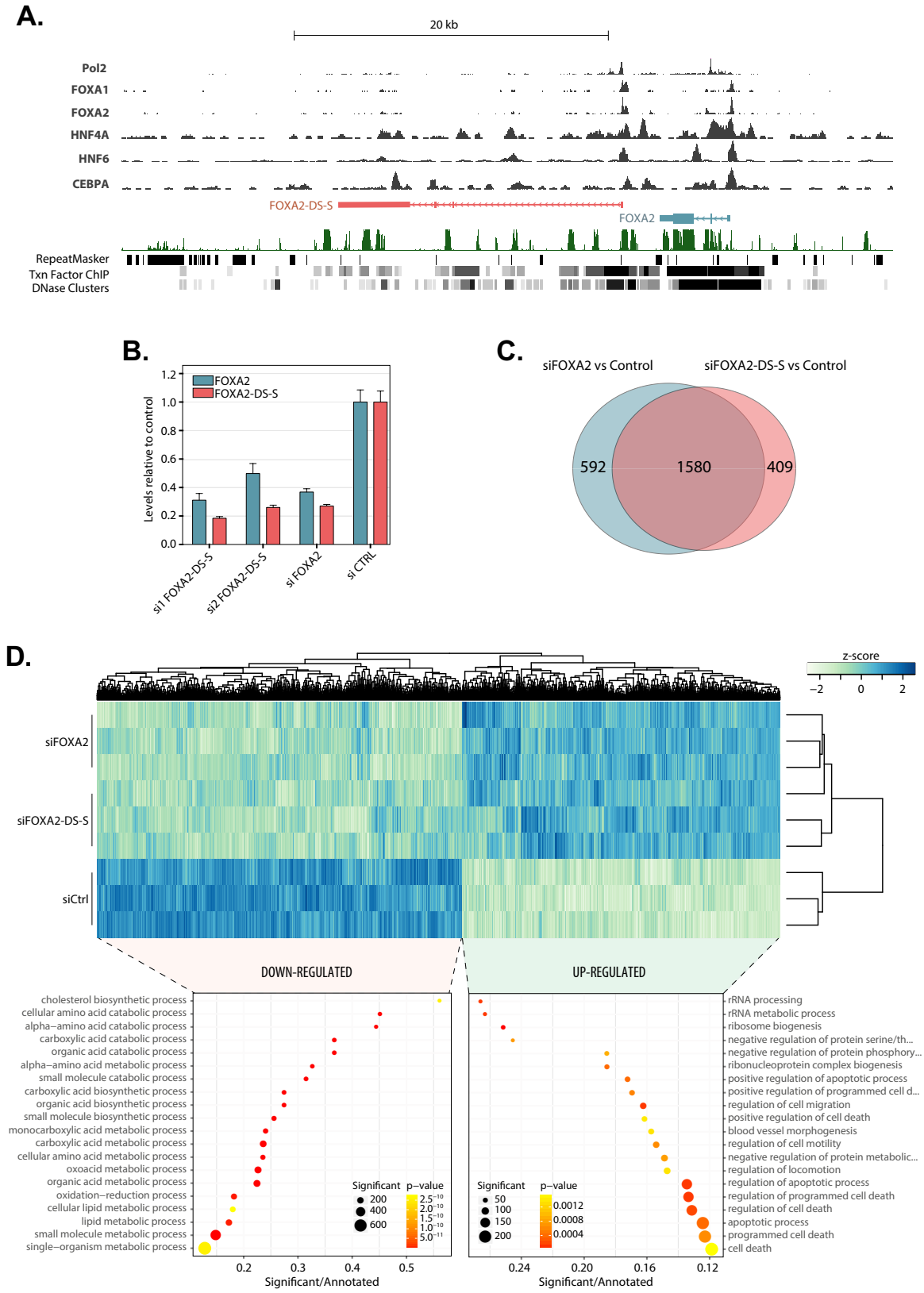


Figure 3.19 legend on next page

Figure 3.19 *(previous page)* ***A:*** *Screenshot from the Dalliance genome browser (Down et al., 2011) showing the FOXA2 locus with tracks displaying coverage data for ChIP-Seq experiments for Pol2, FOXA1, FOXA2, HNF4A, HNF6 and CEBPA. The ChIP-Seq tracks were produced by the ENCODE project on HepG2 cells.* ***B:*** *Real Time PCR data showing the expression of FOXA2 and FOXA2-DS-S in Huh7 cells upon knock-down. Si1- and si2- FOXA2-DS-S indicate two different, non-overlapping siRNAs designed against FOXA2-DS-S. The data is expressed relative to the expression of the control transfected with scrambled siRNAs; the error bars indicate the SEM across three replicate experiments.* ***C:*** *Venn diagram showing the number of significantly differentially expressed genes (adjusted p-value <0.05 and log_2 fold change > or < 1.25) in the microarray experiment on Huh7 knock-down of FOXA2 or FOXA-DS-S.* ***D:*** *Heatmap showing microarray data upon knock-down of FOXA2 or FOXA-DS-S in Huh7 cells. The color-scale indicates normalized intensities (z-score). The heatmap contains all genes that were significantly altered (adjusted p<0.05) upon knock-down of either FOXA2 or FOXA-DS-S. The scatter plots in the lower part of the panel show GO enrichment data for genes that were significantly down (left) or up-regulated (right) in either siFOXA2 or siFOXA-DS-S. The knock-down experiments (panel B and microarray) have been realised by Dr Amaral, Dr Viré and Ms Büscher in the laboratory of Prof Kouzarides.*

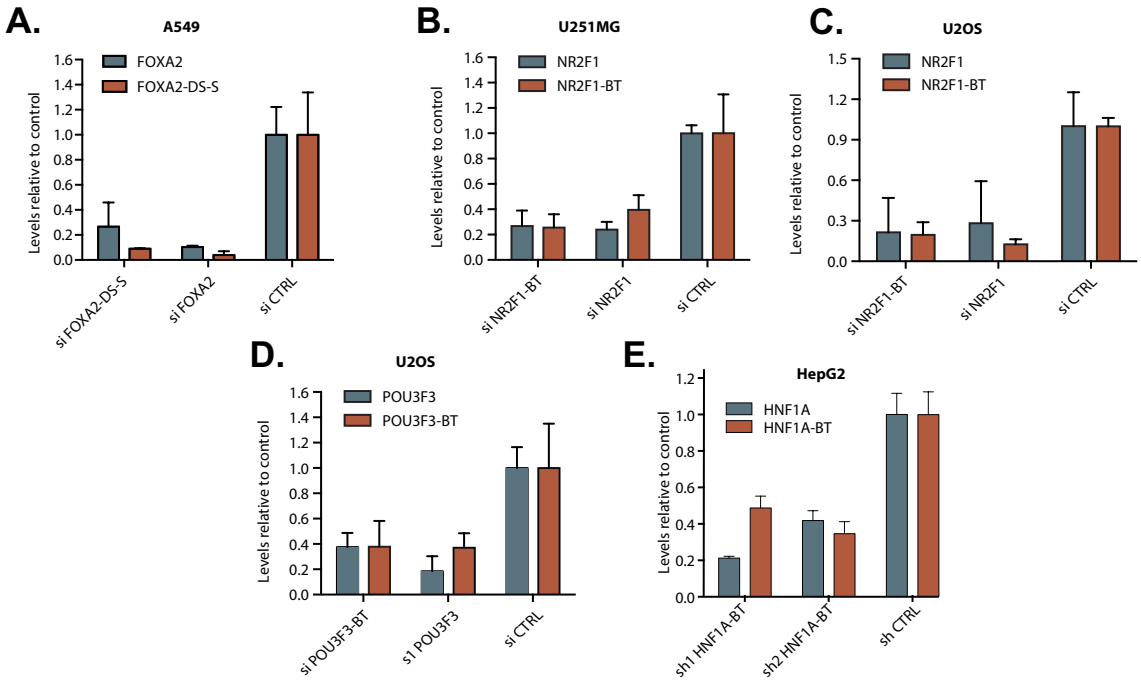


Figure 3.20 ***A-D****: Real time PCR data showing the expression of pcRNAs and associated coding genes upon knock-down of FOXA2-DS-S (**A**), NR2F1-BT (**B, C**), POU3F3-BT (**D**), HNF1A-BT (**E**) and their associated coding genes. These experiments have been realised by Dr Amaral, Dr Viré and Ms Büscher in the laboratory of Prof Kouzarides.*

ing gene (**Figure 3.20A-E**), suggesting that the regulation of neighbouring genes *in cis* might be a common mechanism of action for pcRNAs. To further support this *cis*-based, context-dependent regulation, we observed that the ectopic over-expression of full-length *HNF1A-BT* in human liver cells has no effect on the expression of the associated coding gene *HNF1A*.

Involvement in cancer The NanoString® analysis of pcRNA expression revealed that numerous pcRNAs and tapRNAs are differentially expressed – together with their associated coding genes – in different cancer cell lines (**Figure 3.21A-E**). These results, together with the notion that many lncRNAs have been implicated in cancer and other diseases (Amaral et al., 2013; Balbin et al., 2015), prompted us to further investigate the role of pcRNAs in this context.

By querying the expression of pcRNAs in a panel of 63 normal vs. tumour microarray studies, we found that the expression of 203 pcRNAs is significantly altered in cancer (**Figure 3.22A**). Some of these are well known lncRNAs with established roles in cancer (e.g. *GAS5*, *DLEU2*, *PART1* and *MEG3*, Pickard and Williams, 2015), while other, such as *FOXA2-DS-S*, had no previous cancer-related role.

Specifically, we observed that *FOXA2-DS-S* and the associated coding gene were both significantly downregulated in lung cancer compared to controls (p-value = 3×10^{-16} and 2×10^{-22} respectively, **Figure 3.22B**), in line with the recent observation that *FOXA2* might act as a tumour suppressor gene in hepatocellular carcinoma by inhibiting epithelial-to-mesenchymal transition (Tang et al., 2011; Wang et al., 2014). Our data also imply the pcRNA *FOXA2-DS-S* in this process, and we further show that its knock down in Huh7 and A549 cells greatly increases cell invasion and cell migration capacities *in vitro* (**Figure 3.22C** and **Figure 3.23A**).

In addition, we find consistent results for *NR2F1-BT* and *POU3F3-BT*, which upon knock down decrease the invasion and migration characteristics of glioblastoma (U251MG) and osteosarcoma (U2OS) cells in the same way as the knock down of their respective coding genes (**Figure 3.23C-D**).

Taken together, these results further support the idea that pcRNAs exert their effects through the regulation of the neighbouring protein coding genes *in cis*, and highlight their potential use as therapeutic targets or diagnostic markers in cancer and/or in any other pathological condition where the associated coding genes have a role.

In conclusion, in this project we have identified and characterised positionally conserved lncRNAs, showing that they are genomically associated, co-expressed and co-induced with developmental transcription factors. The majority of pcRNAs were found to possess binding sites for CTCF in their pro-

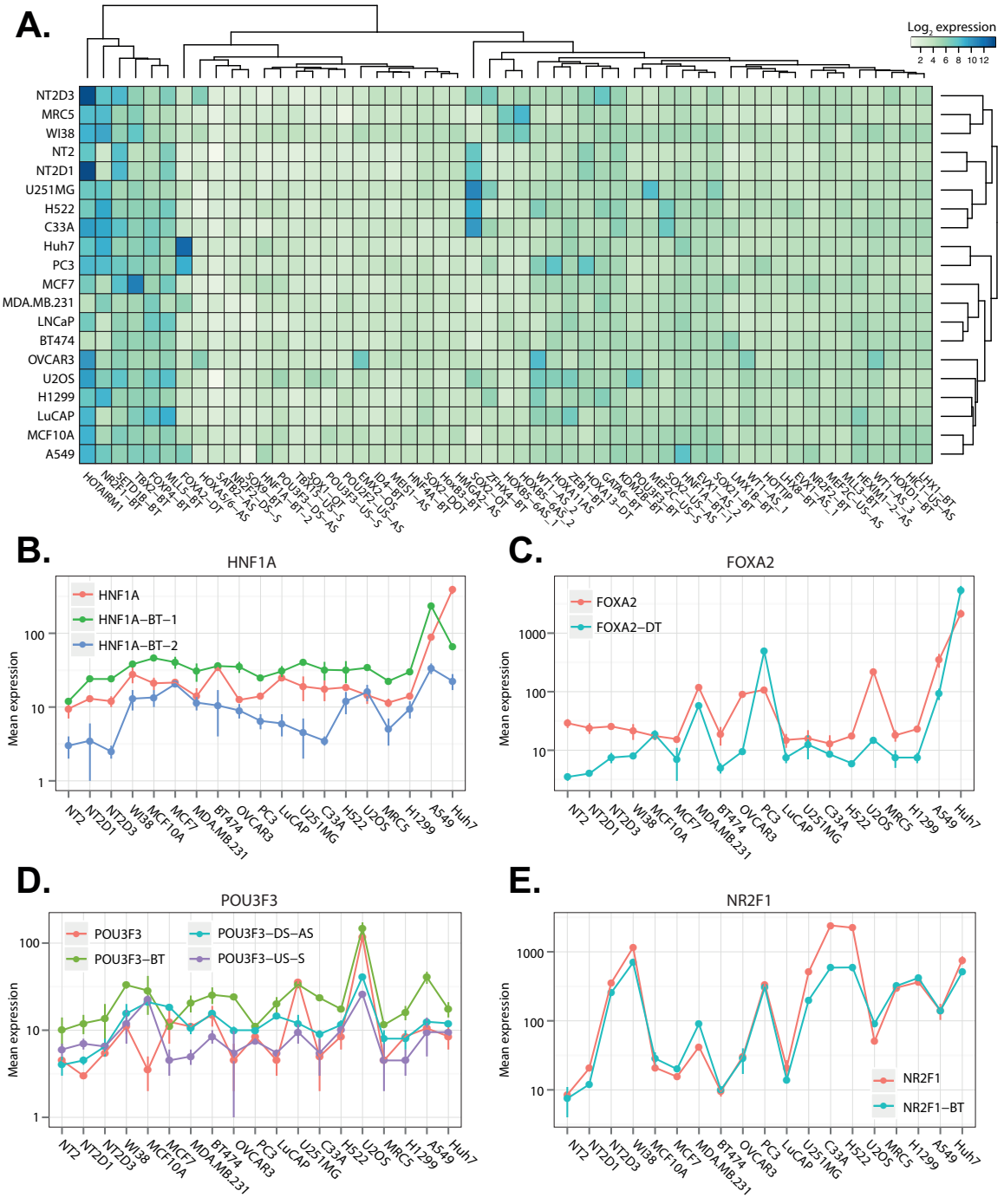


Figure 3.21 *A: Heatmap showing the NanoString® expression profiles of human pcRNAs across all the cancer cell lines included in the assay. (**B-E**) NanoString® expression profiles of human pcRNAs HNF1A (**B**), FOXA2 (**C**), POU3F3 (**D**) and NR2F1 (**E**) across all the cancer cell lines included in the assay.*

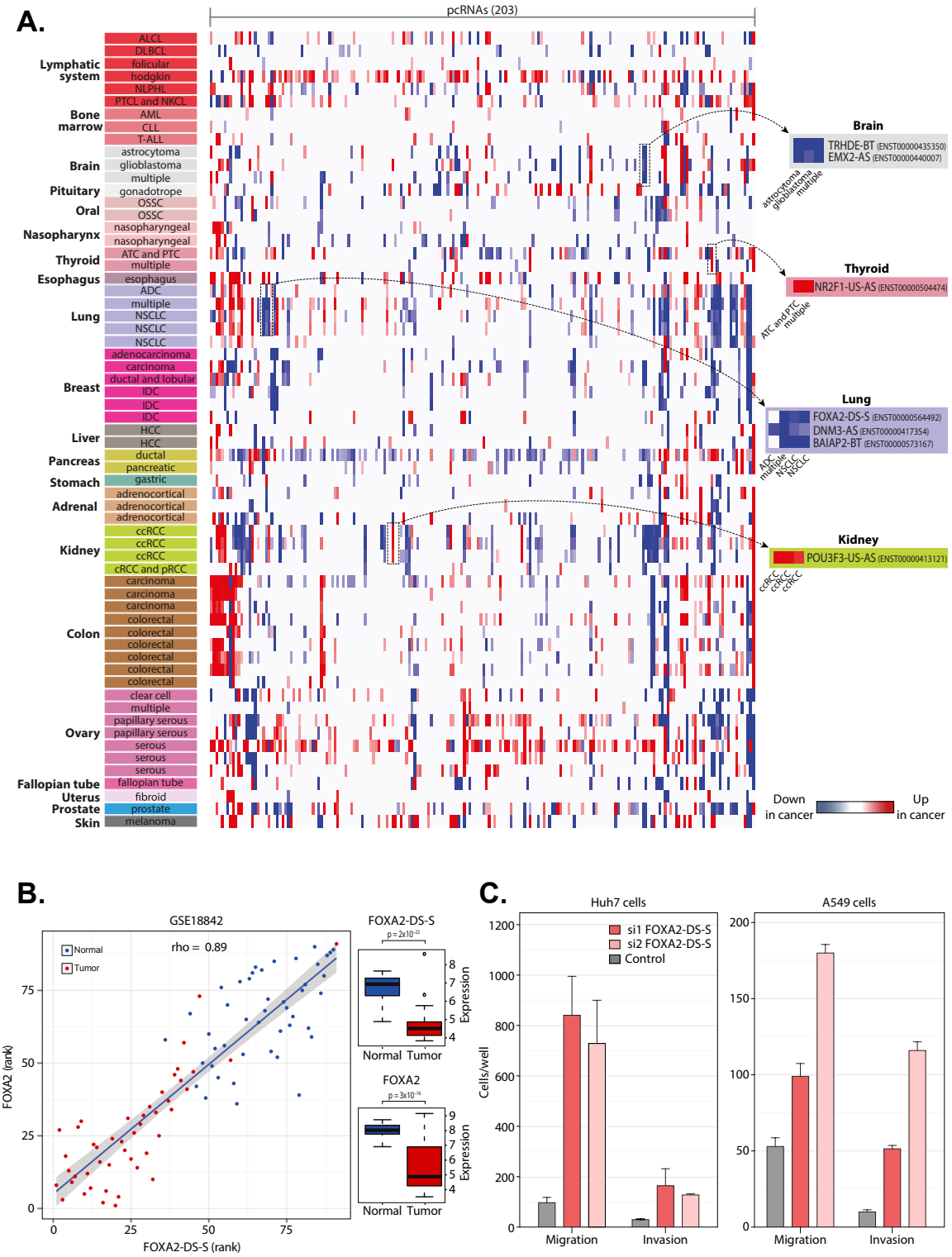


Figure 3.22 A: *Heatmap showing pcRNAs differentially expressed in cancer microarray studies. Student t-test (p-value < 0.005 and fold-change > 1.25) was used to identify pcRNAs (columns) that were up (red) or down-regulated (blue) in tumours compared to normal tissues (rows). Examples of pcRNAs associated with specific loci are shown.* ***B:*** *Spearman correlation between the expression of FOXA2 and FOXA2-DS-S in lung cancers (GSE18842 dataset). Tumour and normal individual samples are represented as blue and red dots, respectively. Boxplots on the right show that both transcripts are down-regulated in tumour compared to normal samples (Student's t-test p-values are indicated).* ***C:*** *Invasion and migration assay analysis of Huh7 cells upon knock-down of FOXA2-DS-S using two different siRNAs (si1 and si2) compared to negative control siRNA. The analysis in panels A and B have been realised by Mr Arias-Carrasco in the laboratory of Dr Maracaja-Coutinho.*

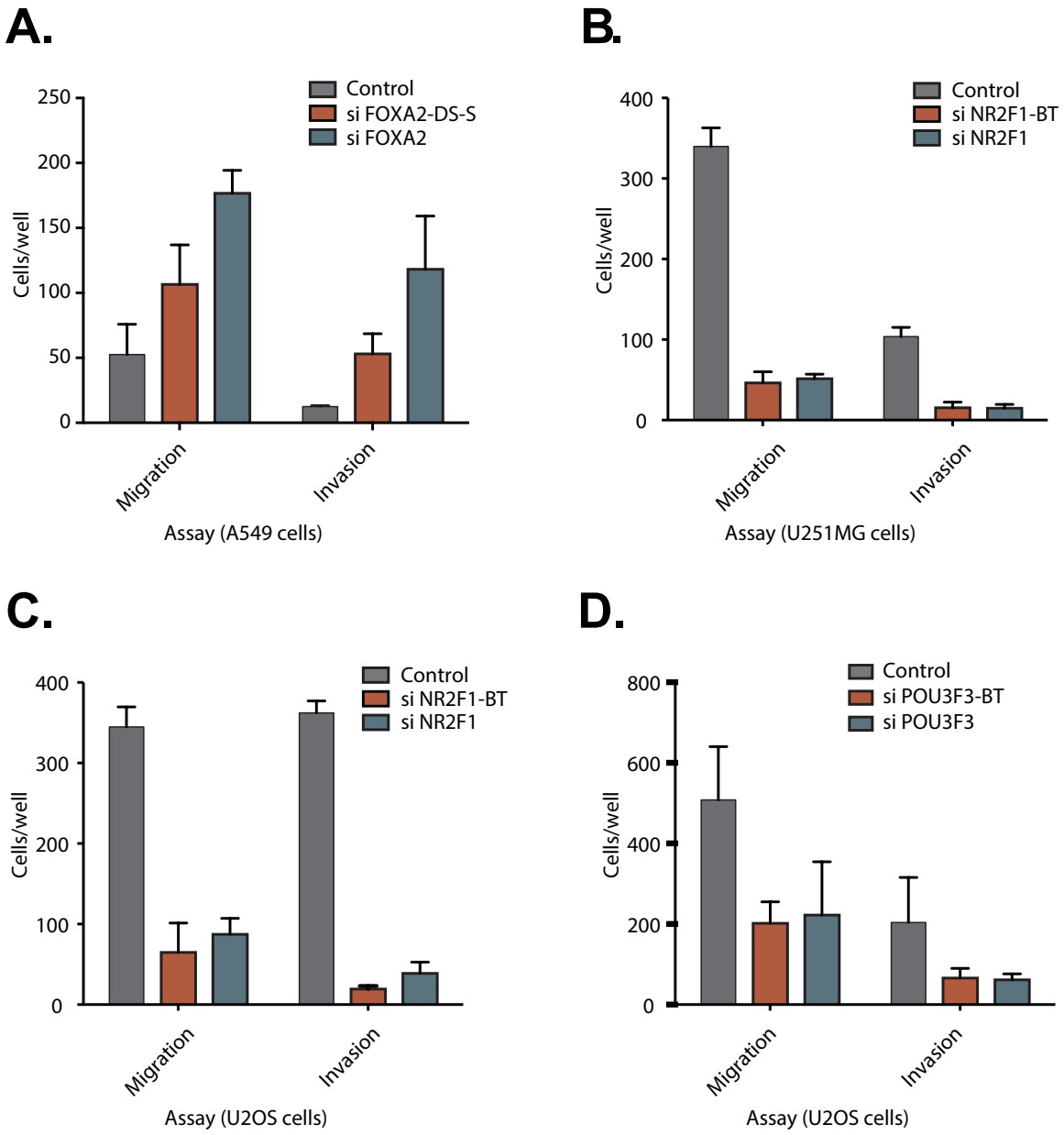


Figure 3.23 *(**A-D**) Invasion and migration assay analysis upon knock-down of FOXA2 (**A**), NR2F1 (**B,C**), POU3F3 (**D**) and their pcRNAs compared to negative control siRNA. These experiments have been realised by Dr Amaral, Dr Viré and Ms Büscher in the laboratory of Prof Kouzarides.*

moters and to overlap chromatin loop anchor points. The functional analysis of several pcRNAs revealed that they positively regulate the expression of the neighbouring protein coding genes and they have similar effects of the invasion and migration property of cancer cell lines. These results will be discussed in detail in Chapter 6. The next chapter will describe our parallel efforts to characterise the roles of extracellular vesicles in cell-to-cell communication.

Tables

Table 3.1 *RNA-Seq datasets analysed in this study. In the column "Source", 1 refers to Brawand et al. (2011) and 2 refers to ENCODE Project Consortium et al. (2012).*

Species	Type	Sex	ID	Library	Material	Notes	Source
mmu	br	F	SRR306757	single	tissue	.	1
mmu	br	M	SRR306758	single	tissue	.	1
mmu	br	M	SRR306759	single	tissue	.	1
mmu	br	M	SRR306760	single	tissue	.	1
mmu	br	M	SRR306761	single	tissue	.	1
mmu	br	M	SRR306762	single	tissue	.	1
mmu	cb	F	SRR306763	single	tissue	.	1
mmu	cb	M	SRR306764	single	tissue	.	1
mmu	cb	M	SRR306765	single	tissue	.	1
mmu	ht	F	SRR306766	single	tissue	.	1
mmu	ht	M	SRR306767	single	tissue	.	1
mmu	ht	M	SRR306768	single	tissue	.	1
mmu	kd	F	SRR306769	single	tissue	.	1
mmu	kd	M	SRR306770	single	tissue	.	1
mmu	kd	M	SRR306771	single	tissue	.	1
mmu	lv	F	SRR306772	single	tissue	.	1
mmu	lv	M	SRR306773	single	tissue	.	1
mmu	lv	M	SRR306774	single	tissue	.	1
mmu	ts	M	SRR306775	single	tissue	.	1
mmu	ts	M	SRR306776	single	tissue	.	1
hsa	br	F	SRR306838	single	tissue	.	1
hsa	br	M	SRR306839	single	tissue	.	1
hsa	br	M	SRR306840	paired	tissue	.	1
hsa	br	M	SRR306841	single	tissue	.	1
hsa	br	M	SRR306842	paired	tissue	.	1
hsa	br	M	SRR306843	single	tissue	.	1
hsa	cb	F	SRR306844	single	tissue	.	1
hsa	cb	M	SRR306845	single	tissue	.	1
hsa	cb	M	SRR306846	single	tissue	.	1
hsa	ht	F	SRR306847	single	tissue	.	1
hsa	ht	M	SRR306848	single	tissue	.	1
hsa	ht	M	SRR306849	single	tissue	.	1
hsa	ht	M	SRR306850	single	tissue	.	1
hsa	kd	F	SRR306851	single	tissue	.	1
hsa	kd	M	SRR306852	single	tissue	.	1
hsa	kd	M	SRR306853	single	tissue	.	1
hsa	lv	M	SRR306854	single	tissue	.	1
hsa	lv	M	SRR306855	single	tissue	.	1
hsa	lv	M	SRR306856	single	tissue	.	1
hsa	ts	M	SRR306857	single	tissue	.	1
hsa	ts	M	SRR306858	single	tissue	.	1
mmu	ESpA	N	SRR496249	single	ES-Bruce	cell_polyA	2
mmu	ESpA	N	SRR496250	single	ES-Bruce	cell_polyA	2
mmu	CH12	N	SRR549363	single	CH12-PSU	cell_polyA	2
mmu	CH12	N	SRR549364	single	CH12-PSU	cell_polyA	2
mmu	MEL	N	SRR496221	single	MEL-LICR	cell_polyA	2
mmu	MEL	N	SRR496222	single	MEL-LICR	cell_polyA	2
mmu	MEL	N	SRR549339	single	MEL-PSU	cell_polyA	2
mmu	MEL	N	SRR549340	single	MEL-PSU	cell_polyA	2
mmu	MEL-dmso	N	SRR549335	single	MEL-PSU	cell_polyA	2

continued …

…continued

Species	Type	Sex	ID	Library	Material	Notes	Source
mmu	MEL-dmso	N	SRR549336	single	MEL-PSU	cell_polyA	2
mmu	C2C12	N	SRR496442	paired	C2C12	.	2
mmu	C2C12_EqS	N	SRR496443	paired	C2C12_EqS	.	2
hsa	EScypA	N	SRR307919	paired	H1-hESC	cytosol_polyA	2
hsa	ESnupA	N	SRR307925	paired	H1-hESC	nucleus_polyA	2
hsa	EScepA	N	SRR307911	paired	H1-hESC	cell_polyA	2
hsa	EScepA	N	SRR307912	paired	H1-hESC	cell_polyA	2
hsa	Hsmm	N	SRR521516	paired	Hsmm	.	2
hsa	Hsmm	N	SRR521517	paired	Hsmm	.	2
hsa	Hsmm	N	SRR521518	paired	Hsmm	.	2
hsa	Hsmm	N	SRR521519	paired	Hsmm	.	2
hsa	GM12878	N	SRR521447	paired	GM12878_R1	.	2
hsa	GM12878	N	SRR521448	paired	GM12878_R1	.	2
hsa	GM12878	N	SRR521449	paired	GM12878_R1	.	2
hsa	GM12878	N	SRR521450	paired	GM12878_R1	.	2
hsa	GM12878	N	SRR521451	paired	GM12878_R2	.	2
hsa	GM12878	N	SRR521452	paired	GM12878_R2	.	2
hsa	GM12878	N	SRR521453	paired	GM12878_R2	.	2
hsa	GM12878	N	SRR521454	paired	GM12878_R2	.	2
hsa	GM12878	N	SRR521455	paired	GM12878_R2	.	2
hsa	GM12878	N	SRR521456	paired	GM12878_R2	.	2
hsa	K562	N	SRR521457	paired	K562_R1	.	2
hsa	K562	N	SRR521458	paired	K562_R1	.	2
hsa	K562	N	SRR521459	paired	K562_R1	.	2
hsa	K562	N	SRR521460	paired	K562_R1	.	2
hsa	K562	N	SRR521461	paired	K562_R1	.	2
hsa	K562	N	SRR521462	paired	K562_R2	.	2
hsa	K562	N	SRR521463	paired	K562_R2	.	2
hsa	K562	N	SRR521464	paired	K562_R2	.	2
hsa	K562	N	SRR521465	paired	K562_R2	.	2

GO ID	GO Term	Annotated Genes	Genes with pcRNA	Expected	Adjusted p-value
GO:0045944	Positive regulation of transcription fro...	848	90	31.42	1.2×10^{-15}
GO:0000122	Regative regulation of transcription fro...	607	62	22.49	8.5×10^{-10}
GO:0006355	Regulation of transcription, DNA-templat...	2991	206	110.81	6×10^{-4}
GO:0048557	Embryonic digestive tract morphogenesis	21	8	0.78	6.1×10^{-4}
GO:0001709	Cell fate determination	43	10	1.59	2.6×10^{-3}
GO:0060021	Palate development	86	14	3.19	2.6×10^{-3}
GO:0031128	Developmental induction	34	8	1.26	2.6×10^{-3}
GO:0003151	Outflow tract morphogenesis	59	12	2.19	3×10^{-3}
GO:0048844	Artery morphogenesis	50	13	1.85	3.1×10^{-3}
GO:0048701	Embryonic cranial skeleton morphogenesis	42	11	1.56	8.5×10^{-3}
GO:0045665	Negative regulation of neuron differenti...	55	10	2.04	1.4×10^{-2}
GO:0035019	Somatic stem cell maintenance	48	9	1.78	2.5×10^{-2}
GO:0051145	Smooth muscle cell differentiation	48	12	1.78	2.5×10^{-2}
GO:0061140	Lung secretory cell differentiation	13	5	0.48	2.5×10^{-2}
GO:0048704	Embryonic skeletal system morphogenesis	93	20	3.45	2.5×10^{-2}
GO:0060425	Lung morphogenesis	52	12	1.93	2.7×10^{-2}
GO:0021536	Diencephalon development	119	15	4.41	3.4×10^{-2}
GO:0010470	Regulation of gastrulation	31	7	1.15	3.4×10^{-2}
GO:0001701	In utero embryonic development	357	27	13.23	3.4×10^{-2}
GO:0007420	Brain development	621	67	23.01	3.7×10^{-2}

Table 3.2 *GO enrichment of pcRNA-associated protein coding genes*

4

Extracellular vesicles from Neural Stem Cells transfer IFN-ɣ via Ifngr1 to activate Stat1 signalling in target cells

The work reported in this chapter was published in the journal "Molecular Cell" in 2014 (Cossetti et al., 2014a) and is the result of a collaboration between the Pluchino and Enright laboratories. All the experimental work apart for the electron microscopy and mass spectrometry has been performed by Dr Cossetti, Dr Iraci and co-authors in the laboratory of Dr Pluchino. All the bioinformatics analysis have been done in the laboratory of Dr Enright by Dr Saini, Dr Davis and myself. I have driven the design and analysis of the RNA-Seq experiments, the gene ontology analysis and the analysis of microarray and SILAC data in NIH 3T3 cells, which led to the identification of Stat1 as the key pathway activated in Th1 NPCs and transferred to recipient cells via EVs. I have also contributed to the design, data analysis and data interpretation of the experimental work presented in this chapter.

4.1 Introduction

The discovery of neurogenesis in the adult mammalian brain dates back to the early 1960s (Altman, 1963), when it was shown that in the rodent Central Nervous System (CNS) there is constitutive production of neurons in the hippocampus and in the olfactory bulb. However, the neurogenic potential of the adult brain was not widely accepted until the 1990s, when the scientific community started to recognize that neurogenesis persists in many areas of the postnatal brain (reviewed in Ming and Song, 2005). Subsequently, several studies have established that the adult mammalian brain, under physiological conditions, has at least two specific areas of active neurogenesis, which have been defined neural stem cell niches: the subventricular zone (SVZ) of the lateral ventricle and the subgranular zone of the dentate gyrus in the hippocampus (Gage, 2000; Palmer, 1997; Doetsch et al., 1999).

The discovery of adult neural stem cells led to the speculation that the adult CNS could have an intrinsic potential to repair itself after injury. However, in the last 20 years, several works have shown that the endogenous stem cell compartment in most cases fails to repair the damage and undergo regeneration (Ekdahl et al., 2011; Monje et al., 2003; Butovsky et al., 2006; Rolls et al., 2007; Pluchino et al., 2008). This observation, together with recent advances

in stem cell biology, have prompted the idea that CNS diseases and/or injuries could be treated with the local or systemic delivery of stem cells, in the hope that the injected cells would migrate toward the lesioned tissue, proliferate and differentiate, leading to a restoration of the function of the damaged areas.

Indeed, several works have shown that the local injection of Neural Progenitor Cells (NPCs) promotes a functional recovery in various CNS disease models, such as those of stroke, multiple sclerosis and traumatic brain injury (Martino et al., 2011). However, in contrast with the initial expectations, it was also shown that injected NPCs usually fail to proliferate, differentiate and induce repair of the damaged tissue (Cao et al., 2002; Jeong et al., 2003; Lu et al., 2003; Chu et al., 2004; Fujiwara et al., 2004; Pluchino et al., 2005); following these observations, several works have now clearly demonstrated that the therapeutic mechanism of transplanted somatic stem cells does not depend solely on cell replacement, but rather it is mainly due to their capacity to engage a complex mechanism of cell-to-cell communication with the host immune system that mediates neuroprotection and immunomodulation (Pluchino et al., 2003; Pluchino et al., 2005; Mueller et al., 2006; Rampon et al., 2008; Bacigaluppi et al., 2008; Pluchino et al., 2009; Martino et al., 2011). Among the possible routes of cell-to-cell communication, Extracellular Vesicles (EVs) offer an interesting possibility, because they have both the capacity to reach compartments distant from their place of production as well as the capacity to transport a broad range of molecules, such as metabolites, lipids, mRNAs and miRNAs.

In this work we investigated the properties of NPC-derived EVs, characterised their content and assessed their capacity to respond to perturbations of the microenvironment. We found that a Th1-like proinflammatory environment deeply modifies the transcriptome and proteome of NPC-derived EVs and exosomes, and induces the secretion of mRNA and protein components of the IFN-ɣ pathway. Using an *in vitro* model of target cells we found that EVs are able to transfer IFN-ɣ via the IFN-ɣ/Ifgnr1 complex, which induces the activation of the IFN-ɣ pathway in the target cells.

This work describes a novel mechanism of EV-mediated cell-to-cell communication which allows NPCs to propagate the activation of a signalling pathway at a distance. This process might constitute one of the strategies used by NPCs to communicate with the immune system *in vivo*, providing a potential molecular explanation for their therapeutic immunomodulatory effects.

4.2 Characterisation of NPC derived EVs

We first sought to provide a physical characterisation of the vesicles secreted by NPCs under normal culture conditions. Primary cultures of NPCs were es-

tablished from the Subventricular Zone (SVZ) of adult SJL mice and expanded as neurospheres in chemically defined serum-free media as previously described (Pluchino et al., 2005). Scanning Electron Microscopy (SEM) of the NPC surface showed numerous membranous structures of small and medium size, which were compatible with previous descriptions of nanotubes and membrane vesicles (**Figure 4.1A,B**). Similarly, Transmission Electron Microscopy (TEM) imaging of NPCs showed the presence of numerous electron-dense vesicles in contact with the NPC plasma membrane, which suggested that NPCs were actively shedding EVs from the plasma membrane (**Figure 4.1C**).

We then collected EVs from the supernatant by differential centrifugation (Théry et al., 2001) and characterised their physical and biochemical properties by electron microscopy, Dynamic Light Scattering (DLS), Nanoparticle Tracking Analysis (NTA) and Western blot. Transmission electron microscopy showed that the purified EV population consists of vesicles of heterogeneous size (**Figure 4.1D**), with a sub-population of cup-shaped electron-dense vesicles in the size range 40 nm–120 nm (**Figure 4.1E**). To resolve the heterogeneous vesicle population into distinct subpopulations, we applied DLS on EVs, since it is the preferred method to routinely determine the size of nanoparticles (Bootz et al., 2004). The size distribution analysis of EVs by DLS identified three distinct particle size classes (SC): an SC1 with a peak size diameter of 136.7 nm ($\pm$31.36 nm) consistent with exosomes (Théry et al., 2001); a SC2 with a peak size diameter of 667.2 nm ($\pm$100.22 nm), similar in size to a previously described class of shedding membrane particles (Heijnen et al., 1999), and an SC3 minor population with a peak size diameter of 5087 nm ($\pm$24.74 nm), likely reflecting aggregated particles (**Figure 4.1F**). However, due to the intrinsic limitations of DLS in determining the precise sizes of polydispersed samples, we also decided to employ NTA (Dragovic et al., 2011; Filipe et al., 2010). NTA further supported the presence of a multimodal size distribution with a major peak corresponding to smaller particles (mean diameter 167$\pm$2.82 nm) and a significantly less represented peak of larger particles (mean diameter 342.4$\pm$36.65 nm, **Figure 4.1G**, red curve).

Given the heterogeneity of the vesicles purified in the EV preparation, we next aimed to obtain a refined vesicle preparation enriched in exosomes. To this end we applied a sucrose gradient fractionation protocol on the EV samples as previously described (Théry et al., 1999) and we pooled together the fractions ranging from a density of 1.13 g/mL to 1.20 g/mL, which are generally accepted as exosomes (Bobrie et al., 2011); such pooled fractions will be referred to as EXOs hereafter. NTA of EXOs showed an enrichment in particles

with a size range of 100 nm to 150 nm (**Figure 4.1G**, black curve), which is in line with previous descriptions of exosomes (Théry et al., 2001).

To further characterise the EXO fractions we performed Western blots using antibodies directed against *bona fide* exosomal markers (**Figure 4.1H**). We observed that EXOs were enriched in the endosomal sorting complex protein AIP-1/ALIX and the tumour susceptibility gene 101 (TSG 101) (Théry et al., 2001) compared to EVs. EXO preparations also showed an enrichment for the heat shock proteins HSP70 and HSP90 (Théry et al., 2001), as well as for the exosome-associated protein argonaute 1 (AGO1), but not for the non-vesicle-associated AGO2 (Arroyo et al., 2011). Additionally, we also observed an increase of the two tetraspanins CD9 and CD63, as well as the exosome-like vesicle marker tumour necrosis factor 1 receptor 1 (TNFR1) (Hawari et al., 2004) in both EVs and EXOs as compared to NPCs. Taken together, these data indicate that we have obtained a purified EXO preparation enriched in exosomal markers.

4.3 Modulation of EV and EXO cargo by cytokine signalling

We next aimed to assess the role of NPC-derived EVs and EXOs in the intercellular spreading of inflammatory signals. To this purpose we cultured NPCs in the presence cytokines mixes that mimic *in vitro* a Th1-like pro inflammatory microenvironment or a Th2-like anti inflammatory microenvironment (Pluchino et al., 2008). The Th1 pro inflammatory mix (hereby referred to as Th1 condition) is composed of interferon gamma (IFN-ɣ), tumour necrosis factor alpha (TNFα) and interleukin 1 beta (IL1β), while the Th2 anti inflammatory mix (hereby referred to as Th2 condition) is composed of interleukin 4 (IL4), interleukin 5 (IL5) and interleukin 13 (IL13). After culturing NPCs in the presence of Th1 and Th2 cytokines we purified EVs and EXOs and applied again the NTA analysis verifying that there is no significant difference in the size of vesicles after cytokine conditioning (**Figure 4.2A and B**, respectively). Moreover, we also confirmed that the total RNA and protein content of EVs and EXOs, is not altered in response to cytokine treatment (**Figure 4.2C,D**).

In order to determine the role of NPC-derived EVs as conveyors of immune signals and to assess whether this process is modulated by the microenvironment, we employed RNA-Seq (see Methods, section 11.1) to characterise the transcriptome of NPC and NPC-derived EVs and EXOs in basal, Th1 and Th2 conditions.

We found that Th1 cytokines have a broad impact on the NPC transcriptome; in fact, 686 genes were found up-regulated and 477 down-regulated with a FC > 5 in Th1 vs. Basal NPCs (**Figure 4.3A**). When we performed a Gene

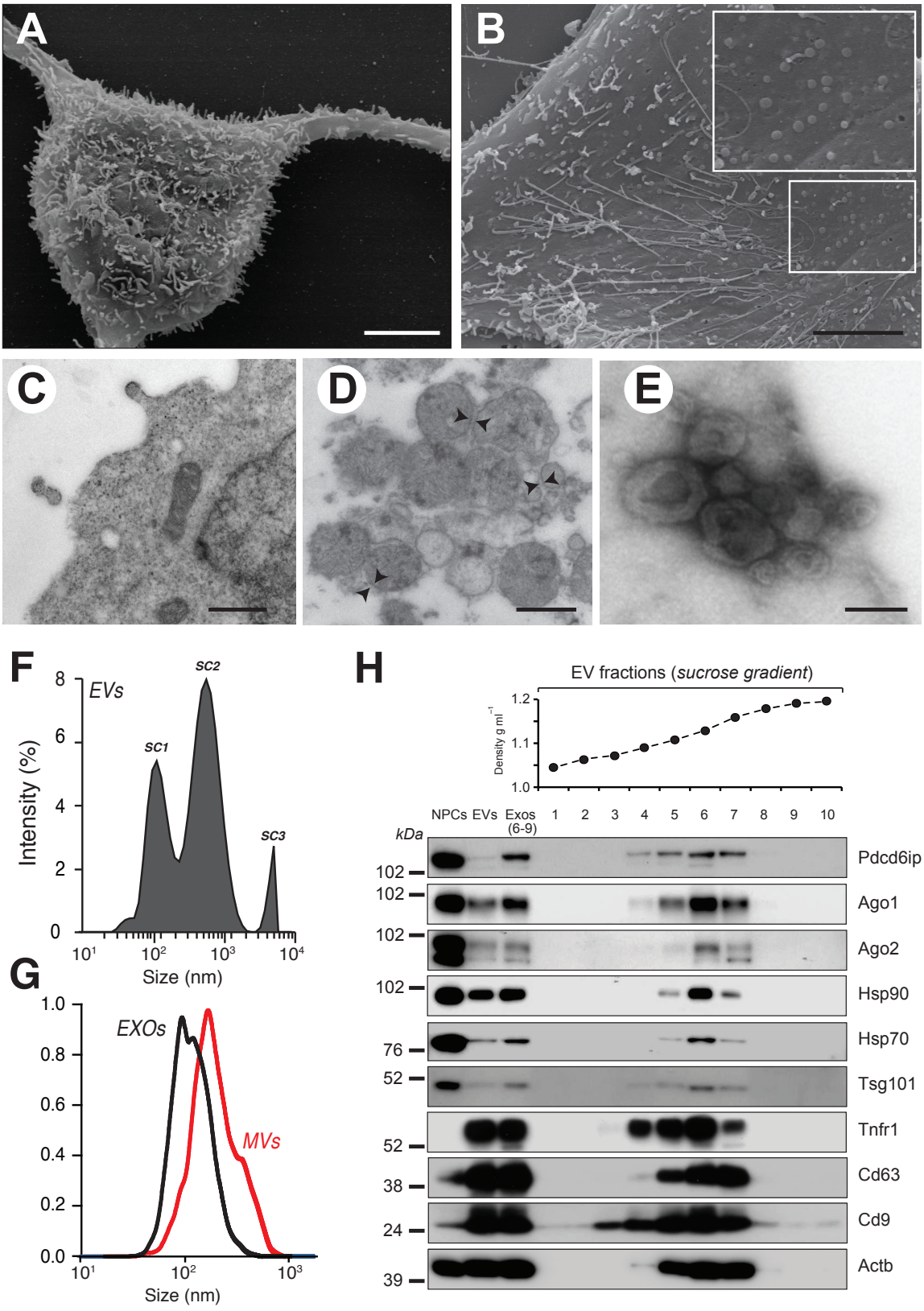


Figure 4.1 *legend on next page*

Figure 4.1 *(previous page)* ***A:*** *SEM photograph of a NPC with long adherent expansions and membrane protrusions on the surface. Scale bar 5 µm.* ***B:*** *SEM of the NPC surface showing nanotubes and round membrane vesicles. The inset shows a magnified detail. Scale bar 5 µm.* ***C:*** *TEM photograph showing two membrane vesicles being released from the plasma membrane of an NPC. Scale bar 500 nm.* ***D:*** *TEM photograph of NPC-derived EVs, showing an heterogeneous population of vesicles of different diameters (range 4 nm to 200 nm) surrounded by a double-layer membrane (arrowheads). Scale bars 500 nm.* ***E:*** *TEM photograph negatively stained EVs. Cup-shaped vesicles of 8 nm to 120 nm can be observed. Scale bar 100 nm.* ***F:*** *EV size distribution by intensity, determined by DLS. The relative intensity of light scattered by EVs in various size classes (SC) are plotted. Data are mean size (nm) ± standard deviation from a total of 12 independent determinations.* ***G:*** *Distribution of the particle sizes in the EV preparation (red) and EXO preparation (black) determined by NTA. The data show the mean from 5 independent experiment and are normalized to 1 for size comparison.* ***H:*** *Western blot analysis of exosomal markers in NPCs, EVs, and EXOs. The EXO preparation results from pooling the fractions 6-9 of the sucrose gradient, corresponding to a density range of 1.13 g mL^{-1} to 1.20 g mL^{-1}. This figure is representative of 3 independent experiments. Data published in Cossetti et al. (2014a).*

Ontology (GO) enrichment analysis on the differentially expressed genes we found that the most altered pathways in Th1 NPCs were *Response to IFN-ɣ* (p-value = 3.8×10^{-22}) and *Antigen processing and presentation* (p-value = 2.1×10^{-17}; **Figure 4.4A**). In particular, we observed a remarkable upregulation of genes belonging the IFN-ɣ signalling pathway, such as signal transducer and activator of transcription 1 (*Stat1*, FC = 37.7 in Th1 NPCs vs. Basal NPCs), *Stat2* (FC = 14.8), interferon regulatory transcription factor 1 (*Irf1*, FC = 47.7), *Irf2* (FC = 3.6) and guanylate-binding protein 9 (*Gbp9*, FC = 26.2) (**Figure 4.4B**). Moreover, we found that these effects were specific for the Th1 cytokines, since Th2 cytokines only induced the differential expression of a small number of genes (**Figure 4.3A**) without any enrichment for specific GO categories. Additionally, we also investigated whether Th1 cytokines also regulate mRNA abundance within EVs and EXOs, and we found that in both EVs and EXOs the majority of genes belonging to the IFN-ɣ signalling pathway were enriched in response to Th1 cytokines (**Figure 4.4C,D**), while they were not in response to Th2 cytokines (**Figure 4.3B,C** and **Figure 4.4C,D**).

We next asked whether the secretion of mRNA components of the IFN-ɣ signalling pathway could be paralleled by the secretion of their protein counterparts. We therefore extracted total proteins from basal, Th1 and Th2 EVs and EXOs as previously described (Witwer et al., 2013) and measured the abundance of selected components of the IFN-ɣ pathway by Western blot analysis. These data show a Th1-specific increase of total STAT1 in NPCs, which was paralleled in Th1 EVs and – to a lesser extent – Th1 EXOs (**Figure 4.4E**). The full activation of STAT1 requires two phosphorylations, on Y701 and S727 (Wen et al., 1995); we found that both forms were upregulated in Th1 NPCs, while the former was also upregulated in Th1 EVs and EXOs (**Figure 4.4E**).

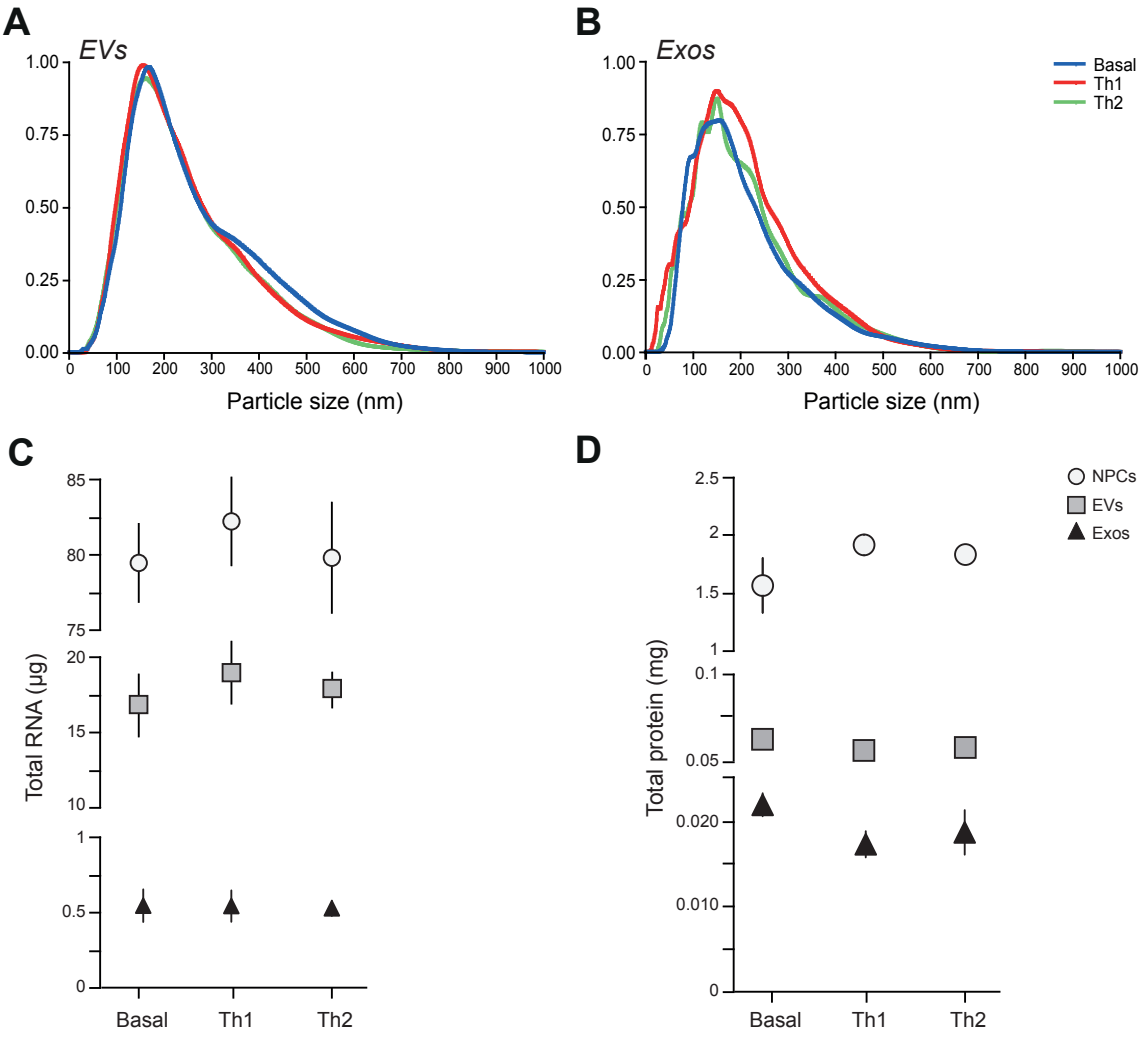


Figure 4.2 A,B: *Plot showing the results of NTA analysis on EVs (A) and EXOs (B) purified from NPCs grown in Basal (blue), Th1 (red) or Th2 (green) condition. Data are normalized means from 4 independent experiments. Normalization of the data was made by dividing the concentration value at every particle size in the distribution by the largest concentration value within the distribution.* ***C:*** *Quantification of total RNA purified from 12 × 10^6 NPCs, EVs or EXOs, in basal, Th1 or Th2. The data represent the mean (±SEM) of 3 independent experiments.* ***D:*** *Quantification of total proteins purified from 12 × 10^6 NPCs EVs or EXOs, in basal, Th1 or Th2. The data represent the mean (±SEM) of 3 independent experiments. Data published in Cossetti et al. (2014a).*

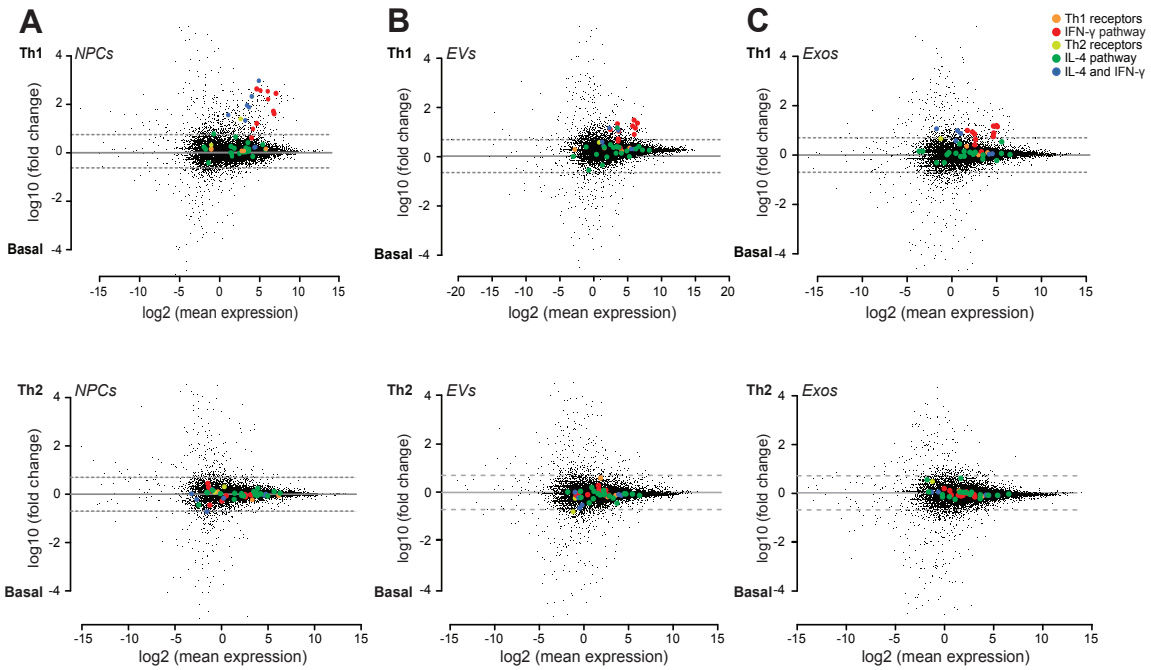


Figure 4.3 A-C: *MA plot of Th1 (top) and Th2 (bottom) NPCs (A), EVs (B) and EXOs (C) vs Basal. Genes belonging to manually curated gene sets were highlighted: Th1 receptors (IFNgR1, IFNgR2, TNFRsf1a, IL-1R1); IFN- γ pathway (Stat1, Stat2, Irf1, Irf2, Irf9, Ifi44, Ifi47, Ifit1, Ifit3, Gbp1, Isg15); Th2 receptors (IL-4Ra, IL-13Wa1); IL-4 and IFN- γ pathway (Socs1, Jak1, Jak2, H2-Ab1, H2-Aa, H2-Eb1, Cd74, Ciita); IL-4 pathway (Src, Tyk2, Shc1, Irs1, Irs2, Irs4, Inpp5d, Grb2, Ecm1, Sos1, Sos2, Pik3ca, Pdk1, Rps6kb1, Akt1, Bad, Jak3, Il4i1, Nfil3, Maf, Bhlhe41). Data published in Cossetti et al. (2014a).*

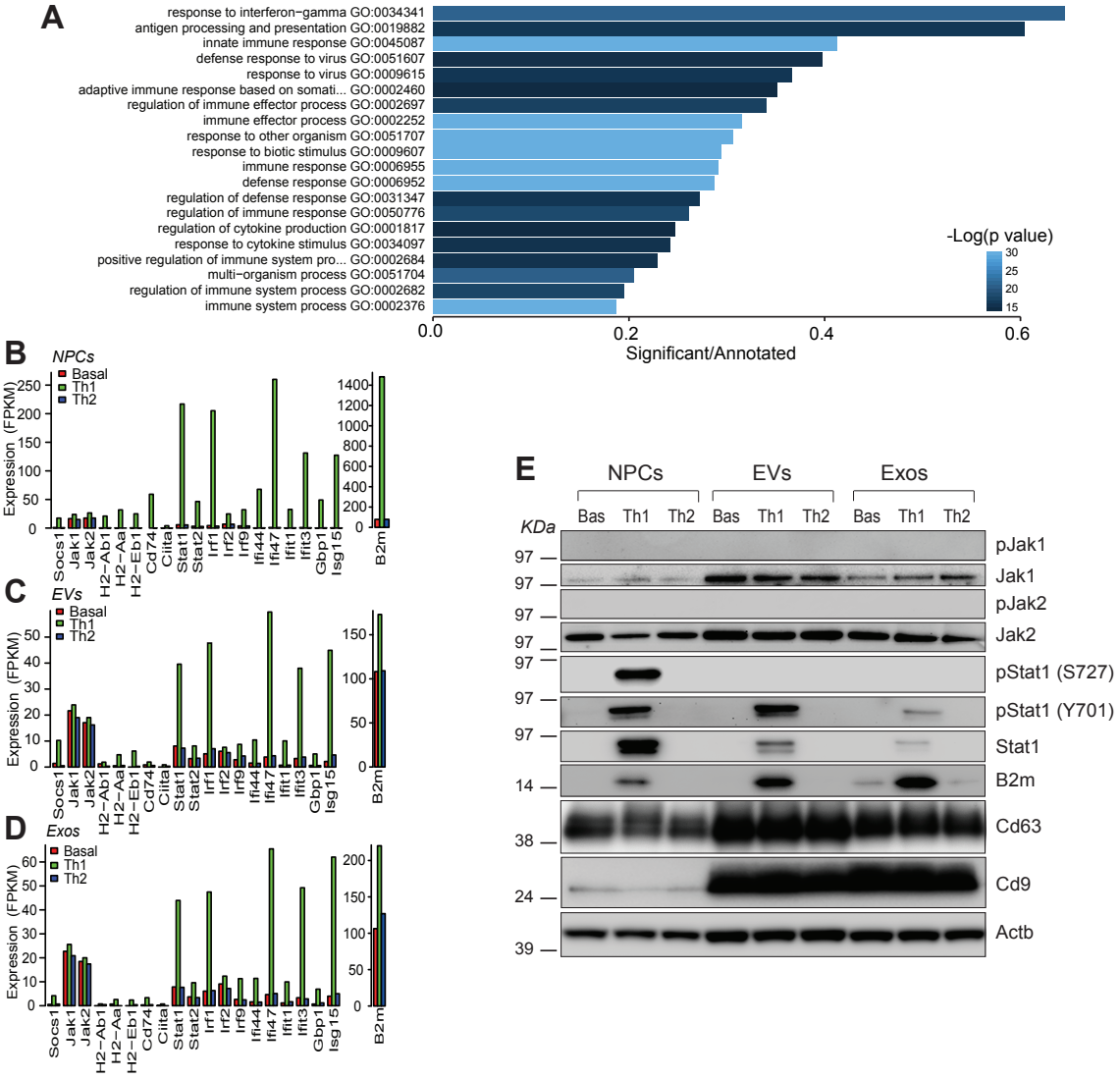


Figure 4.4 A: *GO enrichment analysis of genes upregulated in Th1 NPCs vs basal. The x-axis shows the fraction of differentially expressed genes vs the total number of genes in each GO category. The colour scale represents the negative log_{10} of the adjusted p-value.* ***B-D:*** *Histogram plots showing the expression (FPKM) of genes of the IFN-ɣpathway in Basal (red), Th1 (green) and Th2 (blue) NPCs (B), EVs (C) and EXO (D). The y-axis indicates FPKMs for each gene.* ***E:*** *Western blot of components of the Jak/Stat signalling pathways in NPCs, EVs and EXOs in basal, Th1 and Th2 conditions. Panel is representativeaof 4 independent experiments. Data published in Cossetti et al. (2014a).*

Additionally, we also found a Th1-specific upregulation in NPCs, EXOs and EVs of beta-2-microglobulin (B2M) (**Figure 4.4E**), a downstream component of the IFN-ɣ signalling pathway (Fellous et al., 1981; Nachbaur et al., 1988; Wong et al., 1984). On the other hand, we did not detect in any sample the upregulation of janus kinase 1 and 2 (JAK1/2) nor of their phosphorylated forms (**Figure 4.4E**). Similarly, we did not detect in any sample the upregulation of STAT6 or its phosphorylated form. This is in line with the lack of expression detected by long RNA-Seq for this transcription factor and for the Th2 cytokine receptors. Also, this finding supports the specificity of STAT1 activation for the Th1 cytokines. Finally we observed that the exosomal markers CD63 and CD9 were enriched in EVs and EXOs compared to NPCs, but their abundance was not altered in response to Th1 or Th2 treatment (**Figure 4.4E**).

Taken together, these results suggest that stimulation of NPCs with Th1 cytokines induces the activation of IFN-ɣ responsive pathways and promotes the packaging of activated components of the pathway (i.e. phosphorylated STAT1) in EVs and EXOs. These observations prompt the hypothesis that such activated members of the IFN-ɣ pathway might be transferred to other cells. Therefore, we next sought to determine if EVs and EXOs secreted by Th1 NPCs have any biologically relevant effect on recipient cells.

4.4 Th1 EVs mediate the activation of the Stat1 pathway in target cells

To assess whether NPC-derived EVs had an impact on target cells we used the NIH 3T3 cell line as a model of vesicle-recipient cells. To model *in vitro* the dynamics of EV internalisation by target cells we collected EVs from NPC lines transfected with either a farnesylated Enhanced Green Fluorescent Protein (fEGFP) or the CD63 protein fused with Red Fluorescent Protein (RFP). The labelled EVs were then added to the culture media of NIH 3T3 cells and the internalisation was assessed by confocal microscopy and Stimulated Emission Depletion (STED) microscopy. We could detect internalisation of the labelled EVs as early as 2 hours after incubation (**Figure 4.5A**), which reached its peak around 9 hours after incubation (**Figure 4.5B**)

We then profiled the transcriptome and proteome of NIH 3T3 cells exposed to EVs using NCode™ Mouse RNA microarrays and tandem mass spectrometry (MS/MS) combined with Stable Isotope Labelling of Aminoacids in Cell culture (SILAC) (**Figure 4.6A**). When comparing NIH 3T3 cells exposed to Basal EVs to unexposed NIH 3T3 cells, we detected 408 genes and 96 proteins whose expression underwent significant changes ($B \geq 3$ and p-value ≤ 0.01 for genes and proteins respectively, **Figure 4.6B,C**).

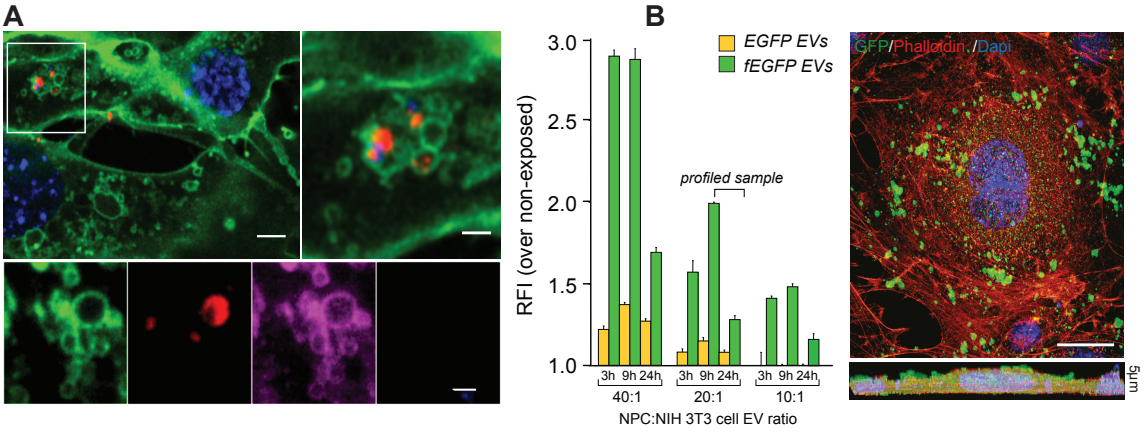


Figure 4.5 A: *Uptake of CD63-RFP EVs in fEGFP NIH 3T3 cells at 2 h after EV transfer. EVs are in red in confocal microscopy and magenta in STED microscopy. Scale bars: top left 5 µm; top right 2 µm; bottom 1 µm.* ***B:*** *Flow cytometry time course analysis of the internalization of fEGFP-labelled EVs by target cells. Data are represented as mean relative fluorescence intensity ± SEM from 3 independent experiments. Right: Representative confocal microscopy photograph of an NIH 3T3 cell (red) exposed to fEGFP EVs (green) is shown. The lower panel is a Z stack of 5 slices taken at a distance of 1 µm. Scale bar 10 µm. Data published in Cossetti et al. (2014a).*

We next considered whether EVs secreted in response to cytokines also induce specific responses in target cells. The exposure of NIH 3T3 cells to Th1 EVs led to the differential expression of 443 genes and 130 proteins ($B \geq 3$ and p-value ≤ 0.01 for genes and proteins respectively), compared to untreated NIH 3T3 cells (**Figure 4.6B,C**). Furthermore, when we compared the impact of Th1 EVs vs. Basal EVs on NIH 3T3 cells, we identified 24 genes and 95 proteins whose expression was specifically regulated by Th1 EVs only. On the other hand, the exposure of NIH 3T3 cells to Th2 EVs led to significant changes in 554 genes and 97 proteins, ($B \geq 3$ and p-value ≤ 0.01 for genes and proteins respectively). However, these changes were largely similar to those elicited in NIH 3T3 cells after the exposure to Basal EVs (no genes with $B \geq 3$ in NIH 3T3 cells exposed to Th2 EVs compared to NIH 3T3 exposed to Basal EVs).

To better resolve the functional trends elicited by EVs we used the tool GeneMANIA (Mostafavi et al., 2008) to obtain an integrated pathway analysis combining microarray and SILAC data. We first focused on the set of genes and proteins differentially expressed in NIH 3T3 cells exposed to Basal EVs, and we found a significant enrichment for the GO categories *Nuclear Chromosome* (p-value = 9.76×10^{-6}), *DNA Replication* (p-value = 2.47×10^{-4}) and *Fibrillar collagen* (p-value = 2.47×10^{-4}). We also noticed that several of the genes and proteins regulated by Basal EVs displayed the same variation pattern also in response to Th1 or Th2 EVs, regardless of the type of conditioning imposed on donor NPCs. When we repeated the GeneMANIA analysis on this set of genes and proteins common between Basal, Th1 and Th2 we found that the most

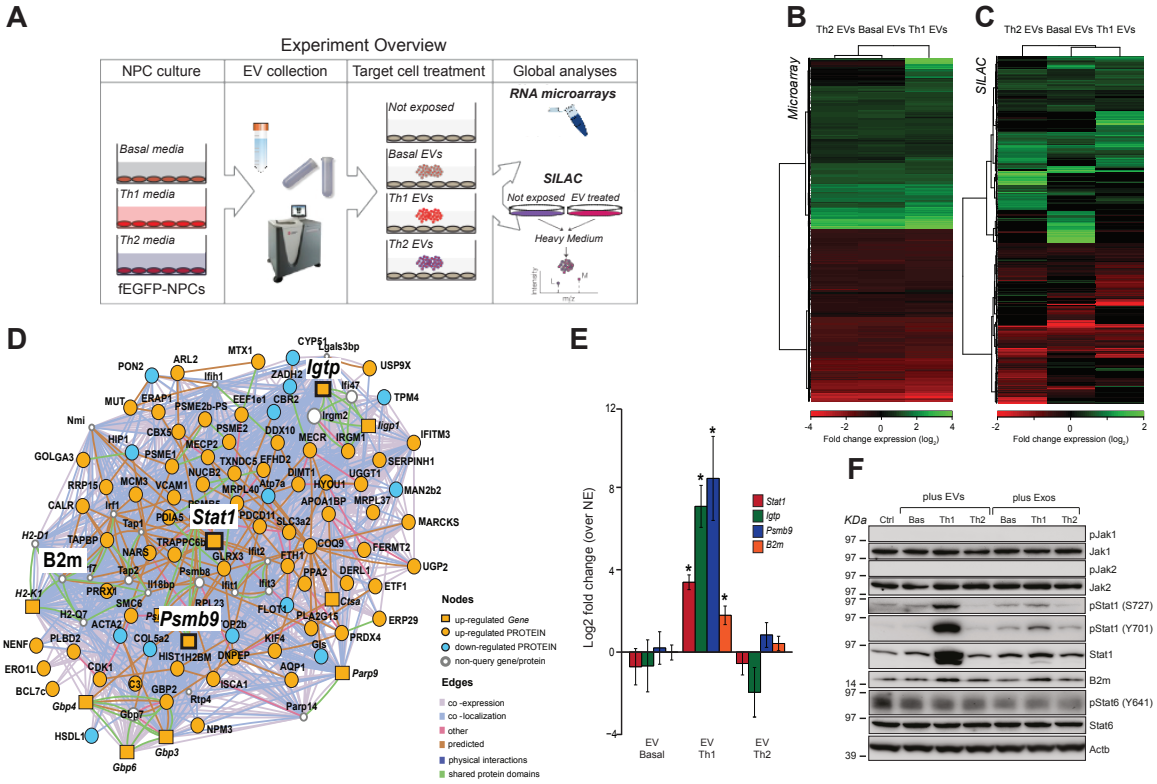


Figure 4.6 ***A:*** *Diagram representing the experimental design of EV transfer to target cells.* ***B:*** *Heatmap of the data obtained with NCode™ Mouse RNA microarrays. The color-scale represents the gene fold-changes in NIH 3T3 cells exposed to basal, Th1 or Th2 EVs for 24h relative to NIH 3T3 cells not exposed to EVs.* ***C:*** *Heatmap of the SILAC data in target cells exposed to EVs as in (B).* ***D:*** *GeneMANIA network of genes and proteins differentially expressed in NIH 3T3 cells exposed to Th1 EVs only.* ***E:*** *Bar chart showing qRT-PCR data for Stat1, Igtp, Psmb9, and B2M expression in target cells exposed to EVs as in (B). Data are represented as mean log_2fold change (± SEM, 3 independent experiments) over target cells not exposed to EVs. *p-value < 0.05, compared to not exposed.* ***F:*** *Western blot of components of the Jak/Stat pathway in target cells treated with EVs as in (B). This Western blot is representative of 4 independent experiments. Data published in Cossetti et al. (2014a).*

significantly enriched GO term is Extracellular Matrix, followed by Collagen and Chemokine Receptor Binding. We then focused on genes and proteins specifically regulated by Th1 or Th2 EVs only. When we analysed the Th1-specific GeneMANIA network (**Figure 4.6D**) we found that the most significantly enriched GO term was *Antigen processing and presentation* (p-value = 3.19×10^{-14}) followed by *response to interferon beta* (p-value = 2.17×10^{-9}), *response to interferon gamma* (p-value = 5.26×10^{-7}), and *response to cytokine stimulus* (p-value = 1.57×10^{-7}).

On the other hand, we did not find any significant GO enrichment in the Th2-specific network, suggesting that Basal and Th2 EVs elicit broadly similar effects on recipient NIH 3T3 cells.

The activation of an Interferon response in NIH 3T3 cells exposed to EVs derived from Th1 NPCs suggests that specific components exclusively present in Th1 EVs might be functionally active and responsible for inducing an IFN-like intracellular response in recipient cells. To further investigate this point, we used qRT-PCR to verify the induction of *Stat1* and other components of the IFN-γ pathway in NIH 3T3 cells exposed to Th1 EVs. These experiments confirmed that Th1 EVs specifically induce the upregulation of *Stat1, Igpt, Psmb9* and *B2M* mRNA (**Figure 4.6E**). Furthermore, we also found by Western Blot analysis that the increase in *Stat1* mRNA was also paralleled by an increase in the protein level of total STAT1 as well as its two phosphorylated forms (Y701 and S727) (**Figure 4.6F**). We did not observe any change in either total STAT6 nor its phosphorylated form, indicating the specificity of the STAT1 signal. Looking at proteins downstream of STAT1, β2M shows an increase in cells exposed to Th1 EVs, consistently with a functional activation of STAT1. Importantly, all the changes observed in NIH 3T3 cells exposed to EVs were mirrored, although to a lesser extent, by cells exposed to EXOs (**Figure 4.6F**).

Taken together, these results show that Th1 EVs induce the activation of the IFN-γ signalling pathway in recipient cells. Our data suggest that this effect might be mediated by two (non mutually exclusive) mechanisms, namely the direct transfer of RNAs and/or proteins (Valadi et al., 2007; Kwon et al., 2014) or via the indirect induction of genes or activation of genes and/or proteins in target cells (Li et al., 2013a).

4.5 The EV-associated IFN-γ/Ifngr1 complex activates the Stat1 signalling pathway in target cells

To evaluate the contribution of the direct transfer of STAT1 protein and/or mRNA via Th1 EVs we measured the levels of STAT1 and other components of the IFN-γ pathway in NIH 3T3 cells *Stat1*$^{-/-}$ exposed to EVs derived from

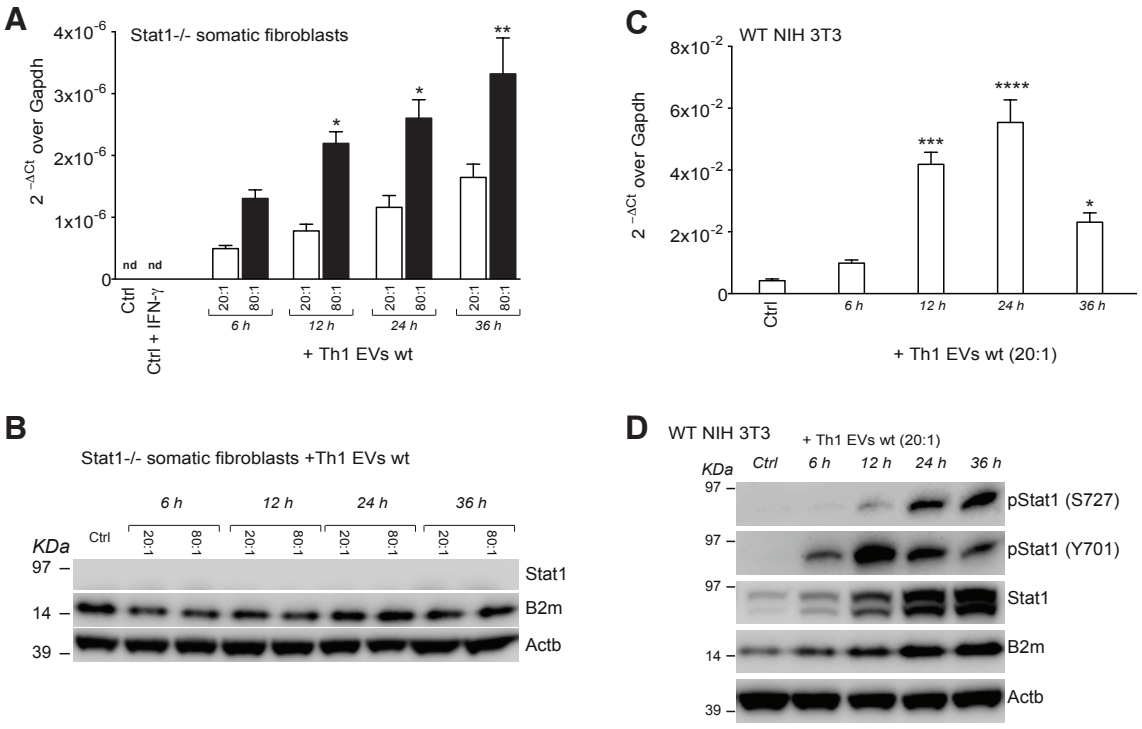


Figure 4.7 A,B: *qPCR (A) and Western blot (B) data showing the the expression of Stat1 and key RNAs of its pathway in Stat1$^{-/-}$ somatic fibroblasts exposed to two different ratios of wild type Th1 EVs for as long as 36 h in vitro.* ***C,D*** *Experiments as in A and B but on wild type recipient cells. The real-time PCR data are expressed as mean 2^{-dCt} (± SEM) over Gapdh from 3 independent experiments (****p-value<0.0001; ***p-value<0.001; *p-value<0.05, compared to 20:1). Western blot panels are representative of 5 independent experiments. Data published in Cossetti et al. (2014a).*

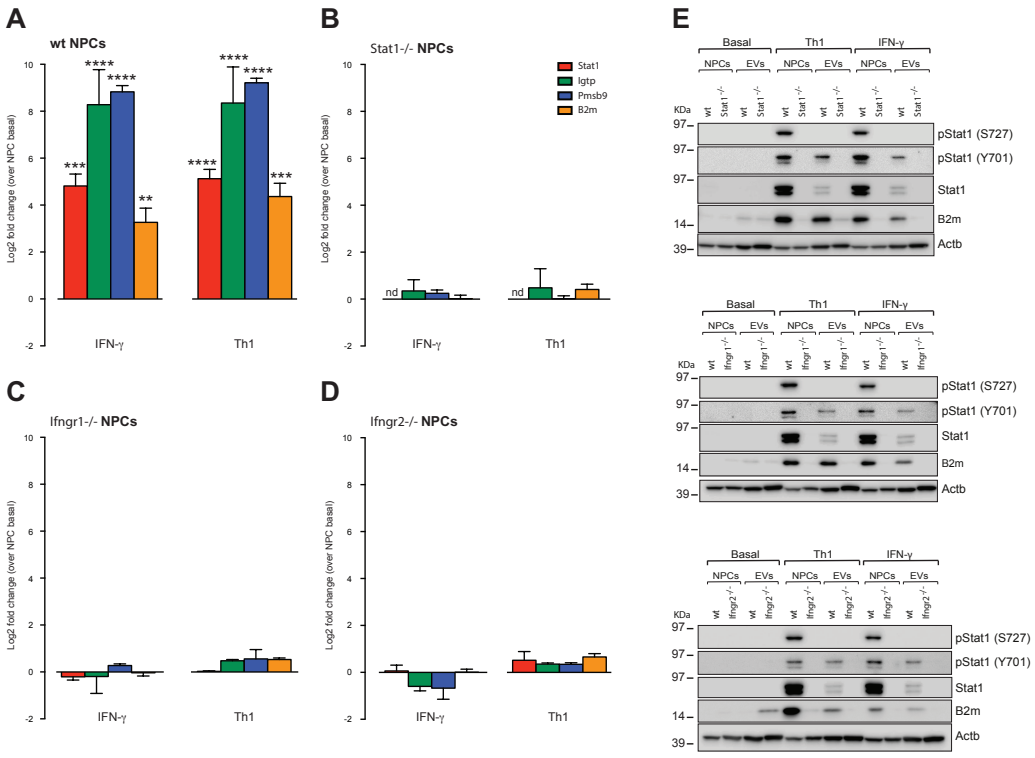


Figure 4.8 A-D: *qPCR quantification of key elements of the Stat1 pathway in wild type (A), Stat1$^{-/-}$ (B), Ifngr1$^{-/-}$ (C) and Ifngr2$^{-/-}$ (D) NPCs treated with either IFN-ɣor Th1 cytokines for 24h in vitro. The data represent the mean fold changes from 3 independent experiments (****p< 0.0001; ***p< 0.001; **p< 0.01, compared to NPC basal. nd= not detectable.)* ***E:*** *Western blots for component of the Stat1 pathway in NPCs and EVs as in A-D. Panels representative of 4 independent experiments. Data published in Cossetti et al. (2014a).*

Wilde Type (WT) NPCs treated with Th1 cytokines. Using qRT-PCR we found that the concentration of exogenous *Stat1* in *Stat1*$^{-/-}$ target cells increases proportionally with the incubation time ($p < 0.001$; $R^2 = 0.7684$ and $R^2 = 0.6061$ for EV:3T3 ratios of 20:1 and 80:1 respectively, **Figure 4.7A**), showing that Th1 EVs directly transfer *Stat1* mRNA to target cells. However, we could not detect by Western Blot a corresponding increase in STAT1 protein (**Figure 4.7B**), suggesting that the quantity of transferred exogenous STAT1 protein as well as STAT1 translated from transferred exogenous *Stat1* mRNA are below the detection limit of our method.

In order to precisely dissect the molecular events that led to the activation of STAT1 signalling in target cells we first decided to discriminate the relative contribution of each cytokine in the Th1 mix. We found that the effect of IFN-ɣ alone is comparable to that of the Th1 mix both in NPCs (**Figure 4.8A-E**) as well as in NIH 3T3 cells exposed to NPC-derived EVs . This result further confirmed that the EV component responsible for activating the STAT1 pathway in target cells is downstream of IFN-ɣ signalling.

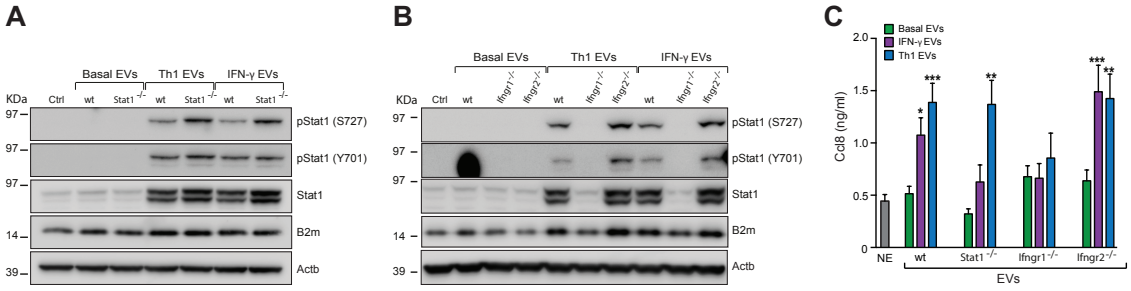


Figure 4.9 A,B: *Western blot of the Stat1 pathway in NIH 3T3 cells exposed to EVs from WT, Stat1$^{-/-}$ (A), Ifngr1$^{-/-}$ and Ifngr2$^{-/-}$ (B) NPCs. Panels representative of 3 independent experiments.* ***C:*** *ELISA measuring the release of CCL8 by target cells exposed to EVs as in (A) and (B). Data represented as mean ± SEM from a total of n ≤ 3 independent experiments. *p ≤ 0.01, **p ≤ 0.001, ***p ≤ 0.0001, compared to target cells not exposed (NE) to EVs. Data published in Cossetti et al. (2014a).*

Next, we generated Knock Out (KO) NPC lines lacking *Stat1* (Durbin et al., 1996), *Ifngr1* (alpha chain) (Huang et al., 1993) and *Ifngr2* (beta chain) and tested their capacity to respond to the Th1 cytokines mix as well as IFN-ɣ alone. As expected, we found that neither Th1 cytokines nor IFN-ɣ are able to activate STAT1 signalling in these knock-out NPC cell lines (**Figure 4.8A-E**). We then purified EVs from *Ifngr1*$^{-/-}$, *Ifngr2*$^{-/-}$ and *Stat1*$^{-/-}$ NPCs treated with IFN-ɣ and tested their capacity to activate STAT1 signalling in NIH 3T3 cells. We found that both *Ifngr2*$^{-/-}$ and *Stat1*$^{-/-}$ IFN-ɣ EVs were still capable of inducing STAT1 activation in target cells (**Figure 4.9A,B**). However, we also found that *Ifngr1*$^{-/-}$ IFN-ɣ EVs did not induce STAT1 activation when seeded on recipient NIH 3T3 cells.

In addition, to further analyse the functional relevance of STAT1 activation in target cells, we evaluated the level of the chemokine C-C motif ligand 8 (CCL8), which has been shown to be secreted by fibroblasts in response to pro-inflammatory cytokines such as IFN-ɣ and IL-1β (Gouwy et al., 2005; Struyf et al., 2009). We used Enzyme-Linked Immunosorbent Assay (ELISA) to assess the CCL8 production by target fibroblasts exposed to EVs derived from WT and KO NPCs in either basal, Th1 or IFN-ɣ only conditions. Interestingly, we found that the level of CCL8 produced by NIH 3T3 cells is significantly higher when exposed to Th1 and IFN-ɣ EVs derived from WT, *Stat1*$^{-/-}$ and *Ifngr2*$^{-/-}$ NPCs compared to basal conditions, but not from *Ifngr1*$^{-/-}$ NPCs (**Figure 4.9C**).

These data show that IFNGR1 is necessary to elicit the EV-mediated activation of IFN-ɣ signalling in target cells.

To determine whether the IFN-ɣ cytokine is itself trafficked via EVs together with IFNGR1, we measured the concentration of IFN-ɣ in the EV preparations using an ELISA assay. We determined that the amount of IFN-ɣ is in

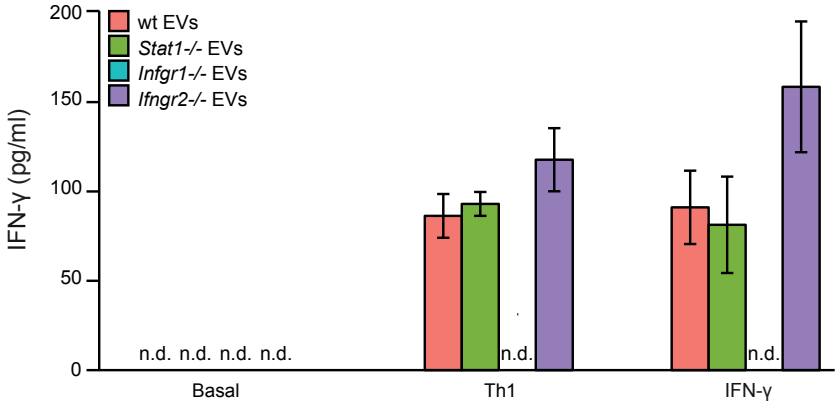


Figure 4.10 *Quantification of IFN-γ in EVs collected from 6 × 10^6 WT, Stat1$^{-/-}$, Ifngr1$^{-/-}$ or Ifngr2$^{-/-}$ NPCs. Data expressed as mean (± SEM) from a 3 independent EV preparations; nd: not detectable. Data published in Cossetti et al. (2014a).*

the range 52.13 pg/mL to 210.3 pg/mL (corresponding to ~15 pg of IFN-γ per ~30 µg of EV proteins) in WT, *Stat1*$^{-/-}$ and *Ifngr2*$^{-/-}$ EVs in both the Th1 and IFN-γ conditions, and we could not detect any significant difference between these samples (**Figure 4.10**). On the contrary, the concentration of IFN-γ was below the ELISA detection limit in EVs from *Ifngr1*$^{-/-}$ NPCs in both the Th1 and IFN-γ conditions.

To verify whether IFN-γ directly binds IFNGR1 on the surface of EVs we purified EVs from WT and *Ifngr1*$^{-/-}$ NPCs in basal conditions, and pretreated them directly with the same concentrations of IFN-γ used to treat NPCs in the previous experiments. We found that IFN-γ-treated EVs (basal$^{\text{IFN-γ}}$ EVs) from WT – but not from *Ifngr1*$^{-/-}$ – NPCs can recapitulate the effects on target cells of the EVs derived from Th1-treated NPCs, in terms of both STAT1 activation and CCL8 production (**Figure 4.11A,B**). Importantly, this experiment excludes the possibility that nonspecific interactions between the IFN-γ cytokine and the vesicles could lead to a passive carry-over of cytokines used to condition NPCs. Furthermore, these data also demonstrate that the IFN-γ/IFNGR1 complex on EVs is necessary to induce the activation of the STAT1 signalling pathway in target cells.

4.6 Target cells require Ifngr1 to sustain the EV-mediated activation of the Stat1 pathway

To further characterise the mechanisms through which the IFN-γ/IFNGR1 complex engages signalling in target cells, we generated somatic fibroblasts *Ifngr1*$^{-/-}$ and analysed their STAT1 activation in response to EVs from IFN-γ-

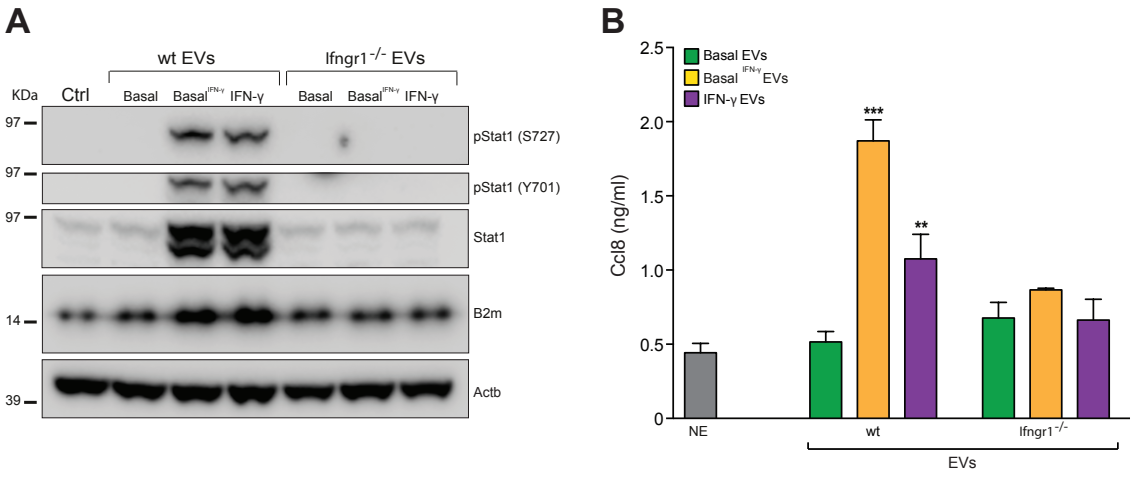


Figure 4.11 A: *Western blot of the Stat1 pathway in target cells exposed to basal EVs pretreated with 100 ng/mL IFN-ɣ (basal IFN-ɣ). Panel representative of 3 independent experiments.* ***B:*** *ELISA assay measuring Ccl8 release by target cells exposed to EVs as in (D). Data represented as mean ± SEM from a total of n ≤ 3 independent experiments. *p ≤ 0.01, **p ≤ 0.001, ***p ≤ 0.0001, compared to target cells not exposed (NE) to EVs. Data published in Cossetti et al. (2014a).*

treated WT NPCs as well as EVs from basal NPCs *in vitro* treated with IFN-ɣ (basal$^{IFN\text{-}ɣ}$ EVs). We found that both treatments induced a modest and dose-dependent upregulation of STAT1 and B2M in *Ifngr1*$^{-/-}$ target cells (**Figure 4.12A,B**). The upregulation of total STAT1 in *Ifngr1*$^{-/-}$ recipient cells was significantly lower than that of wild type recipient cells at all concentrations tested, and was not accompanied by an increase of the phosphorylated form. These data suggest that target cells require IFNGR1 to undergo the full activation of STAT1 signalling in response to EVs.

Finally, we estimated the kinetics of the binding of IFN-ɣ to its receptors to determine whether IFN-ɣ bound to the receptor on EVs could engage the receptor on target cells. Based on evidence reported in the literature that measured the K_{off} of the IFN-ɣ/IFNGR1 complex (Sadir et al., 1998) we estimated that the half-life of the ligand-receptor complex is ~139 s. Consequently, every ~2 minutes half of the receptor-bound IFN-ɣ dissociates from the complex becoming available for binding other competing receptors. Considering that the incubation time of EVs with target cells is 24 h it is reasonable to assume that our experimental design allows enough time for IFN-ɣ to reach a binding equilibrium between the competing receptors on EVs and target cells.

In conclusion, our data show that NPC-derived EVs stimulated with IFN-ɣ are able to shuttle IFN-ɣ to target cells through the receptor IFNGR1. In turn, EV-transferred IFN-ɣ engages the endogenous IFNGR1 of the target cells activating STAT1 signalling (**Figure 4.12C**). These results highlight a novel mechanism of cell-to-cell communication where the direct transfer of a cytokine

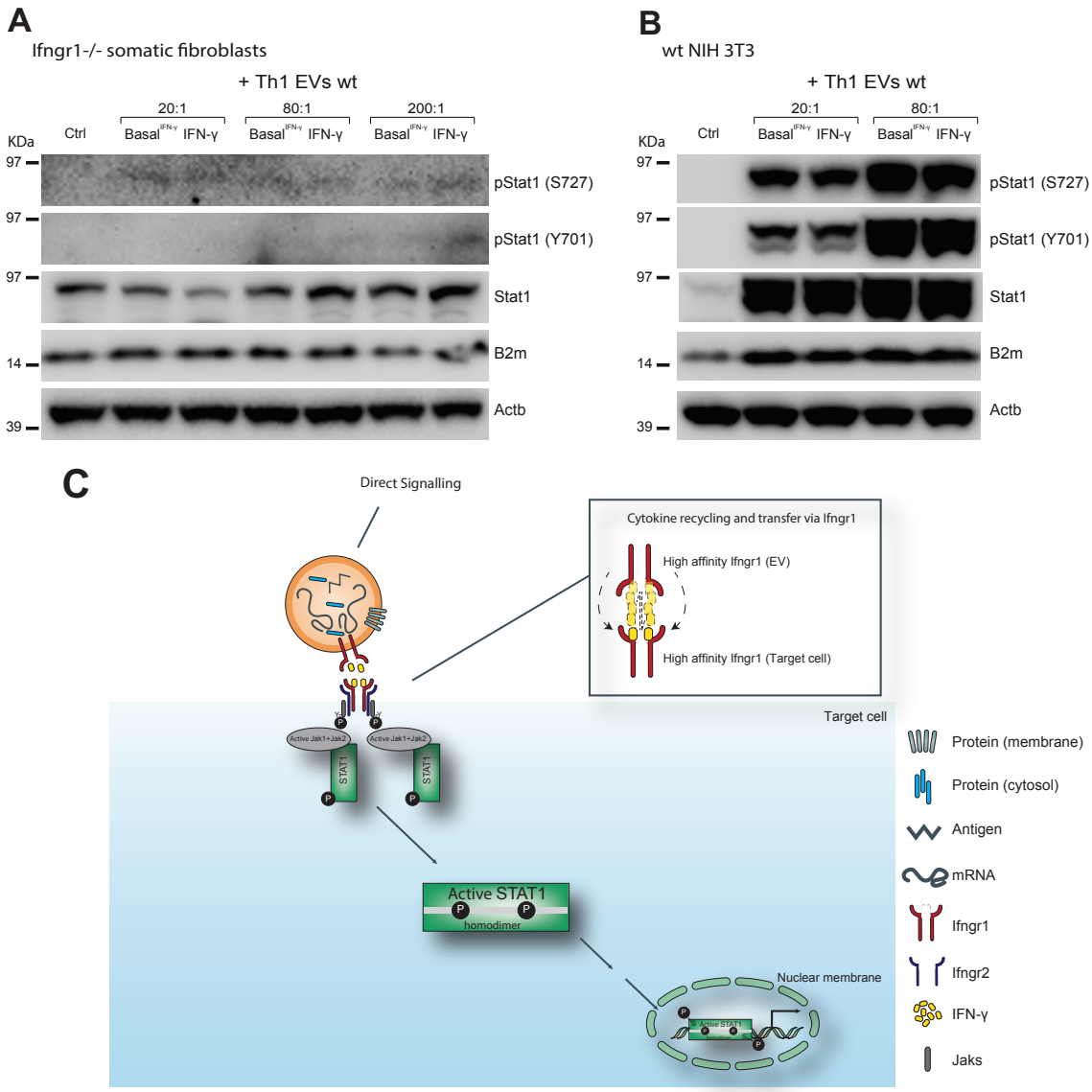


Figure 4.12 A,B: *Western blot of components the Stat1 pathway in Ifngr1$^{-/-}$ (A) or WT (B) somatic fibroblasts (A) or NIH 3T3 cells (B) target cells treated with different ratios (3 µg to 300 µg of total EV protein per treatment) of WT IFN-γ-induced and basal IFN-γ EVs for 24 hr in vitro. Data representative 3 independent experiments.* ***C:*** *Proposed mode of action of EV-based IFN-γ signalling spreading. Data published in Cossetti et al. (2014a).*

via its receptor is able to propagate at a distance the activation of a signalling pathway. These results will be discussed in detail in Chapter 7, whereas the next chapter will report on the investigation of the molecular mechanisms that drive small non-coding RNA secretion in stem cells.

5

Secretion mechanisms of extracellular microRNAs in Neural Stem Cells

The work presented in this chapter is the result of a collaboration between the Enright and Pluchino laboratories. The samples for RNA-Seq have been prepared by Dr Cossetti and Dr Iraci in the laboratory of Dr Pluchino. The luciferase and ChIP assays have been done by Dr Iraci as specified in the figure legends.

5.1 Introduction

The transplantation of somatic Neural Progenitor Cells (NPCs) ameliorates the clinical outcome of several neuroinflammatory disorders of the Central Nervous System (CNS) (Martino et al., 2011). The molecular mechanisms responsible for the beneficial function of transplanted stem cells are still not completely clear, however increasing evidence suggests that they mediate neuroprotection and immunomodulation by engaging a complex cross talk with the immune system (Pluchino et al., 2005; Martino et al., 2011).

The possible routes of intercellular communication between NPCs and the immune system include the secretion of soluble factors (such as cytokines or growth factor), direct cell-to-cell contacts or the exchange of intracellular molecules through gap junctions (Pluchino and Cossetti, 2013). The secretion of Extracellular Vesicles (EVs) is an additional mode of cell-to-cell communication that has sparked great interest in recent years. Compared to the other classical mechanisms of communication, EVs are of particular interest because of their capacity to transfer at a distance a variety of molecules, such as RNAs, proteins, lipids and metabolites. Moreover, EVs and exosomes play important roles in the exchange of signals between immune cells (Raposo et al., 1996; Bobrie et al., 2011; Mittelbrunn et al., 2011), suggesting that they might also be important players in the communication between transplanted stem cells and the host immune system.

In this work we focused on the EV-mediated transfer of miRNAs from murine NPCs to the surrounding microenvironment. We first provide a comprehensive characterisation of miRNAs contained within EVs and exosomes, finding that a specific subset of them is significantly enriched in vesicles compared to NPCs. This finding suggested the existence of a dedicated sorting machinery that selectively routes some miRNAs towards exosomes, as previ-

ously described by others for other species and cell types (Villarroya-Beltri et al., 2013). We hypothesized that this mechanism might act concurrently on two levels:

1. through the concerted transcriptional regulation of secreted miRNAs by a dedicated set of transcription factors.

2. through carrier proteins that recognise specific short motif in secreted miRNAs and facilitate their localisation toward exosomes.

In pursuit of the first hypothesis, we first characterised the genomic locations of murine miRNA promoters, and then analysed their sequences in search of Transcription Factor Binding Sites (TFBSs) enriched in the promoters of secreted miRNAs. However, this analysis revealed that under our experimental conditions no specific TFBSs are enriched in the promoters of secreted miRNAs.

We therefore set to identify short sequence motifs enriched in the mature sequence of secreted miRNAs. This analysis revealed the presence of two significantly enriched motifs whose presence correlates with miRNA secretion. RNA immunoprecipitation experiments revealed that secreted miRNAs that possess these motifs are bound by hnRNPA2/B1, suggesting that this protein might act as an exosomal miRNA carrier, as previously described by others in a human T lymphocytes cell line (Villarroya-Beltri et al., 2013).

5.2 Characterisation of the small RNA population of NPC-derived Exosomes and EVs

To characterise the smallRNA population of NPCs and NPC-derived vesicles, we modelled *in vitro* the inflammatory environment that NPCs are likely to encounter when transplanted in an animal model of a CNS inflammatory disorder. To this purpose, NPCs were harvested from the SVZ of SJL mice and cultured *in vitro* for 16 hours with either serum free media (hereafter referred to as basal condition), or serum free media added with Th1 cytokines (IFN-ɣ, TNF-α, IL-1β, hereafter referred to as Th1 condition) or Th2 cytokines (IL-4, IL-5, IL-13, hereafter referred to as Th2 condition) to mimic a pro-inflammatory or anti-inflammatory environment respectively, as previously described (Pluchino et al., 2008).

In the effort to fully characterise the content of NPC-derived EVs, we performed an RNA-Seq experiment for small RNAs. To this end, total RNA was purified from NPCs, EVs and Exosomes (EXOs) in Basal, Th1 and Th2 conditions and used to build standard Illumina small RNA-Seq libraries which were then sequenced (see Methods, section 10.1).

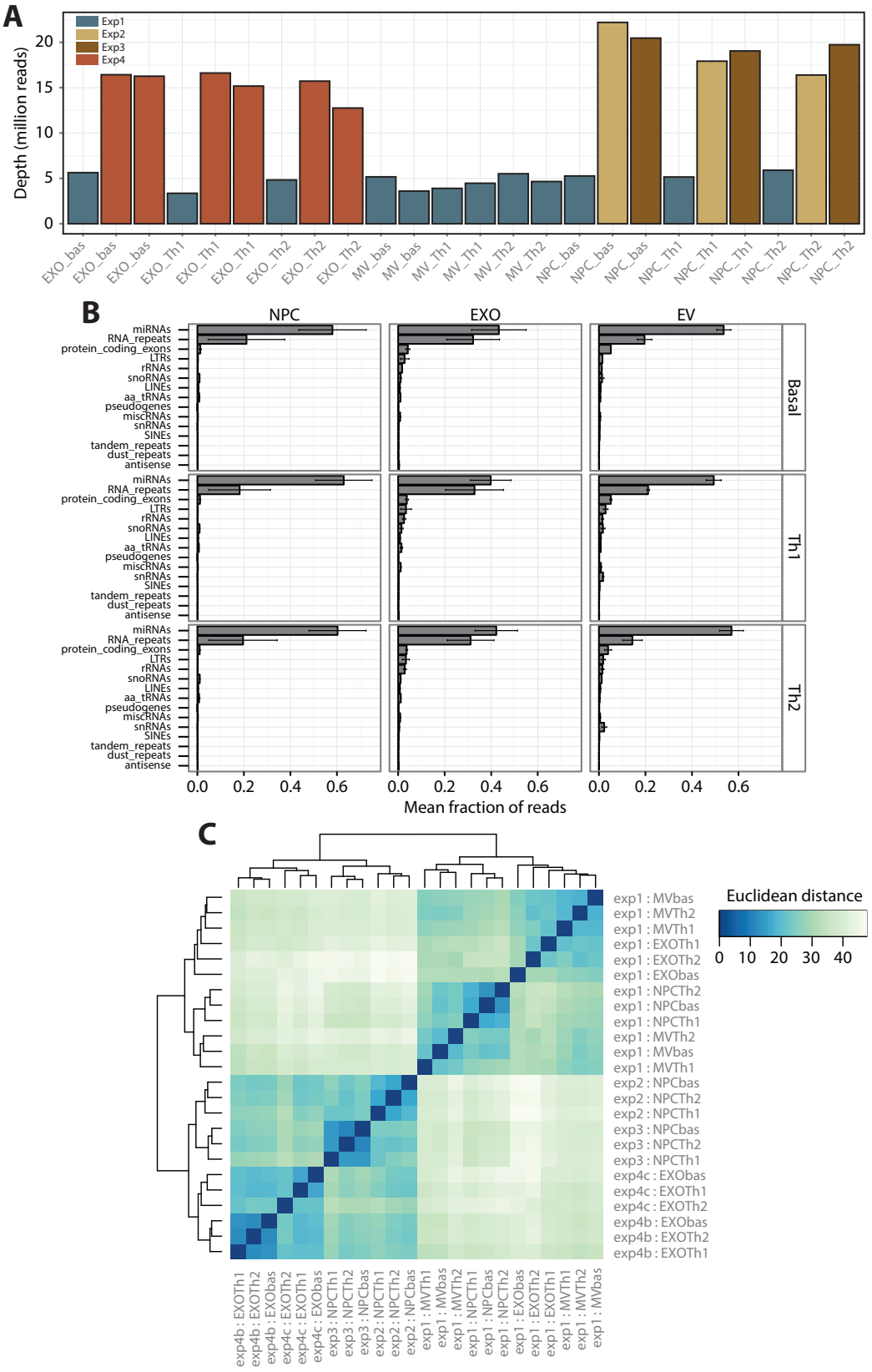


Figure 5.1 ***A:****Bar chart showing the number of reads sequenced for each sample.* ***B:*** *Graph showing the fraction of reads mapping to each type of genomic feature.* ***C:*** *Heatmap showing the Euclidean distance between small RNA-Seq miRNA expression profiles in each sample.*

On average, we sequenced 11 million reads per sample (**Figure 5.1A**), and we used the Kraken pipeline (Davis et al., 2013) to map them to the mouse reference genome and quantify small RNAs (see Methods, section 10.1). We found that microRNAs (miRNAs) are the most abundant class of sequenced small RNAs in all samples (**Figure 5.1B**), consistently with the RNA size selected during library preparation. Based on this observation and the established role of miRNAs in EXOs, we focused our subsequent analysis on this class of small RNAs.

Interestingly, we noticed that the major factor that separated samples was the batch (i.e. samples in the same replicate tend to cluster together, **Figure 5.1C**). Therefore, to identify miRNAs significantly up- or down-regulated across samples or conditions we used DESeq2 (Love et al., 2014) to apply a generalised linear model that takes into account batch effects (see Methods, section 10.1).

We first investigated whether cytokine treatment alters miRNA expression in NPCs, EVs or EXOs. In general, we found that cytokines have a modest impact on miRNA expression, with only 12 miRNAs being differentially expressed in NPCs (p-value$<$0.01, **Figure 5.2A**) upon Th1 stimulation and none upon Th2 stimulation. Similarly, also the miRNA repertoire of EVs and EXOs is remarkably stable in response to cytokine stimulation and we only found one miRNA differentially expressed in Th1 EVs (**Figure 5.2B,C**).

Next, we compared the expression of secreted miRNAs in EVs and EXOs to their expression in parent cells. Considering the modest effect of cytokine stimulation, we now analysed each sample disregarding the cytokine treatment (i.e. using the treatment condition as a co-factor in the linear model, see Methods, section 10.1); this approach has a twofold advantage: first, it allowed us to identify changes that are independent of cytokine stimulation; second, it increased the statistical power of the analysis by increasing the sample size. Interestingly, we found that a remarkable number of miRNAs are significantly more abundant inside EVs and EXOs than in NPCs (**Figure 5.3A**). Specifically, we observed that 69 miRNAs are significantly upregulated in EXOs vs NPCs (p-value$<$0.01, **Figure 5.3A,C**), while 41 are significantly upregulated in EVs vs NPCs (p-value$<$0.01, **Figure 5.3B,C**). For example, miR-181c-5p is significantly more abundant in EXOs than in NPCs ($\log_2$ fold change 2.5, adjusted p-value = 6.8×10^{-12}, **Figure 5.3D**). On the other hand, other miRNAs such as miR-222-3p show an opposite trend, with high expression in NPCs and low abundance within EXOs and EVs ($\log_2$ fold change -1 in EXO vs NPC, adjusted p-value = 0.98, **Figure 5.3E**).

Finally, to evaluate the functionality of secreted miRNAs, we transfected NIH/3T3 embryonic fibroblasts with reporter luciferase constructs carrying

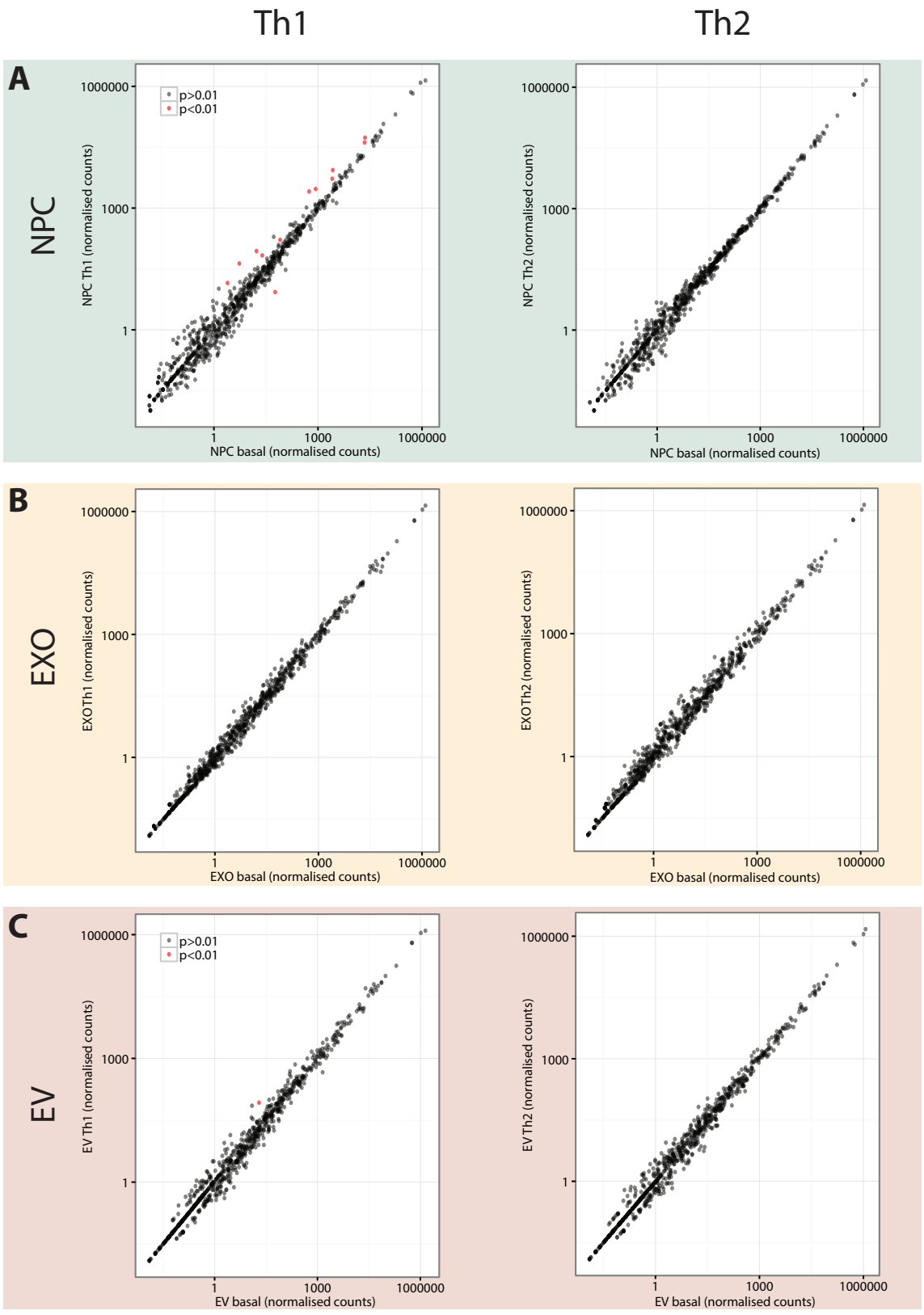


Figure 5.2 A-C: Effect of stimulation with Th1 (left) or Th2 (right) cytokines on the expression of miRNAs in NPCs (A), EXOs (B), and EVs (C). Points in red indicate miRNAs with adjusted p-value <0.05

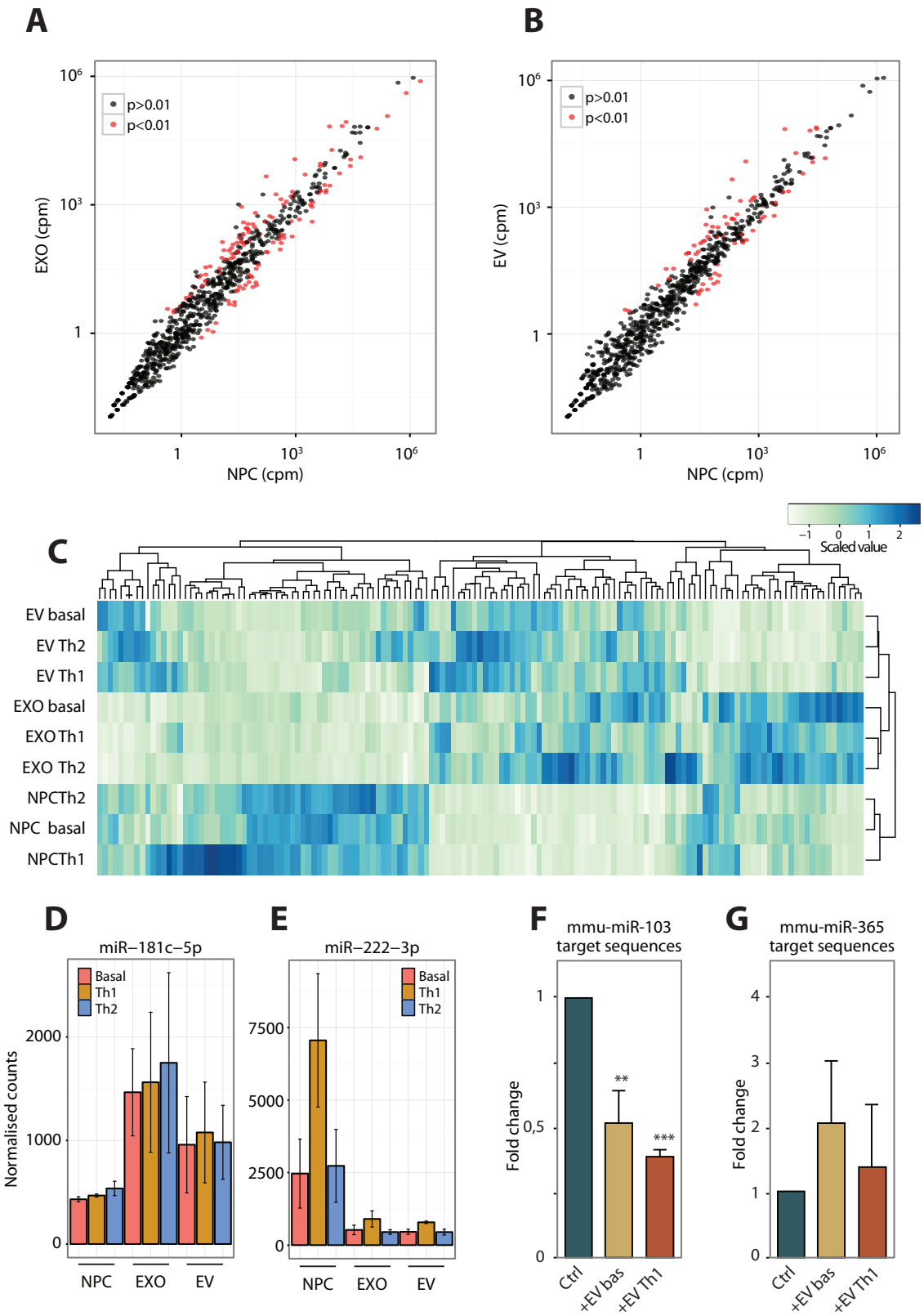


Figure 5.3 legend on next page

Figure 5.3 (previous page) ***A,B****: Scatterplots comparing the expression of miRNAs in EXOs (****A****) or EVs (****B****) to NPCs.****C****: Heatmap showing the scaled expression level of all miRNAs that are significantly up- or down-regulated in at least one condition.****D, E****: Bar chart showing the expression profiles of miRNA 181c-5p (****D****) and miRNAs 222-3p (****E****) across conditions.* ***F, G****: Reporter luciferase assay on NIH-3T3 cells responsive to the levels of either miR-103 (highly abundant in exosomes) (****F****) or miR-365 (low abundance in exosomes) (****G****) by cloning their target sequence downstream of the coding region in the 3'UTR. Luciferase activity was monitored as a function of EXO treatment compared to untreated cells. Data are expressed as fold change (+/- SEM) from a total of 4 independent experiments. *** p-value<0.001; ** p-value<0.01 (Anova, treated vs Ctrl). The experiments for panels* ***F*** *and* ***G*** *were performed by Dr Nunzio Iraci in the laboratory of Dr Pluchino (Department of Clinical Neuroscience, University of Cambridge).*

miRNA binding sites for either miR-103-3p or miR-365-3p, which are present in EVs and EXOs at a high and low levels respectively (average normalised counts in EXOs of 93 305.26 and 83.43 for miR-103-3p and miR-365-3p respectively). We found that in NIH/3T3 cells incubated *in vitro* with EVs purified from NPC, the luciferase activity of the miR-103-3p reporter was significantly decreased (**Figure 5.3F**), while the activity of the reporter for miR-365-3p was not affected (**Figure 5.3G**). These data suggest that miRNAs contained within NPC-derived EVs are functional and are capable of downregulating their target genes in recipient cells.

Overall, these results show that a subset of miRNAs is enriched in EVs and EXOs and their level is remarkably stable to cytokine treatment, suggesting that the EXO/EV miRNA profile is the result of a tightly regulated process and does not simply reflect the transcriptional landscape of the parental cell. We reasoned that the molecular machinery responsible for controlling the secretion of miRNAs might act on two non-exclusive levels, either by controlling the transcription of secreted miRNAs or by directly shuttling miRNAs toward exosomes through a specific RNA-binding protein (carrier). The following paragraphs describe the results obtained in pursuit of these two hypothesis.

5.3 Mechanisms of transcriptional regulation of secreted miRNAs

We hypothesized that miRNAs secreted in exosomes and EVs might be transcriptionally controlled by the concerted action of a specific set of transcription factors. Furthermore, there is evidence in *Saccharomyces cerevisiae* showing that the presence of a specific sequence in the promoters of a set of mRNAs can influence their subcellular localisation, possibly through the recruitment of RNA binding proteins or alterations in the RNA secondary structure (Zid and O'Shea, 2014). Although there is no evidence of such mechanism in other eukaryotes, we speculated that the presence of sequence motifs in the pro-

moters of secreted miRNAs might influence their subcellular localisation and favour their loading in EVs and exosomes.

To verify the validity of this hypothesis, we first implemented a computational pipeline to determine the genomic location of murine microRNA promoters. We then scanned their sequences in search of sequence motifs enriched in the promoters of secreted miRNAs.

5.3.1 *Annotation of miRNA promoters in the mouse genome*

The canonical promoter of protein coding genes is usually located in genomic proximity of their Transcriptional Start Sites (TSSs). However, miRNA primary transcripts have a short half-life, hence their TSS is seldom identified by traditional RNA-Seq experiments. Numerous works have attempted to identify the TSSs and promoters of miRNAs in human (Corcoran et al., 2009; Ozsolak et al., 2008; Saini et al., 2007; Zhou et al., 2007; Barski et al., 2009; Marsico et al., 2013; Georgakilas et al., 2014) , mouse (Alexiou et al., 2009), worm (Zhou et al., 2007) and plants (Megraw, 2006; Zhou et al., 2007). However, to the best of our knowledge, most of the studies in mouse only made use of the limited number of experimental techniques and/or sources of information available at the time of publication. Therefore, we implemented an integrative approach based on computational predictions, sequence annotations and high-throughput chromatin modification data generated by the ENCODE project to produce an accurate and comprehensive annotation of miRNA promoters in the mouse genome.

For each murine pre-miRNA annotated in miRBase (Kozomara and Griffiths-Jones, 2011) we scanned a genomic region of 100 kb upstream looking for six genomic features that could indicate the presence of a promoter. The features that we considered where:

1. CpG islands, from the CpG track of the UCSC genome browser (Gardiner-Garden and Frommer, 1987). CpG islands are known promoter features and are found in the promoters of approximately 40 % of the human genes (Larsen et al., 1992).

2. DNaseI Hypersensitivity Sites (DHSs) from ENCODE/University of Washington on 48 cell lines. DNase Hypersensitivity is a property of chromatin regions bound by transcription factors and regulatory proteins and is often associated with gene promoters (Thurman et al., 2012).

3. ChIP-Seq data for trimethylation of histone 3 lysine 4 (H3K4me3) on 32 cell lines from the ENCODE project (PSU, LICR and Caltech; ENCODE Project Consortium et al., 2012) and on Neural Stem Cells from

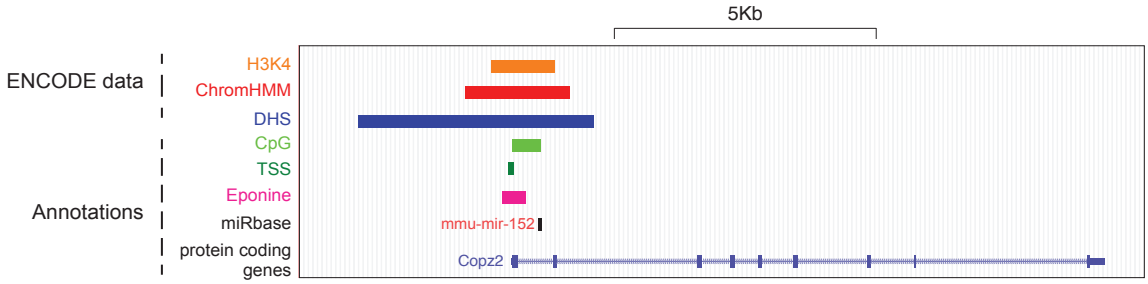


Figure 5.4 *Visualization of the genomic region surrounding miR-152. This intronic miRNA is located within the second intron of the gene Copz2. Overlapping the TSS of Copz2 there are an H3K4me3 peak, ChromHMM peak, Cpg Island, DHS, host gene TSS and Eponine prediction, supporting the presence of the miRNA TSS in this region.*

Mikkelsen et al. (2007). H3K4me3 is a histone modification associated with active promoters that has been widely used to identify promoters of protein-coding genes (Kim et al., 2005).

4. Eponine predictions of TSSs. Eponine is a tool that uses DNA weight matrices to identify specific sequence motifs indicative of transcriptional start sites (Down and Hubbard, 2002). This tool was already successfully used to identify miRNA promoters in the human genome (Saini et al., 2007).

5. Annotated TSS of host gene for intronic miRNAs. Out of 734 miRNA transcripts mapped in the mouse genome (miRBase v18), 416 (56.7 %) are contained within an intron of a protein-coding gene (host gene). Many – although not all – intronic miRNAs are co-transcribed with their host gene (Rodriguez, 2004), therefore the TSS of the host can be used as a proxy for locating the miRNA promoter.

6. Genome-wide chromatin segmentation using chromHMM based on data produced by the ENCODE consortium (ENCODE Project Consortium et al., 2012; Ernst and Kellis, 2012). This method assigned a putative function to each region of the human genome based on the patterns of chromatin modifications. We selected the regions identified as promoters and identified their corresponding syntenic regions in the mouse genome.

Taken together, these data can support the presence of a promoter in a given chromatin region. For example, upstream of pre-miR-152 we find a peak of H3K4me3, a ChromHMM promoter, a CpG island, a DHS peak and an Eponine TSS predictions (**Figure 5.4**). In this case, the pre-miRNA resides in the first intron of the host gene *Copz*, suggesting that the transcription of the pri-miRNA is driven by the host gene's promoter.

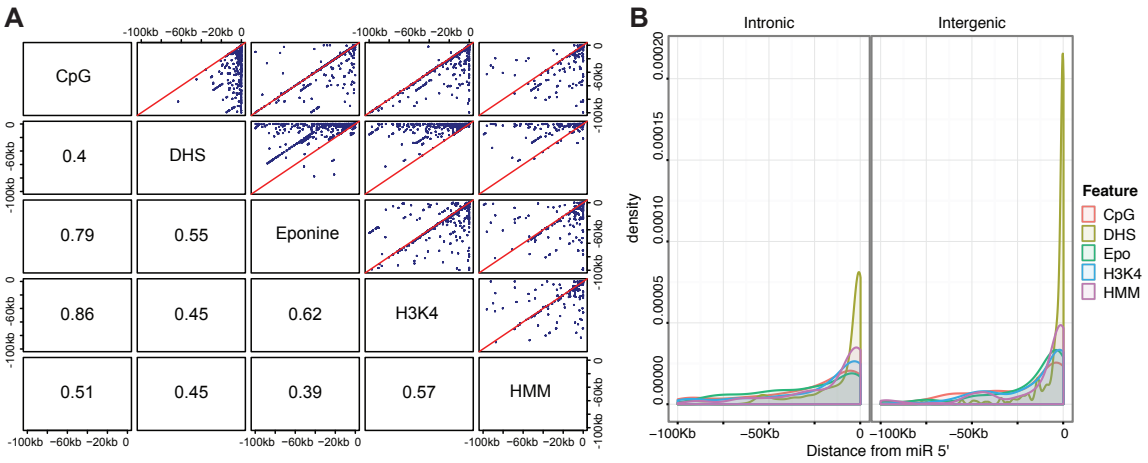


Figure 5.5 A: *Scatterplot showing the position of each feature relative to the annotated miRNA 5' end. Pearson correlation coefficients between features are shown.* ***B:*** *Density distribution of distances of each feature from the annotated miRNA 5' end.*

When we measured the distance between miRNAs and the closest upstream features of each type, we found a high concordance in their position, suggesting that multiple features often appear in the same genomic region (**Figure 5.5A**). Interestingly, we also found that the closest upstream features typically reside ~5kb upstream of the annotated miRNAs (**Figure 5.5B**), in accordance with previous analysis on the location of human miRNA promoters (Saini et al., 2007).

In order to identify high confidence promoters, we merged features less than 200bp apart (approximating the length of DNA in one nucleosome) into unique clusters (i.e. candidate promoters), and assigned a score to each cluster to reflect how many features supported it (**Figure 5.6A**). This score was then weighted according to its distance from the miRNA, so that in cases where two candidate promoters had the same score the closest would be preferred (see Methods, section 10.3). Additionally, candidate promoters that overlapped the annotated TSS of an upstream transcript that did not overlap the miRNA where given a negative score, to avoid annotating as miRNA promoters the promoters of extraneous upstream genes (see Methods, section 10.3).

We found, on average, 15.7 candidate promoters for each miRNA (**Figure 5.6B**) with a mean size of 1.4 kb (**Figure 5.6C**) in accordance with the typical size of known promoters. In general, candidate promoters tend to be uniformly distributed in the region 100kb upstream of miRNAs (**Figure 5.6D**). However, after selecting for each miRNA only the candidate promoter with the highest score, we found a clear peak ~5kb upstream of the miRNA (**Figure 5.6E**). This supports the validity of our approach and is in accordance with previous findings on human miRNA promoters (Saini et al., 2007).

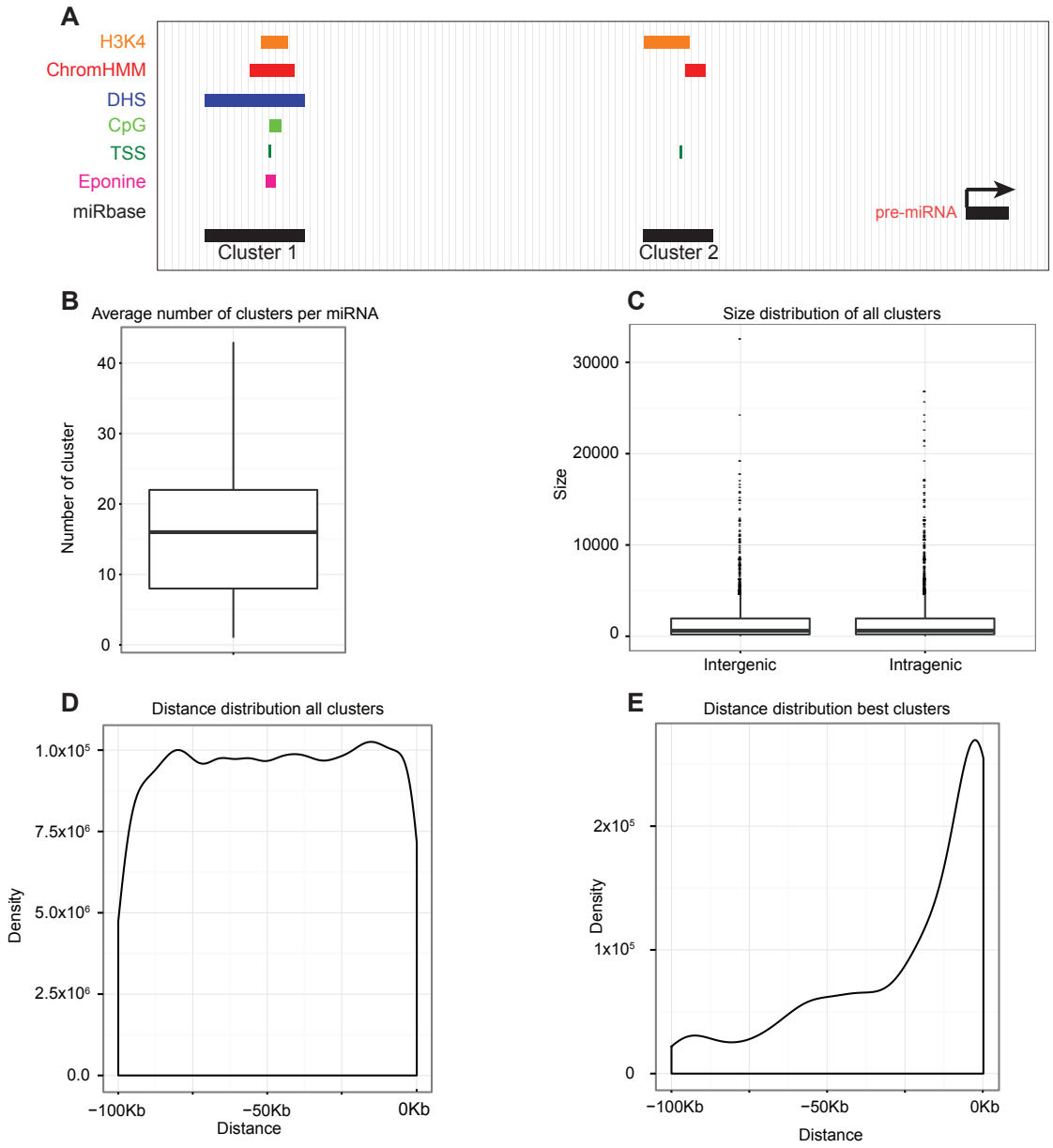


Figure 5.6 ***A:*** *Schematic diagram showing two examples of clusters.* ***B:*** *Distribution of average number of clusters per miRNA.* ***C:*** *Distribution of cluster size.* ***D:*** *Distance of all clusters from the corresponding miRNA.* ***E:*** *Distance of each miRNA's highest scoring cluster from the miRNA itself.*

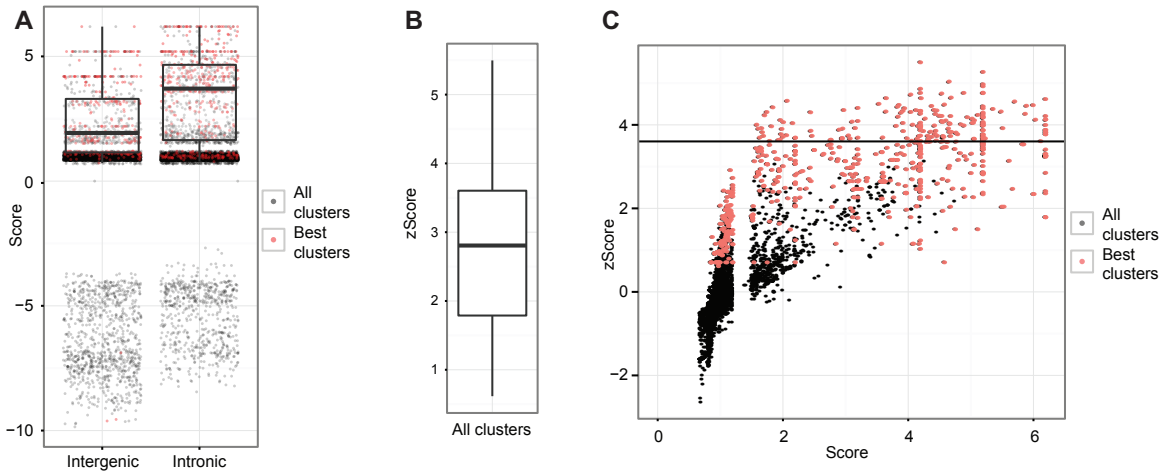


Figure 5.7 ***A:*** *Distribution of cluster scores for Intergenic and Intronic miRNAs. Best clusters are highlighted in red.* ***B:*** *Distribution of z-scores for all clusters.* ***C:*** *The Score of each cluster is plotted against its z-score. Best clusters are highlighted in red. The horizontal line indicates the third quartile of the distribution of z-scores.*

The scores of candidate promoters for intronic and intergenic miRNAs were similarly distributed (**Figure 5.7A**), although the best scoring promoters for intronic miRNAs tend to have on average a higher score. This is likely due to the fact that for intronic miRNAs we consider the TSS of the host gene as an extra feature that, by definition, doesn't apply to intergenic miRNAs.

As previously mentioned, we find on average 15.7 candidate promoters for each miRNA, of which the highest scoring one is the most likely promoter. In order to further characterise the best scoring promoters we introduce a z-score metric to measure how much stronger is the support for the best scoring promoter compared to the other candidate promoters of a given miRNA. We find that the best promoters have an average z-score of 2.7 (**Figure 5.7B**), i.e. their score is on average 2.7 standard deviations above the mean score of all the candidate promoters of a given miRNA. Nevertheless, we find cases where the same miRNA has multiple candidate promoters with high score and high z-score, suggesting that in some cases one miRNA could have multiple alternative promoters (**Figure 5.7C**).

The inspection of the individual best promoters revealed that the vast majority of them are supported by a DHS mark, more than 50 % have both a DHS and an H3K4me3 mark and 30 % have a DHS mark, an H3K4me3 mark, an Eponine prediction and a CpG island (**Figure 5.8A**). Additionally, the highest scoring candidate promoters are on average more conserved than all candidate promoters and more conserved than random intergenic regions (**Figure 5.8B** p-values = 1.03×10^{-5} and 2.2×10^{-16} respectively).

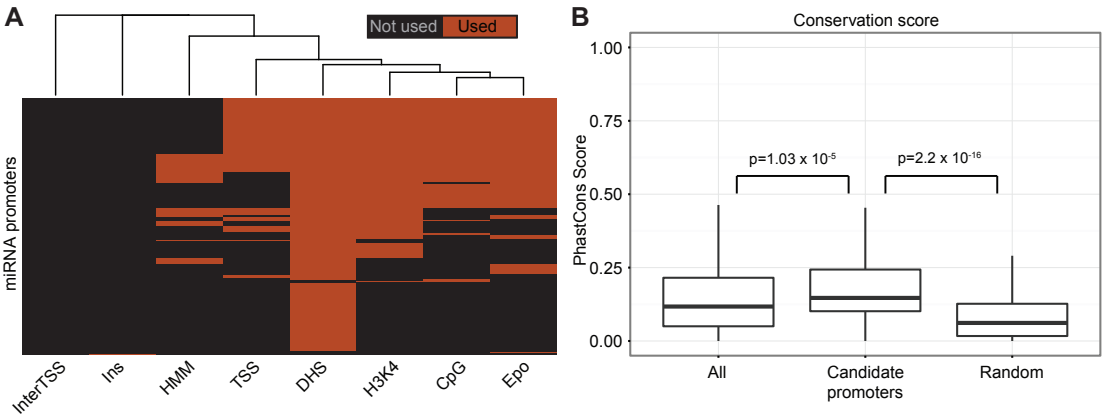


Figure 5.8 A: *Heatmap showing what features support each candidate promoter (best cluster). Each row represents a best cluster while each column represents a feature. Yellow cells indicate that the given feature was used to support the given best cluster.* ***B:*** *PhastCons conservation score of all clusters, best clusters and random genomic sequences.*

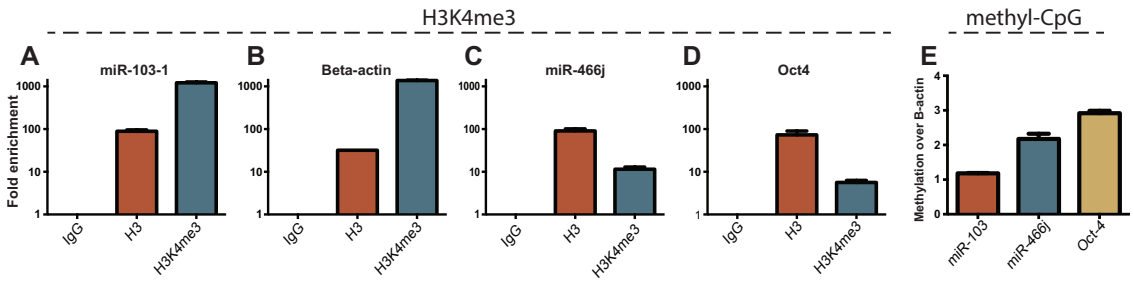


Figure 5.9 A-D: *In vivo ChIP analyses of selected miRNA and mRNA promoters in NPCs. Results show one experiment in which each region was amplified by qPCR in triplicate. SEM is indicated. Relative enrichment of a given promoter region obtained with a specific antibody was compared with that obtained with pre-immune serum (IgG), which was set to 1 in the graph. The genes beta actin and Oct4 are respectively expressed and not expressed in NPCs and their promoters acted as a positive and negative control respectively.* ***E:*** *mCIP analyses of selected miRNA and mRNA promoters in NPCs. Results show one experiment in which each region was amplified by qPCR in triplicate. SEM is indicated. Relative methylation level of a given promoter region was compared with that obtained on beta-actin promoter, which was set to 1 in the graph. These experiments were performed by Dr Iraci in the laboratory of Dr Pluchino.*

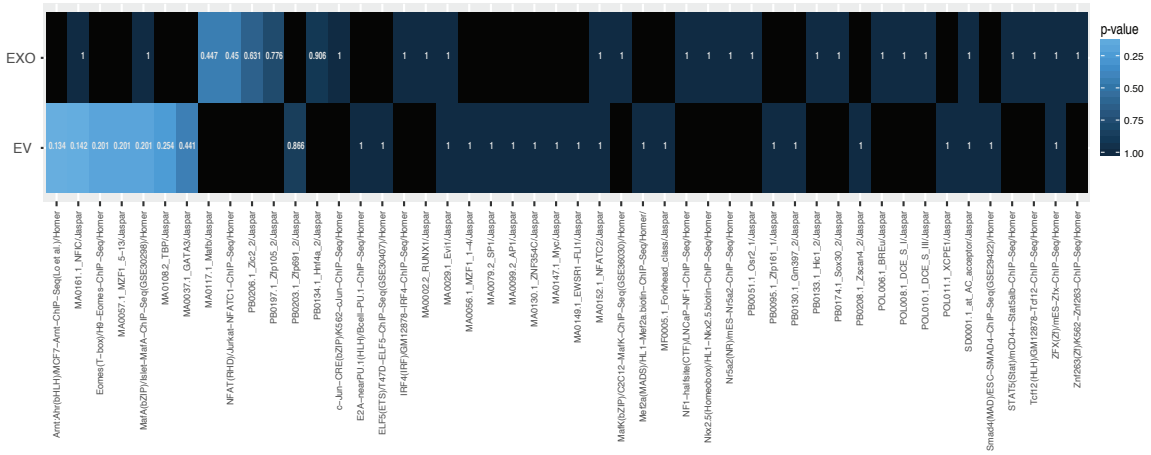


Figure 5.10 *Heatmap showing the p-values of enrichment of Transcription Factor Binding Sites (x-axis) in the promoters of EV-enriched or EXO-enriched miRNAs.*

Finally, by chromatin immunoprecipitation assay (ChIP) and methyl-CpG immunoprecipitation (mCIP) we validated H3K4me3 and CpG methylation of the predicted promoters of miR-103, which is expressed in NPCs and miR-466j, which instead is not expressed in NPCs (**Figure 5.9A-E**); this experiment confirmed high H3K4me3 and low methyl-CpG for miR-103, and an opposite trend for miR-466j, supporting the validity of our promoter prediction approach. Taken together, these results provide a valuable resource that will allow us to investigate the presence of enriched TFBSs in the promoters of secreted miRNAs.

5.3.2 *Enriched TFBS in the promoters of secreted miRNAs*

After identifying the genomic location of all murine miRNA promoters we analysed the sequence of the promoters of miRNAs significantly enriched in EXOs and EVs (vs NPCs, p-value$<$0.05) in search of enriched TFBSs. To this end we used Homer (Hypergeometric Optimization of Motif EnRichment; Heinz et al., 2010), an algorithm that performs *de novo* motif discovery and reports those enriched in a given set of sequences compared to a background set (see Methods, section 10.4). This analysis revealed that none of the transcription factors motifs identified in miRNA promoters is significantly enriched in the promoters of secreted miRNAs (**Figure 5.10**). This result suggests that under our experimental conditions exosomal miRNAs are not likely to share common mechanisms of transcriptional regulation.

In light of these data, we focused our analysis on potential post-transcriptional mechanisms that might drive the secretion of miRNAs.

5.4 Mechanisms of post-transcriptional regulation of secreted miRNAs

5.4.1 *Identification of short motifs enriched in secreted miRNAs*

The analysis of the expression profiles of NPC-derived exosomal miRNAs suggests the existence of a dedicated machinery that is able to recognise a subset of miRNAs and promote their sorting towards exosomes. To investigate the molecular determinants responsible for this specific secretion mechanism we analysed the sequence of EXO-enriched miRNAs in search of short sequence motifs that could be recognised by candidate carrier proteins.

To this end, we used the R/Bioconductor package BCRANK (Ameur et al., 2009), a tool that scans an ordered list of sequences and reports short motifs that are over represented at the top of the list. After filtering miRNAs with low expression levels, we selected those that were significantly enriched or depleted in EXOs (p-value$<$0.05, see Methods, section 10.5) and ordered their sequences according to the fold change in EXOs vs NPCs. BCRANK identified various motifs (reported in Table 5.1) which are enriched in the sequence of EXO-enriched miRNAs.

We found that the highest scoring motif has the consensus sequence `KGVGH`[1] (**Figure 5.11A**) and is present in 65 miRNAs, while the second and third highest scoring motifs (consensus sequences `VVKGVG` and `DCBCM`, **Figure 5.11B,C**) are present in 37 and 46 miRNAs respectively. All three motifs showed a clear preference for secreted miRNAs, although they are also present in a small number of retained miRNAs (**Figure 5.11D-F**). The comparison of the consensus sequences suggested that `KGVGH` and `VVKGVG` are offsets of the same G-rich motif, while the third consensus sequence (`DCBCM`) represents a separate C-rich motif.

5.4.2 *Differential secretion of miRNA 5p and 3p arms.*

The analysis of the smallRNA-Seq data showed that 12 miRNAs undergo arm-switch between cells and EXOs, meaning that one of the two arms of the mature miRNA is significantly (p$<$0.05) higher in EXO vs NPC while the other arm is significantly higher in NPC vs EXO (**Figure 5.12A,B**). Table 5.2 reports the names, the fold changes and the p-values of miRNAs that undergo arm switch.

[1] The notation used for the consensus sequences follows the IUPAC guidelines (Cornish-Bowden, 1985).

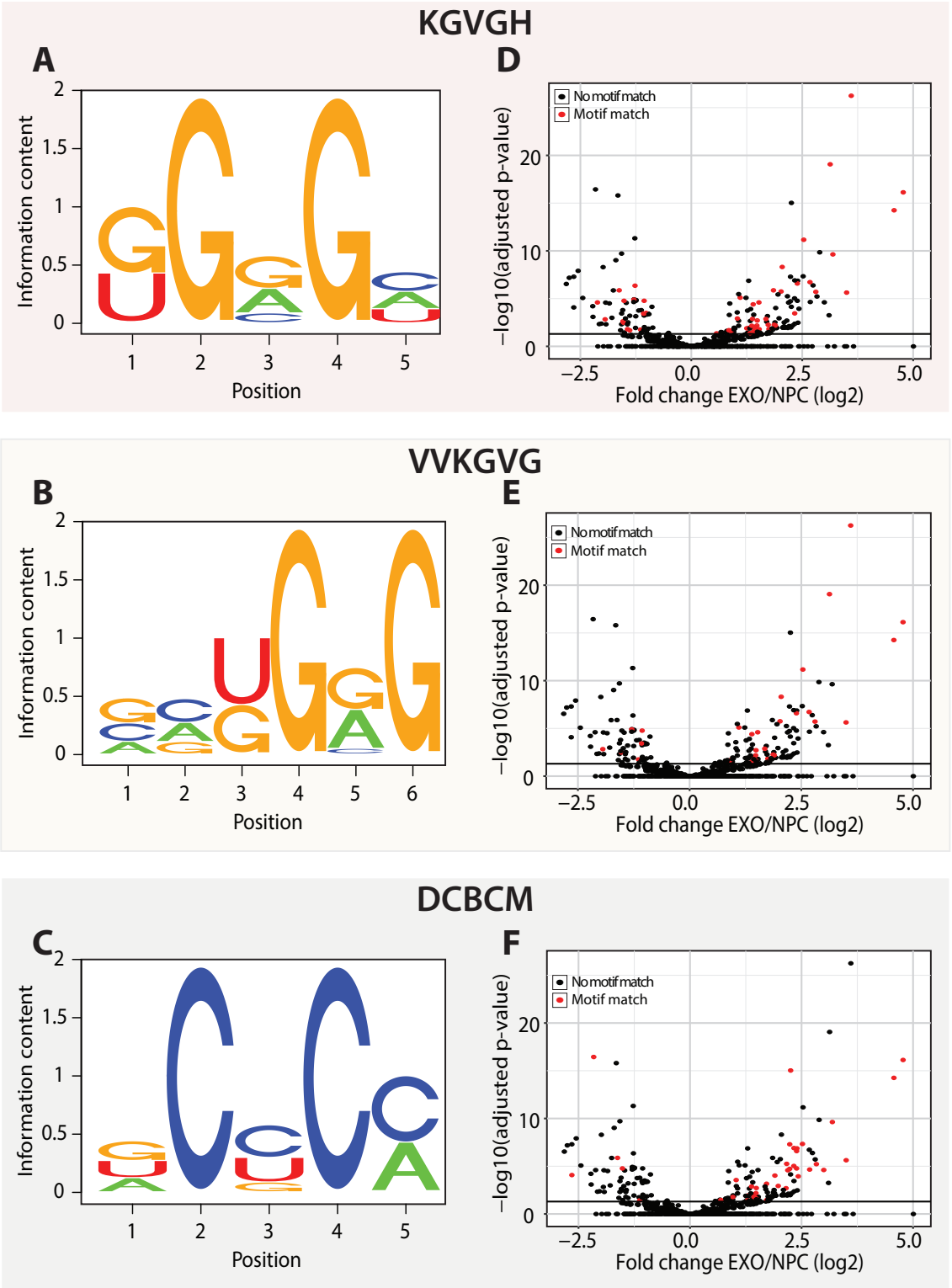


Figure 5.11 *A-C: Sequence logos of the three top scoring motifs identified by BCRANK as enriched in the sequence of secreted miRNAs.* ***D-F:*** *Volcano plots showing the $\log_2$ fold change of miRNAs expression in EXO vs NPC (x-axis) plotted against the -$\log_{10}$ of the p-value (y-axis). The red points indicate miRNAs that contain a match for the motifs KGVGH (**D**), VVKGVG (**E**) and DCBCM (**F**).*

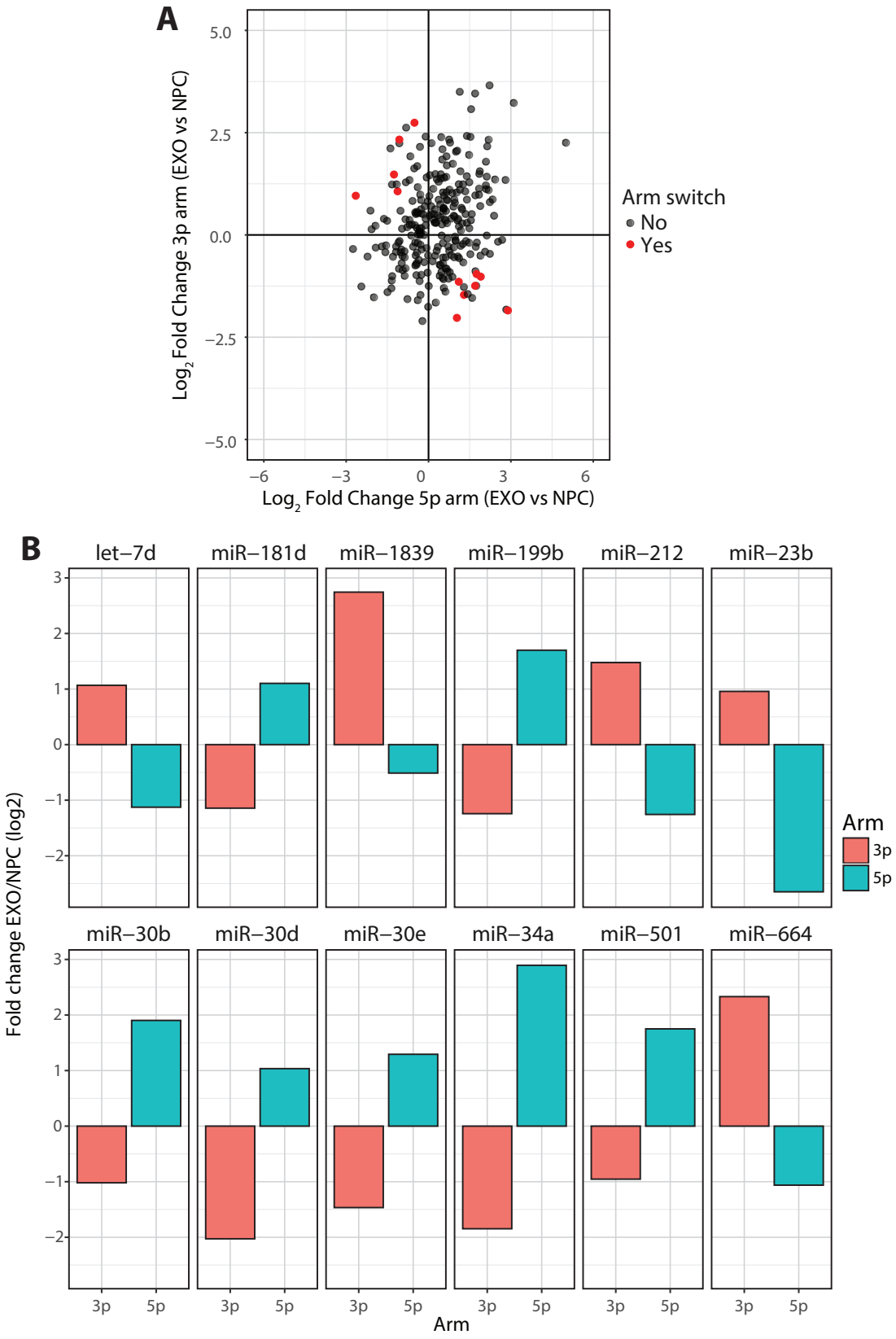


Figure 5.12 *A: Scatter plot showing for each miRNA the fold change (log_2) in EXOs vs NPCs of its 5p arm (x-axis) and 3p arm (y-axis). The red points represent miRNAs that undergo arm switch, i.e. miRNAs for which both arms have a statistically significant ($p<0.05$) fold change.* ***B****: Bar chart showing the fold change for the 5p and 3p arms of miRNAs that undergo arm switch.*

	Consensus	Score
1	KGVGH	142.39
2	VVKGVG	140.83
3	DCBCM	135.63
4	HNSDGRG	135.54
5	DCBCM	132.17
6	HNSDGRG	132.08
7	HBYAWD	129.01
8	HDRTH	128.78
9	MDRTHD	126.44
10	HDRTH	125.51

Table 5.1 *Ten highest scoring motifs identified by BCRANK*

Considering that the 5p and 3p arms of the same miRNA are transcribed at the same level, any difference in their abundance in cells or exosomes is necessarily the result of post-transcriptional processes. We therefore reasoned that their differential secretion in EXOs and NPCs could result from differences in the 5p and 3p sequences of these miRNAs, which might facilitate the secretion of one arm while retaining the other in the cell. For these reasons, we applied BCRANK to the sequences of arm-switching miRNAs with the purpose of identifying enriched short sequence motifs.

	FC (5p)	p-value (5p)	FC (3p)	p-value (3p)
mmu-miR-30d	1.035	2.75×10^{-4}	−2.028	4.16×10^{-3}
mmu-miR-34a	2.894	1.39×10^{-10}	−1.847	3.13×10^{-5}
mmu-miR-30e	1.293	1.34×10^{-7}	−1.466	2.75×10^{-5}
mmu-miR-199b	1.699	4.87×10^{-2}	−1.243	1.42×10^{-5}
mmu-miR-181d	1.102	7.91×10^{-6}	−1.143	1.77×10^{-2}
mmu-miR-30b	1.903	9.46×10^{-5}	−1.020	3.10×10^{-4}
mmu-miR-501	1.750	1.48×10^{-2}	−0.954	4.21×10^{-4}
mmu-miR-23b	−2.649	5.02×10^{-8}	0.958	3.11×10^{-2}
mmu-let-7d	−1.127	1.61×10^{-4}	1.068	6.95×10^{-3}
mmu-miR-212	−1.257	1.16×10^{-2}	1.477	3.83×10^{-2}
mmu-miR-664	−1.062	1.68×10^{-5}	2.331	1.23×10^{-7}
mmu-miR-1839	−0.512	3.88×10^{-2}	2.744	4.08×10^{-7}

Table 5.2 *Table showing miRNAs that undergo arm switch between EXOs and NPCs. Fold changes are expressed as log_2 and are relative to EXO vs NPC.*

The arm-switching miRNAs in Table 5.2 (in total 24 mature sequences) were sorted according to their fold change in EXOs vs NPCs. We then run BCRANK with a starting length of 3 (see Methods, section 10.5). We observed that the top motifs identified (Table 5.3, **Figure 5.13A-D**) were similar to those obtained in the analysis of all secreted miRNAs (**Figure 5.11A-C**), and contained either a strong G-rich or a C-rich motif. These results confirm the validity of

	Consensus	Score
1	KTRWHD	88.19
2	KYKSHT	87.14
3	VNWCVM	86.96
4	VYCYHHV	86.62
5	BSDKNGT	86.22

Table 5.3 *Highest scoring motifs identified by BCRANK in EXO-enriched miRNAs that undergo arm switch*

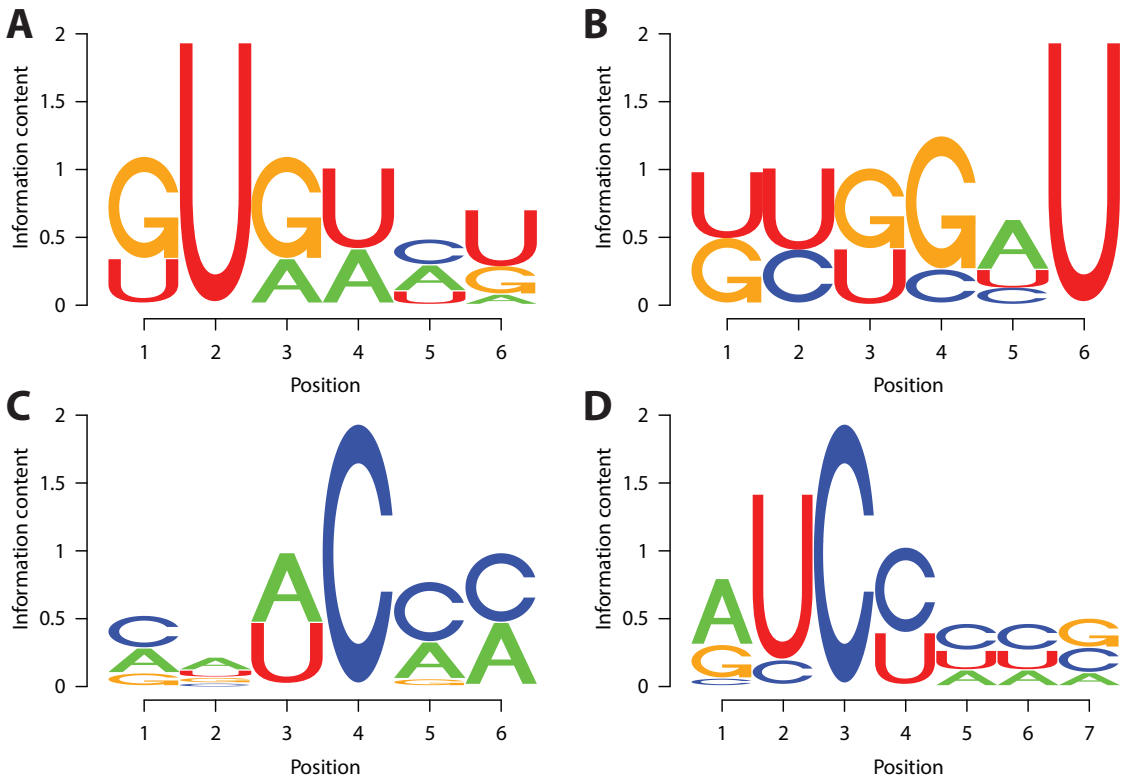


Figure 5.13 A-D*: Sequence logo of the four highest scoring motifs identified by BCRANK in EXO-enriched miRNAs that undergo arm switch.*

the previously identify motifs as candidate binding sites for molecular carriers of exosomal miRNAs.

5.4.3 *Analysis of GC content bias in secreted miRNAs*

We thought that the observed enrichment of motifs rich in Guanosines and Cytidines in secreted miRNAs might reflect a more general GC bias (either of technical or biological origin) rather than being the consequence of the enrichment of specific motifs. To verify this hypothesis we calculated the GC content of each miRNA and verified whether it has an influence on miRNA expression and/or miRNA secretion.

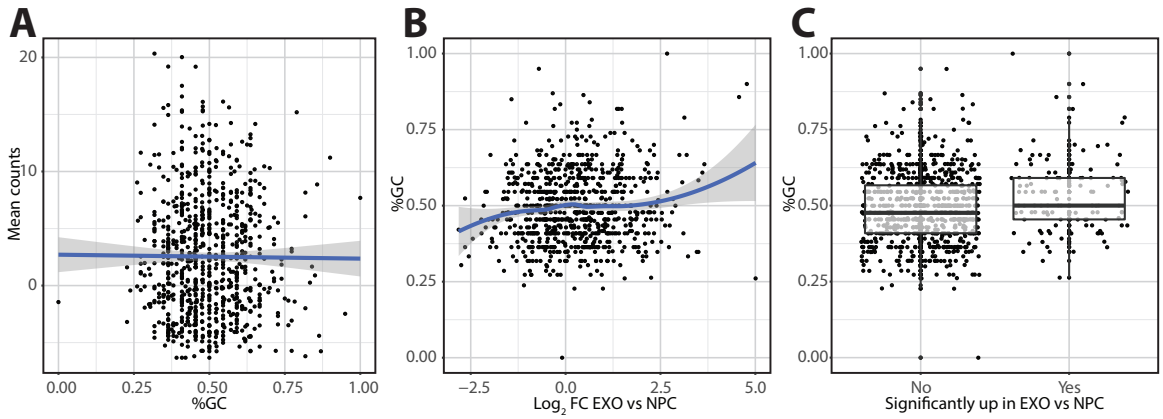


Figure 5.14 ***A****: Scatter plot showing the percentage of Guanosines and Cytidines (x-axis) in the sequence of miRNAs plotted against their mean counts in NPCs and EXOs (y-axis). The blue lines shows the linear fit of the data with the corresponding standard error (shaded area).* ***B****: Scatter plot showing the log_2 fold change (x-axis) in EXOs vs NPCs of miRNAs plotted against the percentage of Guanosines and Cytidines in their sequence (y-axis).* ***C****: Boxplot and overlaid scatterplot showing the distribution of the percentage of Guanosines and Cytidines for miRNAs significantly secreted or retained.*

We observed a negligible but significant effect of the GC content of each miRNA on its mean expression (**Figure 5.14A**, Pearson correlation coefficient = 0.005, p-value = 0.041). Similarly, we found a very modest but significant effect of GC content on the secretion coefficient of each miRNA (**Figure 5.14B,C**, Pearson correlation coefficient = 0.018, p-value = 1.249×10^{-4}). These data indicate that the GC content of a miRNA alone has a negligible effect on its secretion, suggesting that the motifs identified do not a reflect a generic GC bias.

5.4.4 *Refinement of secretion motifs*

The motifs identified by BCRANK as enriched in exosomal miRNAs show a preference for motifs containing either a CC region or a KG (G/T G) region. However, the enrichment analysis is highly dependent on the parameters used to filter the list of input sequences (namely the expression and p-value thresholds). To overcome this limitation and gain more confidence in the predicted motifs we therefore sought to optimise the two motifs identified using a number of input lists produced by varying the filtering thresholds.

To this end, instead of using BCRANK to search all possible motifs, we restricted its search space to motifs containing either the CC seed or the KG seed, then extending them and scoring each extended motif with 1000 reorderings of the input list of miRNAs. This procedure was repeated multiple times, providing as input a list obtained by varying the filtering thresholds as outlined in Table 5.4.

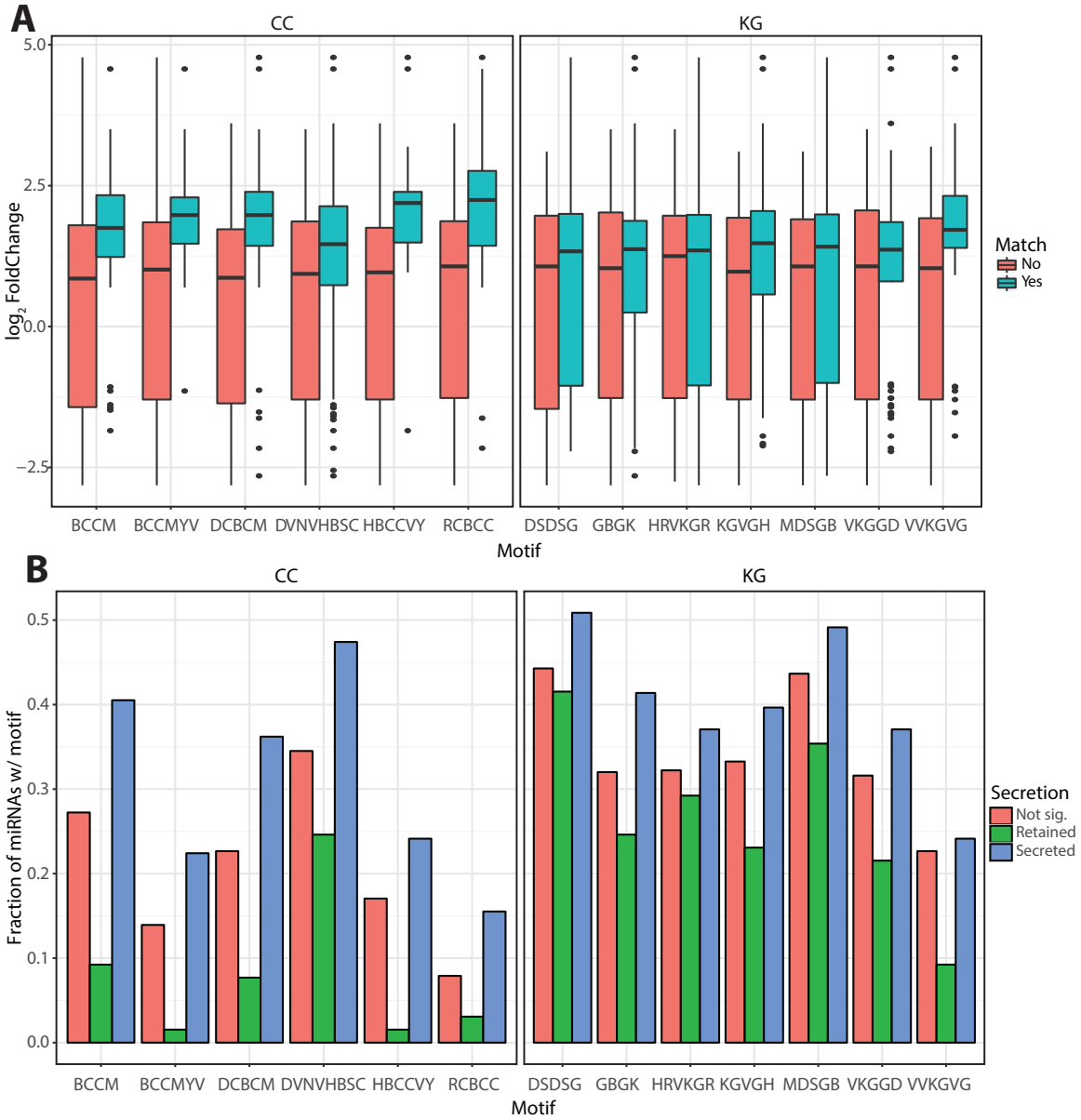


Figure 5.15 *A: Boxplot showing the distribution of log_2 fold changes in EXOs vs NPCs for miRNAs match (green) or do not match (red) the refined motifs obtained by BCRANK.* ***B****: Bar chart showing for each motif (x-axis) the fraction of miRNAs with a match that are significantly secreted (blue), significantly retained (green) or with a non-significant fold change in EXOs vs NPCs (red).*

Expression thr.	p-value thr.
0.2	0.01
0.2	0.05
0.5	0.01
0.5	0.05
0.5	1
0.75	0.01
0.75	0.05
0.75	1

Table 5.4 *Table showing the thresholds used to produce filtered lists of miRNAs for the motif optimisation with BCRANK. We selected miRNAs with expression above the specified quantile of the distribution of expressions and p-value in EXO vs NPC lower than the specified threshold. The list was then ordered according to the fold change in EXOs vs NPCs and provided as input for BCRANK.*

	Base	Motif	Stable	Retained	Secreted	Ratio
1	CC	BCCM	131	6	47	0.89
2	CC	BCCMYV	67	1	26	0.96
3	CC	DCBCM	109	5	42	0.89
4	CC	DVNVHBSC	166	16	55	0.77
5	CC	HBCCVY	82	1	28	0.97
6	CC	RCBCC	38	2	18	0.90
7	KG	DSDSG	213	27	59	0.69
8	KG	GBGK	154	16	48	0.75
9	KG	HRVKGR	155	19	43	0.69
10	KG	KGVGH	160	15	46	0.75
11	KG	MDSGB	210	23	57	0.71
12	KG	VKGGD	152	14	43	0.75
13	KG	VVKGVG	109	6	28	0.82

Table 5.5 *Table of refined motifs with the number of Stable (p-value>0.05 in EXOs vs NPCs), Retained (log_2 fold change<0 in EXOs vs NPCs and p-value<0.05) and Secreted (log_2 fold change>0 in EXOs vs NPCs and p-value<0.05) miRNAs that match each motif. The ratio column reports the number of Secreted/(Secreted + Retained) miRNAs.*

In total we obtained six refined motifs for the CC seed and seven refined motifs for the KG seed (Table 5.5). We observed that the miRNAs that contain one of the identified CC motifs have on average a higher secretion fold change (EXOs vs NPCs) compared to those that do not contain the motif (**Figure 5.15A**). In particular, the motifs `HBCCVY` and `RCBCC` give the highest median fold changes; however a smaller number of miRNAs had matches for these motifs compared to the other CC motifs (**Figure 5.15B**). On the other hand, a higher number of EXO-enriched miRNAs have a match for the motifs `BCCM` and `DCBCM` (Table 5.5), and these two motifs also display a higher ratio of secreted vs retained miRNAs (**Figure 5.16A**).

Compared to the CC motifs, miRNAs containing the KG motifs tend to have lower fold changes in EXOs vs NPCs and the number of EXO depleted

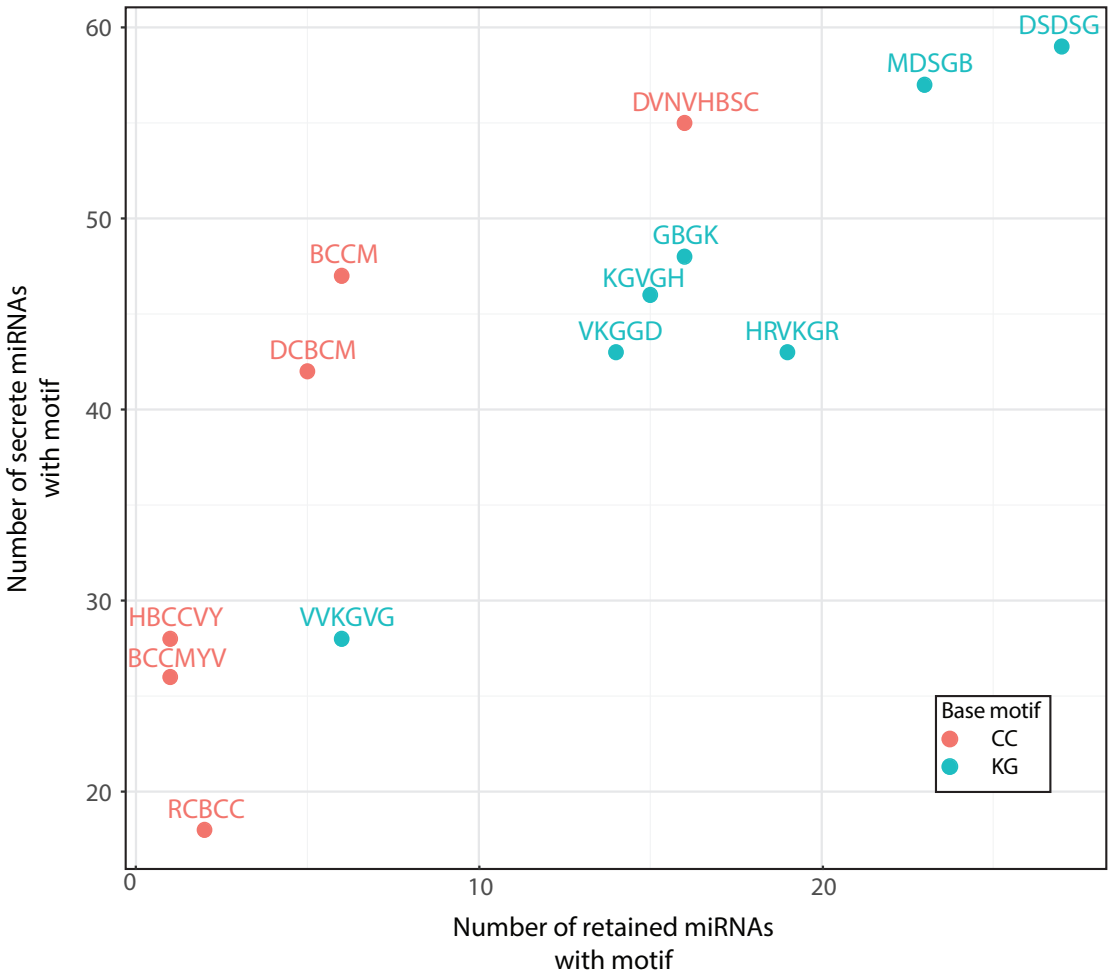


Figure 5.16 A: Scatter plot showing for each refined motif the number of retained miRNAs that contain the motif (x-axis) plotted against the number of secreted miRNAs that contain the motif (y-axis).

(down) miRNAs with matches to these KG motifs is higher compared to the CC motifs (Table 5.5 and **Figure 5.15A,B Figure 5.16A**). Overall, the motifs `GBGK`, `KGVGH` and `VKGGD` show the best compromise between sensitivity and specificity.

We selected as best motifs: `BCCM` for the CC seed and `GBGK` for the KG seed, based on the observation that they have the highest ratio between the number of secreted vs retained miRNAs that match the motifs (**Figure 5.16A** and **Figure 5.17A-C**). These results are in line with the observations reported by Villarroya-Beltri et al. (2013), which identify two similar motifs in human primary T lymphoblasts as binding sites for hnRNPA2B1.

Interestingly, a preliminary unreplicated RNA immunoprecipitation experiment showed that hnRNPA2B1 is present in NPC-derived EVs (**Figure 5.18A**) and it has a greater binding affinity for the secreted miRNA miR-361-5p - which contains the `BCCM` secretion motif - compared to miR-32-5p, which instead is not secreted and does not contain the motif (**Figure 5.18B**). Taken together, these data suggest that in murine NPCs hnRNPA2B1 might act as a carrier for secreted miRNAs.

In this chapter we provided a characterisation of the population of miRNAs secreted by NPCs into EVs and exosomes, finding that a subset of miRNAs is selectively enriched in EXOs and EVs. The analysis of the promoter sequences of secreted miRNAs revealed that they are unlikely to be transcriptionally co-regulated. However, by analysing their mature sequences we identified several short RNA motifs that might act as binding sites for carrier proteins. Indeed, it was recently reported by others that the protein hnRNPA2B1 binds these motifs and acts as a molecular carrier for miRNA secretion in human T lymphocytes. Our data confirm these findings and extend this mode of regulation to murine stem cells. These results will be discussed in detail in Chapter 8.

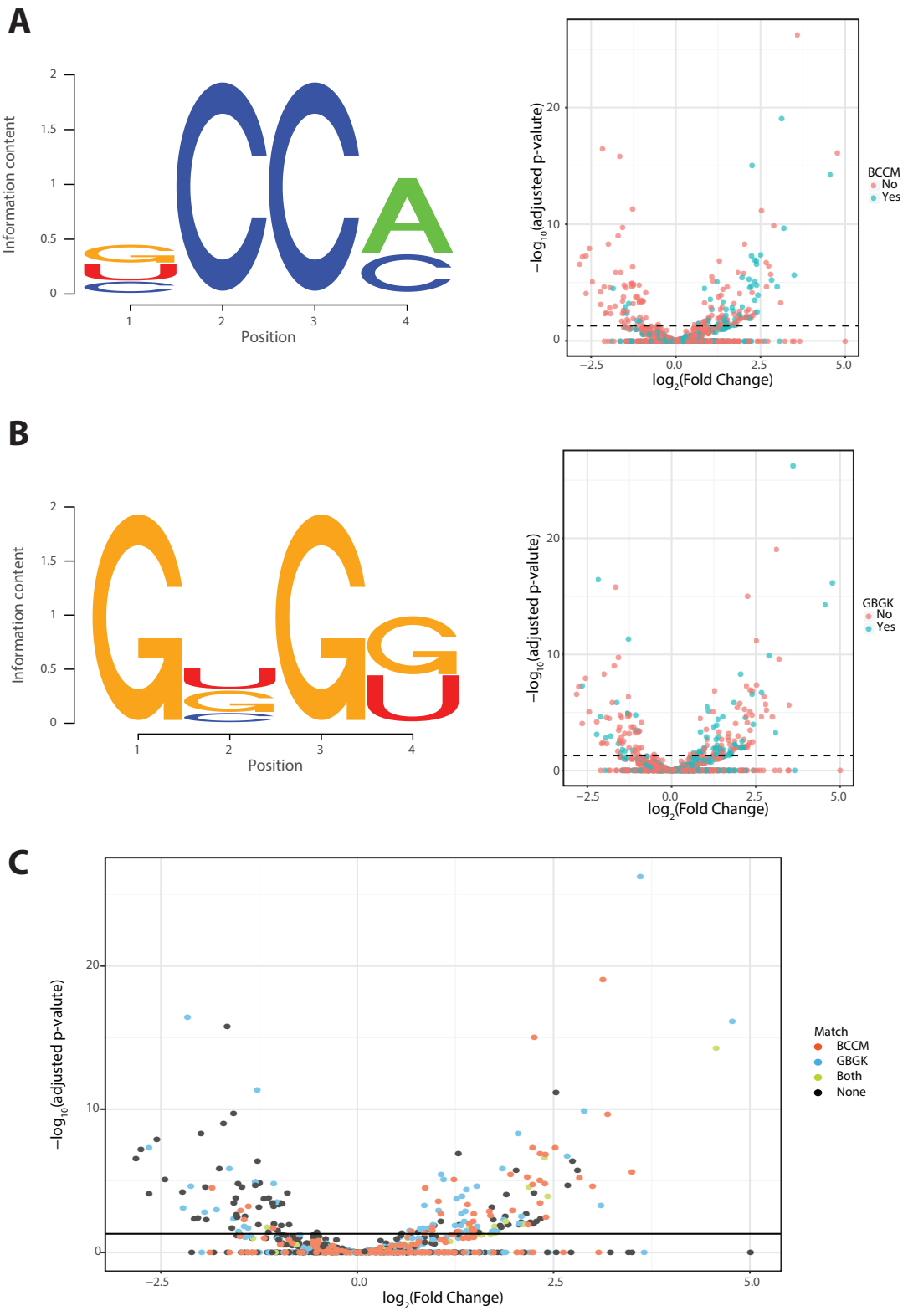


***Figure 5.17 A,B**: Sequence logos of the BCCM (**A**) and GBGK (**B**) motifs (left) and volcano plots (right) showing the log_2 fold change in EXOs vs NPCs plotted against the -log_{10} of their p-value. The blue points indicate the miRNAs that match the BCCM or GBGK motifs. **C**: Volcano plot showing the miRNA log_2 fold change in EXOs vs NPCs plotted against the -log_{10} of their p-value. The colours indicate whether a given miRNA matches the BCCM motif, the GBGK motif, both or neither motifs.*

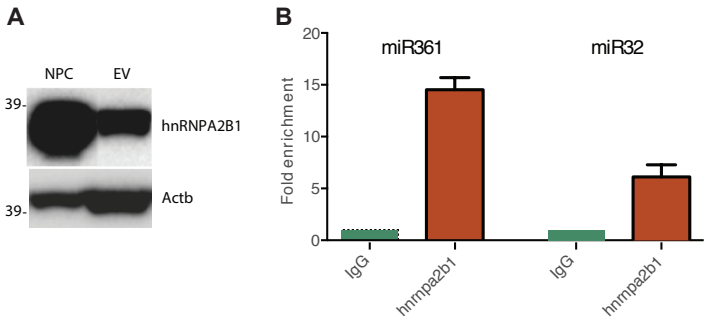


Figure 5.18 *A: Western Blot analysis for hnRNPA2B1 in NPCs and EVs.* ***B****: RNA Immunoprecipitation experiment for hnRNPA2B1, measuring the levels of miR-361-5p and miR-32-5p. The data was normalised over IgG immunoprecipitate. Experiment performed by Dr Iraci in the laboratory of Dr Pluchino.*

Part IV

Discussion

6

Positional conservation identifies topological anchor point (tap)RNAs linked to developmental loci

In this work we have identified and catalogued on a genome-wide scale human and mouse lncRNAs that display positional conservation between the two species. We defined pcRNAs based on their conserved genomic locations relative to neighbouring protein coding genes and the presence of a conserved promoter. Several earlier works described that many lncRNAs possess conserved promoters (Carninci et al., 2005; Guttman et al., 2009; Derrien et al., 2012; Necsulea et al., 2014) and are located in syntenic genomic locations (Carninci et al., 2005; Engstrom et al., 2006; Lipovich et al., 2006; Dinger et al., 2008b; Ulitsky et al., 2011; Necsulea et al., 2014; Hezroni et al., 2015). By systematically applying these two criteria, we were able to identify a set of 665 conserved lncRNA promoters that produce 1700 positionally conserved lncRNAs. We found that the set of 626 neighbouring protein coding genes associated with pcRNAs is strongly enriched in developmental transcription factors, a remarkable fraction of which have well established roles in processes such as neural development (e.g. *SOX2*), endoderm differentiation (e.g. *FOXA2*), cranio-caudal segment identity (e.g. *HOXA1*, *A2*, *A3*, *A11*, *A13, B3, C5* and *D8*) or eye development (e.g. *SIX3*). We observed that pcRNAs display conserved expression patterns in human and mouse somatic tissues and cell lines, and their expression often correlates with that of the neighbouring protein coding genes. The analysis of transcription factor binding profiles in the promoters of pcRNAs and associated genes revealed that they are often regulated by similar sets of transcription factors, providing a possible explanation for the observed co-expression. We also demonstrated that several pcRNAs are co-induced with the associated coding genes when cells are differentiated in the presence of morphogens.

Taken together, these data suggested a conserved functional connection between pcRNAs and associated coding genes, leading to the hypothesis that pcRNAs might regulate *in cis* the neighbouring genes, as previously demonstrated for specific cases such as *WT1-AS* (Dallosso et al., 2007), *EVF-2* (Feng et al., 2006), *SOX2OT* (Amaral et al., 2009; Askarian-Amiri et al., 2014), *HOTTIP* (Wang et al., 2011) and *EVX1AS* (Luo et al., 2016; Bell et al., 2016). The functional analysis of the pcRNAs *FOXA2-DS-S*, *POU3F3-BT*, *NR2F1-BT* and *HNF1-BT1* revealed that in all cases tested the knock down of the coding gene

abrogates the expression of the pcRNA, and in the case of *FOXA2-DS-S*, this effect can be explained by the presence of FOXA2 binding peaks in the promoter of the pcRNA. We also found that in all cases tested the knock down of the pcRNA reduces the expression of the neighbouring protein coding gene. In the case of *FOXA2* the reciprocal relationship between coding gene and pcRNA extends genome-wide, as the majority of the downstream effects of knocking down the pcRNA were largely overlapping with those of FOXA2 knock down. These data demonstrate that some pcRNAs positively regulate the neighbouring coding genes, and suggest that they do so acting *in cis*. This hypothesis is corroborated by the observation that the ectopic overexpression of a pcRNA had no effect on the neighbouring coding gene, which is in line with the idea that the context where a pcRNA is expressed is important for its function. These results are strengthened by preliminary data showing that transcriptional silencing of the pcRNA *TBX2-BIDIR* by CRISPR interference (CRISPRi) causes a reduction in the expression of *TBX2*. This orthogonal technique confirms that the effect observed on the coding gene is independent of small RNA-directed transcriptional gene silencing. It is possible that at least for some pcRNAs the *cis* effect on the coding gene, rather than being mediated by the lncRNA itself, depends on RNA-independent processes, such as the act of transcription or enhancer-like functions of their promoters, as recently shown for some lncRNAs (Paralkar et al., 2016; Engreitz et al., 2016). However, the results of the knock-down and CRISPRi experiments argue against this hypothesis, suggesting that pcRNAs have direct, RNA-mediated effects on the neighbouring coding genes. In an ongoing CRISPR display experiment, we are trying to rescue the knock down of a pcRNA by overexpressing and tethering it to its genomic locus. These data will allow us to prove that pcRNAs have direct, RNA-mediated functions *in cis*.

Our results show that this mode of conserved auto-regulation has preferentially evolved for a small number of developmental transcription factors, suggesting that regulation *in cis* by lncRNAs might provide some advantages compared to other modes of regulation – at least for genes such as developmental transcription factors that require to be expressed in a spatially and temporally regulated manner. *Cis*-acting lncRNAs are particularly suited for this function, as recently proposed by Quinn and Chang (2016). While the classical mechanism of auto-regulation requires a protein to be transcribed, exported to the cytoplasm, translated and then reimported in the nucleus before it can regulate its own locus, RNAs transcribed from the locus itself (either ncRNAs or coding RNAs with regulatory functions) are genomically, genetically and physically associated with the locus, and are therefore in the ideal position to evolve regulatory functions.

Although our data suggest that pcRNAs have a positive *cis* effect on the neighbouring genes, we can't exclude that some of them act both *in trans* and *in cis* or exclusively *in trans*. In fact, we found that ~20 % of the genes differentially expressed in *FOXA2-DS-S* knock-down do not change significantly when the coding gene is knocked-down, suggesting that in addition to its *cis* effects this pcRNA might also have *trans*-acting roles. In support of this scenario, there is evidence showing that some lncRNAs possess both modes of action. For example, the *WRAP53* antisense transcript regulates *in cis P53* expression, but it also acts as an mRNA encoding a protein with functions unrelated to P53 (Mahmoudi et al., 2010; Mahmoudi et al., 2009). On the other hand, the idea that pcRNAs act exclusively *in trans* is harder to reconcile with their positional conservation, although it is possible to speculate that the selective pressure for maintaining the genomic association between a coding gene and a *trans*-acting lncRNA might be due to their convergent functions (e.g. *LINC00598* and *FOXO1*, Jeong et al., 2016) and consequent need of co-regulation. Alternatively, the selective pressure for maintaining positional conservation might arise from a regulatory function exerted *in cis* by other elements of the coding locus on the lncRNA.

One of the most interesting characteristics of pcRNAs is that a large number of them overlaps chromatin loop anchor points, and we have defined this group as topological anchor point RNAs (tapRNAs). Loop anchor points are defined as two distant loci with high interaction probability, and recent works have shown that they are often bound by CTCF (Rao et al., 2014; Tang et al., 2015). Interestingly, the majority of these CTCF sites occur in convergent orientation (Rao et al., 2014), an observation that led to the hypothesis that loops are formed by the extrusion of chromatin through a complex consisting of cohesin and CTCF (Sanborn et al., 2015). There is also increasing evidence showing that several lncRNAs are involved in the formation of higher order chromatin structure, as in the case of the tapRNAs *HOTTIP* and *EVX1-AS* (Wang et al., 2011; Luo et al., 2016; Bell et al., 2016). We show that the promoters of tapRNAs are enriched in CTCF binding sites, and the cumulative position of the loop anchor points within the promoters precisely peaks at their Transcriptional Start Sites (TSSs). In a recent preprint article, Harmston et al. (2016) described that a subgroup of evolutionary ancient Topologically Associating Domains (TADs) is associated with a high degree of non-coding sequence conservation and lies in the proximity of developmental genes, suggesting a remarkable parallelism with what we describe for tapRNA loci. These observations suggest that tapRNAs might be involved in regulating the formation or function of such loops, and mediate the recruitment of distal regulatory sites to the tapRNA/coding gene loci. In accordance with this hypothesis,

we report that the distal sites that are in contact with tapRNAs through the loops are enriched in enhancers. Mechanistically, tapRNAs might act as a scaffold recruiting proteins that mediate the formation of the loops, or they might directly bind the distal regulatory regions through DNA-RNA interactions or RNA-RNA interactions with transcripts produced at the enhancer. We found that the sequence of tapRNAs is enriched in binding motifs for Zinc Finger transcription factors, and this is paralleled by an enrichment of the same motifs in the distal enhancers that are in contact with tapRNA loci. Several Zinc Finger factors have the capacity to bind both DNA and RNA (Brown, 2005), suggesting that tapRNAs might act as scaffolds that recruit or sequester transcription factors at enhancers, therefore increasing or buffering their local concentration. All these proposed mechanisms have a realistic molecular foundation (Arner et al., 2015; Kim et al., 2015b; Mondal et al., 2015; Pnueli et al., 2015; Hacisuleyman et al., 2014; Sigova et al., 2015), however our current data do not allow us to confidently say whether tapRNAs play a role in these contexts.

In conclusion, in this work we have used the positional conservation of lncRNAs as an indicator of their common functionality between species. This approach allowed us to identify a small set of syntenic lncRNAs with conserved promoters that are located in proximity of developmental transcription factors. We further show that the majority of these pcRNAs are associated with loop anchor points and positively regulate the expression of neighbouring coding genes. The observation that certain genes are regulated *in cis* by transcribed elements in their proximity argues in favour of a conceptual framework where nearby non-coding transcripts are considered part of an "extended gene" structure that encompasses a gene and the *cis* regulatory elements in its proximity. Future biochemical studies will allow us to dissect these extended genes and characterise the role of tapRNAs in the establishment and maintenance of a gene's higher order topological organisation.

7

Extracellular vesicles from Neural Stem Cells transfer IFN-ɣ via Ifngr1 to activate Stat1 signalling in target cells

In recent years it has become clear that inflammation plays a central role in a number of neurological disorders. Conditions such as stroke, multiple sclerosis or traumatic brain injury are all characterised by the presence of neuroinflammation, which is triggered by the activation of glial cells, such as astrocytes and microglia, and followed by the induction of a pro-inflammatory micro-environment that recruits immune cells from the circulation (Pocock and Liddle, 2001; Martino et al., 2011). This inflammatory microenvironment plays a double role: on the one hand, it contributes to further, secondary, damage in the brain parenchyma; on the other, it favours brain repair and healing (Martino et al., 2011). In pathological conditions, the fine balance between these two discording effects of inflammation is often altered, and the inflammatory environment in the brain plays a detrimental role. In this context, the transplantation of Neural Progenitor Cells (NPCs) has a remarkable, beneficial role, both from a clinical and pathological point of view (Pluchino et al., 2005). The mechanisms through which NPCs exert these beneficial effects are still unclear, however it is becoming increasingly apparent that they engage in a complex cross talk with cells of the immune system and modulate their activity, achieving the overarching effect of promoting neuroprotection (Martino et al., 2011; Ziv et al., 2006; Pluchino et al., 2005).

Among the possible routes through which NPCs exert their regulatory effects on the host's immune system, EVs appear particularly suited in light of their capacity to transport a broad range or molecules. The importance of EVs in this process is further stressed by the recent observation that the injection of EVs in mice affected by experimental autoimmune encephalomyelitis (a murine model of multiple sclerosis) induces a clinical and pathological recovery comparable to those animals transplanted with NPCs (Peruzzotti-Jametti and Stefano Pluchino, unpublished data).

In this work we have explored whether the secretion of EVs and exosomes constitutes a mechanism of cell-to-cell communication for NPCs. We have first characterised the physical and biochemical properties of the vesicles released in the culture supernatant, finding that NPC-derived exosomes display size distribution and surface markers compatible with previous descriptions in the literature. When we modelled *in vitro* an inflammatory environment

using Th1 cytokines, we observed, by RNA-Seq and Western blot, a remarkable and specific induction of genes and proteins belonging to inflammatory pathways, and in particular of those belonging to the IFN-ɣ response pathway. Similar analysis on EVs and exosomes revealed that Th1 cytokines, in addition to activating the IFN-ɣ pathway in NPCs, also induce the secretion of protein and mRNA components of this pathway in both EVs and exosomes. Although our data do not provide information on the mechanisms behind the secretion of these mRNAs and proteins, it is likely that their upregulation in the cell facilitates their passive diffusion to the vesicle compartments. However, it is also possible that they are actively and selectively sorted toward exosomes and EVs, as has been shown for other mRNAs and proteins in different contexts (reviewed in Villarroya-Beltri et al., 2014).

The presence of STAT1 and other components of the IFN-ɣ pathway in EVs and exosomes suggests that they can transfer the activation of this pathway to other cells. To verify this hypothesis, we used microarrays and quantitative proteomics to measure the effects of EVs derived from Th1-stimulated NPCs on target cells. These assays revealed that Th1 EVs induce in target cells effects very similar to those induced by Th1 cytokines on NPCs, namely the upregulation of protein and mRNA components of IFN-ɣ pathway such as STAT1, IGTP and B2M. Surprisingly, the effect of Th1 EVs on target cells was more pronounced than that of Th1 exosomes, either suggesting that the effect is associated with the non-exosome fraction of EVs or reflecting a bias introduced by the different purification protocols for exosomes and EVs.

The effect of Th1 EVs on target cells might be mediated by two non-exclusive mechanisms: the direct transfer of mRNAs and proteins from NPCs to target cells, as previously demonstrated by others in alternative systems (Valadi et al., 2007; Li et al., 2013a), or the indirect activation of the pathway through the action of other molecules present in EVs. The analysis of *Stat1*$^{-/-}$ target cells revealed that EVs directly transfer *Stat1* mRNA, but this was not accompanied by a corresponding increase in STAT1 protein. These data indicate that endogenous STAT1 in the target cell is required for the activation of the IFN-ɣ pathway mediated by Th1 EVs. Studying the effects of EVs derived from *Ifgnr1*$^{-/-}$, *Ifngr2*$^{-/-}$ or *Stat1*$^{-/-}$ NPCs we were able to show that the EV-mediated activation of IFN-ɣ signalling in the target cells is caused by the direct transfer of IFN-ɣ bound to its receptor IFGNR1. Additionally, the observation that in the absence of endogenous IFGNR1 the target cells do not respond to Th1 EVs, indicates that the activation of IFN-ɣ signalling passes through the target's IFGNR1.

Taken together, these results demonstrate that in response to Th1 cytokines, NPCs release EVs that present on their surface the IFN-ɣ/IFGNR1 complex.

This allows EVs to shuttle IFN-ɣ to target cells, where it engages the target's IFGNR1 and activates the intracellular signalling cascade that leads to the activation and upregulation of STAT1. This represents a novel mechanism that can be exploited by NPCs to propagate at a distance the activation of a signalling pathway.

In our current work we have used NIH 3T3 fibroblasts as a model of target cell, but future studies will need to address whether this type of mechanism also takes place in other more relevant cell types, such as T lymphocytes and macrophages. Furthermore, using immune cells as a more physiological model of target cells will also allow us to better evaluate the biological impact of such a mechanism of cell-to-cell communication. If successful, these works will help to clarify the mechanisms of interaction between transplanted stem cells and the immune system, shedding light on the molecular basis for the therapeutic potential NPC transplantation.

From a translational perspective, EVs offer an exciting avenue for the development of cell-free therapeutics for neuroinflammatory disorders of the central nervous system.

8

Secretion mechanisms of extracellular microRNAs in Neural Stem Cells

In recent years various works have studied and extensively characterised the process of cell-to-cell miRNA transfer via EVs and its functions, finding it to be involved in numerous biological processes, such as inflammation, cancer, angiogenesis and immune-suppression (Mittelbrunn et al., 2011; Pegtel et al., 2010; Fabbri et al., 2012; Zhuang et al., 2012; Okoye et al., 2014). The majority of these works are concordant in the observation that certain miRNAs are enriched in exosomes compared to the parental cell, and the set of enriched miRNAs is often cell-type dependent and reactive to external stimuli (Mittelbrunn et al., 2011; Squadrito et al., 2014).

In the effort to explore the possible range of ncRNA functions, we sought to characterise the miRNA population associated with EVs and exosomes secreted by murine NPCs and to investigate the mechanisms that lead to the selective secretion of specific miRNAs. We found that miRNAs are the main class of small RNAs present in EVs and exosomes, and we also found that their secretion is remarkably stable in response to Th1 or Th2 cytokines. In different experimental settings, the secretion of miRNAs was shown to be reactive to external stimuli (Squadrito et al., 2014), and it is of particular interest that in our biological context cytokines do not alter the miRNA repertoire of vesicles. The high correlation in miRNA concentration between vesicles and cells suggests that the bulk of cellular miRNAs are able to passively diffuse to EVs and exosomes, but the stability to cytokine treatment suggests that precise regulatory mechanisms are also in play. In fact, when we compared miRNA abundance between cells and vesicles we could detect a set of miRNAs selectively enriched in exosomes and a smaller one selectively enriched in EVs.

We were able to verify by luciferase assays that transferred miRNAs are functional and capable of silencing an artificial target construct in an *in vitro* model of recipient cell. Future experiments will allow us to assess whether NPC-secreted miRNAs induce measurable changes in relevant and physiological cell types, such as T lymphocytes or macrophages.

These data showed that certain miRNAs are selectively enriched in EVs and exosomes, suggesting the existence of a dedicated machinery that promotes their secretion. We reasoned that such a mechanism might act on two non-exclusive levels: transcriptionally and post-transcriptionally.

Transcriptional control of secreted miRNAs

The stability of secreted miRNAs in response to cytokines suggests that they might be transcriptionally co-regulated by a set of transcription factors not responsive to cytokines. Additionally, we also reasoned that promoter elements might recruit specific protein factors at transcribed miRNA loci and influence miRNA processing or subcellular localisation. Such hypothesis finds support in the observation that in yeast the presence of Hsf1 binding sites in the promoters of certain genes is necessary and sufficient to induce diffuse cytoplasmic localisation and increased translation of their mRNAs during glucose starvation (Zid and O'Shea, 2014).

On one level, we aimed to determine whether the promoters of secreted miRNAs were enriched in binding sites for specific transcription factors that might regulate their expression; on another level, we also wanted to determine whether the promoters of secreted miRNAs were enriched in binding sites for factors that could mediate their localisation towards vesicles.

In order to verify this hypothesis we first needed to identify the genomic location of miRNA promoters in the murine genome. Several previous studies aimed at finding the location of miRNA promoters, but those available at the time of this analysis were either in different species or relied on a limited number of data sources (Saini et al., 2007; Corcoran et al., 2009; Marsico et al., 2013; Alexiou et al., 2009). Therefore, we decided to leverage data produced by the ENCODE project to systematically identify miRNA promoters in mouse.

We found that miRNA promoters tend to be 1.4kb in size and are preferentially located ~5kb upstream of the annotated miRNA precursor. These observations are in line with the previously observed features of human miRNA promoters (Saini et al., 2007; Marsico et al., 2013). We also found that the vast majority of promoters are marked by DNase Hypersensitivity Sites, and more than 50% of them are also marked by H3K4me3 in the cell types analysed. Interestingly, we also find that the promoters display a modest but significant conservation, supporting the validity of our identification pipeline.

When we analysed the promoters of secreted miRNAs we could not identify any enriched transcription factor binding motif. This negative result suggests that secreted miRNAs are unlikely to be transcriptionally co-regulated, at least under our experimental settings. However, the lack of enrichment might also be a consequence of the limited power of our analysis. First, only a small number of miRNAs is significantly enriched in EXOs or EVs vs NPCs. Second, the large number of possible motifs to test greatly reduces the statistical power of the enrichment analysis. An additional potential confounding factor might be related to the identification of miRNA promoters. Although we were able to ex-

perimentally validate H3K4me3 and CpG methylation levels for two miRNA promoters, we did not cross-validate our predictions on a set of known miRNA promoters. For this reason, it is possible that the set of promoters that we have annotated contains a number of false positives and/or false negatives which would further decrease the power in the motif enrichment analysis.

To overcome these limitations we are now in the process of refining the promoter identification pipeline by implementing a supervised machine learning approach in which we use RNA-Seq data from *Drosha*$^{-/-}$ and *Dgcr8*$^{-/-}$ human cell lines for training (Dhir et al., 2015; Macias et al., 2015), and mouse *Drosha*$^{-/-}$ data for cross-validation (Georgakilas et al., 2014). Although at an early stage, this approach will allow us to refine the promoter predictions and increase the power of the motif enrichment analysis.

Post transcriptional control of secreted miRNAs

An alternative regulatory mechanism for the secretion of miRNAs might act at the post-transcriptional level. To investigate this hypothesis we analysed the sequence of secreted miRNAs in search of enriched short motifs that could act as binding sites for carrier proteins. Such carrier proteins might, directly or indirectly, mediate the sorting of miRNAs to the Multi Vesicular Bodies (MVBs) with their subsequent secretion into Intraluminal vesicles (ILVs) and exosomes.

The initial motif enrichment analysis identified several enriched motifs with varying strength of enrichment. Interestingly, all of the top scoring motifs consisted of variations on a `CNC` or `GNG` theme, and all of them appeared at high frequency in secreted miRNAs while being under-represented in miRNAs retained in the cell.

To strengthen these findings we adopted a different and complementary approach, performing a motif enrichment analysis of secreted miRNAs that undergo arm switch between NPCs and exosomes. We first identified a set of 12 miRNAs that display arm switch, meaning that one arm of the mature miRNA is enriched in exosomes while the other arm is depleted. The 5p and 3p arm of a miRNA are transcribed at the same level, therefore their differential enrichment in exosomes necessarily results from post-transcriptional regulation. When we searched for short motifs enriched in this set of miRNAs, we again found a high occurrence of G rich and C rich motifs.

After excluding the presence of a GC bias in the data, which might arise as a consequence of RNA-Seq technologies, we optimised the motif enrichment strategy and identified a set of short motifs with strong enrichment in secreted miRNAs. These motifs were very similar and were predominantly rich

in cytosines or guanosines. A likely explanation for the multiplicity of motifs identified could be that a small G-rich or C-rich core constitutes a sufficient motif, whereas the nucleotides around it have a limited effect of motif recognition. An alternative explanation might also be that we only identify a small number of miRNAs enriched in exosomes, and the sample size might be too small for confidently identifying a single consensus motif. However, the high similarity between the motifs suggests that they might act as genuine binding sites for exosomal carrier proteins. Interestingly, while our work was in preparation, Villarroya-Beltri et al. (2013) reported the enrichment of two very similar motifs in exosomal miRNAs secreted by human T cells. Furthermore, it was shown that one of the two motifs was recognised by hnRNPA2B1, which acts as molecular carrier shuttling miRNAs toward exosomes. By Western blot analysis we could show that hnRNPA2B1 is present in murine NPC-derived exosomes and preliminary RNA immunoprecipitation experiments suggest that it preferentially binds secreted miRNAs.

Both the two best motifs identified in our analysis and the two motifs reported by Villarroya-Beltri et al. (2013) seem complementary with each other. This observation might suggest that hnRNPA2B1 recognises a double strand motif in the stem of the pre-miRNA. If this was the case, miRNA sorting toward exosomes would happen before pre-miRNA maturation and independently of miRNA strand selection, therefore leading to the enrichment of both motifs. This hypothesis, is in line with the recent observation that miRNA sorting toward exosomes happens at the level of pre-miRNAs (Melo et al., 2014), but it is in contrast with our finding that miRNAs that undergo arm switch contain the secretion motifs. An alternative model that better explains our data could be that the secretion machinery recognises and sorts the double strand miRNA-target complex. In support of this hypothesis, recent works highlight an association between the miRISC-target complex and GW-bodies (Lee et al., 2009; Gibbings et al., 2009). GW-bodies are in turn associated with exosome biogenesis, and localise at the surface of MVBs (Gibbings et al., 2009). Also, knock down of the GW-bodies protein GW182 seems to reduce miRNA secretion in exosomes (Yao et al., 2012). Taken together, these data indicate a strong connection between the process of miRISC target recognition and miRNA secretion. However, the dynamics of selective miRNA enrichment and their relationship with miRNA or pre-miRNA motifs are still unclear. Future work will allow us investigate the expression of miRNA targets in exosomes, shedding light on the pathways that control miRNA secretion in murine NPCs.

A further point of interest is the observation that mRNAs and miRNAs are not the only classes of RNAs enriched in exosomes, as it was recently found that they also contain vault RNAs, Y-RNAs and tRNA fragments (Nolte-'t Hoen

et al., 2012). Interestingly, in a recent work Sharma et al. (2016) reported that vesicles released by epididymal cells (epididymosomes) transfer tRNA fragments to maturing sperm cells. The repertoire of tRNA fragments contained within epididymosomes was shown to be influenced by the paternal diet, and after fertilisation tRNA fragments influence the embryonic expression of genes controlled by the long terminal repeats of the endogenous retrovirus MERVL (Sharma et al., 2016). Similar results were reported at the same time by Chen et al. (2016), whereas an earlier work by Cossetti et al. (2014b) showed in a mouse model that exosomes can mediate the transfer of EGFP RNA from xenografted human melanoma $EGFP^+$ cells to the germline of the host, showing for the first time that circulating exosomes are capable of reaching the testis, crossing the Weismann barrier and delivering their cargoes to sperm cells.

These data have potentially broad implications, because they provide evidence for a molecular mechanism underlying the soma-to-germline transmission of genetic information. In this context, exosomes represent an ideal medium for the transfer of genetic information from the soma to the germline, because 1) they transport complex populations of RNA; 2) their content can be actively modulated; 3) they protect the cargo from RNase degradation; 4) have the potential of targeting specific recipient cells types or organs. Taken together, these data suggest the exciting possibility that exosomes could act as a vehicle to transfer genetic information from somatic compartments – such as the brain, the liver or the immune system – to the germline, and mediate the acquisition of novel inheritable traits.

Part V

Conclusions

My doctoral studies covered a number of aspects pertaining to the broad field of non-coding RNA genomics. I worked on three collaborative projects were the bioinformatics analysis complemented and was complemented by extensive experimental work. The overall aim of these projects was to further our understanding of the function of ncRNAs and their roles in cell-to-cell communication.

We used positional conservation as an indicator of functional conservation between species, and thus characterised a class of long non-coding RNAs that regulate the expression of neighbouring protein coding genes. It emerged that a remarkable number of them is associated with chromatin loops, and we defined a new class of topological anchor point RNAs (tapRNAs). The study of chromatin topology has recently gained great interest in the scientific community, and it is becoming increasingly clear that long-range chromatin interactions play a central role in the organisation and regulation of the genome. In our work we show a link between chromatin topology, non-coding loci and regulation of gene expression, paving the way for future studies aimed at dissecting the complex interplay between these key aspects of genome regulation.

In the effort to explore the possible range of ncRNA functions, we studied the processes of cell-to-cell communication via the exchange of extracellular vesicles. We first studied the process of exosomes and EVs secretion by Neural Stem Cells, finding that they can transfer mRNAs and proteins. This study revealed a novel mechanism of cell-to-cell signalling based on the EV-mediated transfer of protein and mRNA components of signalling pathways. This work was the result of a close interaction between the wet and dry lab, where the bioinformatics generated a series of hypothesis that could then be experimentally confirmed. This project also provided interesting indications on the complexity of vesicle contents and their roles as effector molecules at a distance, and promoted us to investigate the role of ncRNAs in EV-mediated cell-to-cell communication.

In conclusion, my work addressed some unanswered questions on the function of ncRNAs, providing insights into their roles inside the cell, where they regulate protein-coding genes and chromatin topology, and outside the cell, where they exchange signals between distant compartments of an organism. Although addressing unrelated questions, these studies shared a similar functional approach that highlights the importance of using functionality as a criterion for ncRNA classification. The rapidly evolving field of ncRNA biology is shifting from the negative definition of ncRNAs based on lack of coding potential to definitions based on functional relatedness. Thanks to this paradigm shift, the molecular functions and biological roles of ncRNAs are starting to emerge in their full complexity.

Part VI

Methods

9

Positional conservation identifies topological anchor point (tap)RNAs linked to developmental loci

The methods for all the wet-lab experiments performed for this project by Dr Amaral, Dr Viré and Ms Büscher are reported online[1] and accessible at the following Digital Object Identifier (DOI): 10.1101/051052.

9.1 Human and Mouse reference genomes

The reference genomes for human (hg38) and mouse (mm10) were downloaded from the UCSC FTP server (Rosenbloom et al., 2015) in 2bit format and converted to fasta format using the twoBitToFa tool from the UCSC genome browser. The fasta files were indexed using samtools faidx (v1.2).

The Bowtie index for both genomes were built with bowtie2-build (v2.1.0) (Langmead and Salzberg, 2012).

9.2 Human and Mouse reference transcriptomes

The reference gencode transcriptomes for human and mouse (version 21 for human and version M4 for mouse) were obtained from the Gencode website in GTF format (Harrow et al., 2012).

9.3 Genbank all RNAs

The annotation of mouse Genbank mRNAs was obtained from the "Mouse mRNAs from GenBank" track of the UCSC genome browser using the Table Browser (Karolchik et al., 2004).

9.4 RNA-Sequencing data analysis

In order to obtain comprehensive transcriptomes for human and mouse as well as to quantify pcRNA abundance, we integrated the reference Gencode transcriptomes with RNA-Seq data on human and mouse tissues and cell lines. We used RNA-Seq from six matched human and mouse tissues (Brain, Cere-

[1] http://dx.doi.org/10.1101/051052

bellum, Heart, Kidney, Liver and Testis) as well as data produced by the ENCODE project from similar human and mouse cell lines (Supplementary Table 3.1, page 91).

Mapping

The RNA-Seq datasets were mapped to the reference human and mouse genomes (hg38 and mm10 respectively) using Tophat2 (v2.0.10, bowtie2 v2.1.0 Kim et al., 2013; Langmead and Salzberg, 2012) with the option `b2-sensitive` and the Gencode comprehensive GTF files as reference transcriptomes (v21 and M4 for human and mouse respectively). The reference transcriptomes were built with an independent Tophat run without fastq files and then provided to all subsequent mapping runs through the option:

`--transcriptome-index`

The `--library-type` option was set to "fr-unstranded" for unstranded datasets and "fr-firststrand" for stranded datasets.

For two very deep (>120mln reads each), single end 45nt reads mouse datasets (SRR549335 and SRR549339, see Supplementary Table 3.1, page 91) Tophat by default tried to identify splice junction by coverage search, but stopped at the stage "Searching for junctions via segment mapping" probably due to the very high number of reads. To overcome this problem, we disabled only for these two samples the coverage search functionality (`--no-coverage-search`) as suggested by Tophat's standard error.

Assembly

The transcriptomes were assembled independently for each RNA-Seq dataset using Cufflinks (v2.2.1) with the following options:

`--library-type` "fr-unstranded" for unstranded datasets and "fr-firststrand" for stranded datasets.

`-F 0.05`

`--multi-read-correct`

`--frag-bias-correct` pointing to the genome fasta file

`-M` a masking GTF files to exclude ribosomal transcripts and mitochondrial transcripts. This file was produced by selecting from the Gencode GTF files the lines that matched "Mt_" or "rRNA" in field 14.

`-g` exon-cds-filtered reference transcriptome GTF. This file was produced by selecting only exon and CDS features from the Gencode reference GTF files (field 3), therefore excluding the "gene" and "transcript" entries. Such a filtered GTF file contains all the information needed by Cufflinks and provides a significant speed up in cufflinks' running time.

The Cufflinks assembled transcriptomes for each sample were then merged using Cuffmerge (with the same exon-cds reference transcriptome used for Cufflinks) and converted to BED12+ format using the gtfToBed tool (Kent, 2002) with the option "`-a gene_id,oId,class_code`" to preserve Gene ID, Gencode ID and Cufflinks class codes as additional fields.

Abundance estimation and expression normalisation

The human and mouse merged transcriptomes (merged.gtf) were then quantified against each BAM file using Cuffquant (v2.2.1) with the following options:

`--library-type` (see 9.4)

`--multi-read-correct`

`--frag-bias-correct`

`-M` Reference masked regions (see 9.4)

Finally, the Cuffquant binary output files were normalised with Cuffnorm to produce the human and mouse expression matrices. Cuffnorm (v.2.2.1) was run with the following options:

`--output-format cuffdiff`

`--use-sample-sheet`

`--library-type fr-unstranded`

9.5 Identification of pcRNAs

Human Data preparation

The purpose of this data preparation step is to produce an annotation of reference and novel non-coding transcripts from which we will later identify pcRNAs.

1) Annotation of coding transcripts and CDS

From the Gencode BED annotation we selected transcripts containing an annotated CDS. We then used the getCoding tool of Pinstripe (Gascoigne et al., 2012) to obtain a BED annotation of only the coding portion (CDS) of each coding transcript.

2) Reference non coding RNAs

We filtered the Gencode V21 BED file in the following way:

1. We used awk to select all transcripts without an annotated ORF and composed of more than one exon.

2. We used overlapSelect (UCSC genome browser tool, Kent, 2002) to exclude all transcripts that had more than 20bp of sense overlap with the CDS region of a coding transcript.

3) Novel non coding RNAs

We filtered the merged RNA-Seq transcriptome BED file in the following way:

1. We used awk to remove single exon transcripts as well as transcripts that don't map to the primary assemblies of the autosomes or sex chromosomes.

2. We used overlapSelect to discard transcripts with more than 20bp of sense overlap with the CDS region of a coding transcript.

3. We used bedtools intersect (v2.24.0) to discard transcripts with more than 50% sense exonic overlap with reference non-coding transcripts (previous step).

4. We used Pinstripe dedup (version v1.0.4554.32000, with option exEncomp) to remove redundant transcripts.

5. We used CPAT (v1.2, Wang et al., 2013) to calculate the coding potential of each transcript and only retained transcripts with score <0.364 (see CPAT documentation for information on the threshold).

Finally, we combined *the Reference non coding RNA* annotation and the *Novel non coding RNAs* annotation and we used bedtools intersect to remove all transcripts with more than 50% sense exonic overlap with coding transcripts (although in the previous step we had already filtered out CDS-overlapping transcripts, this step ensures that we do not have transcripts that have more than 50% overlap with the UTR of coding genes).

The file that we obtain is a comprehensive annotation of all reference and novel human non-coding RNAs and we will hereafter refer to it as *know+novel ncRNAs*.

Mouse Data preparation

The purpose of this data preparation step is to produce an annotation of reference and novel non-coding transcripts from which we will later identify pcRNAs.

1) Annotation of coding transcripts and CDS

From the Gencode BED annotation we selected transcripts containing an annotated CDS. We then used the getCoding tool of Pinstripe to obtain a BED annotation of only the coding portion (CDS) of each coding transcript.

2) Reference non coding RNAs

We filtered the Gencode M4 BED file in the following way:

1. We used awk to select all transcripts without an annotated ORF and composed of more than one exon.
2. We used overlapSelect to exclude all transcripts that had more than 20bp of sense overlap with the CDS region of a coding transcript.

3) Novel non coding RNAs

We filtered the merged RNA-Seq transcriptome BED file in the following way:

1. We used awk to remove single exon transcripts as well as transcripts that don't map to the primary assemblies of the autosomes or sex chromosomes.
2. We used overlapSelect to discard transcripts with more than 20bp of sense overlap with the CDS region of a coding transcript.
3. We used bedtools intersect (to discard transcripts with more than 50% sense exonic overlap with reference transcripts (previous step).
4. We used Pinstripe dedup (with option --exEncomp) to remove redundant transcripts.
5. We used CPAT to calculate the coding potential of each transcript and only retained transcripts with score <0.44 (see CPAT documentation for information on the threshold).
6. We removed all transcripts with more than 50% sense exonic overlap with a coding transcript to remove UTR overlapping RNAs.

4) *Genbank non coding RNAs*

Given the lower number of lncRNAs annotated by Gencode in mouse (6951 in Gencode M4 vs 15877 in human Gencode V21), we also incorporated in our analysis Genbank non coding RNAs.

To identify them, we downloaded the "all mRNAs" GTF from the UCSC genome browser and processed it in the following way:

1. We used the gffread tool (v2.2.1, part of the Cufflinks suite) to exclude all transcripts with non-canonical splice sites (i.e. not GT-AG, GC-AG or AT-AC) and with introns shorter than 4nt, then we converted the filtered GTF file to BED with Pinstripe gtfToBed.

2. We retained only transcripts with more than one exon.

3. We discarded transcripts with more than 20bp of sense overlap with the CDS region of a coding transcript (overlapSelect).

4. We used CPAT to calculate the coding potential of each transcript and only retained those with score <0.44 (see CPAT documentation for information on the threshold).

5. We removed all transcripts with more than 50% sense exonic overlap with a coding transcript to remove UTR overlapping RNAs.

Identification of conserved promoters

For each transcript in the human *know+novel ncRNAs* annotation (see 9.5) we produced a BED file of their promoter regions by extending their TSSs of 500bp in each direction, then we obtained their FASTA sequence from the reference genome using the Pinstripe getDna tool.

In order to make a blast database of the mouse genome we first used the ncbi-blast convert2blastmask tool (v2.2.30+, options `-masking_algorithm repeat -masking_options "repeatmasker and tandem repeats from UCSC" -outfmt maskinfo_asn1_bin`) to extract masking information from the soft masked genome fasta file, and then the makeblastdb tool (v2.2.30+, `options -mask_data path-to-convert2blastmask --out -dbtype nucl`).

We then used the following command line to align human ncRNA promoters to the mouse genome with blast (v2.2.30+):

```
blastn -task blastn --db path/to/db -out path/to/out -query
path/to/promoters/fasta -outfmt 6 -evalue 0.001 -num_threads
n-processors -db_soft_mask 40 -lcase_masking
```

Finally, we processed the blastn output file with awk to only retain alignments longer 100nt and with E-value $<10^{-10}$ and convert the blast coordinates (1 based) into BED coordinates (0 based).

Non-coding to coding positional annotation

We next aimed to associate each ncRNA identified in human and mouse to its closest protein coding transcripts. To this end we used Pinstripe "closest", which returned – for each input ncRNA – the closest upstream, downstream and overlapping coding transcript. We then processed each entry and compared the non-coding and coding coordinates to annotate their TSS-to-TSS distance as well as the orientation of the non-coding relative to the coding in the following way:

- If the coding and non-coding intervals overlapped we defined the coding-non-coding pair as OLAP if on the same strand or AS if on different strands.
- If there was no overlap and the non-coding was upstream of the coding (relative to the strand of the coding), we defined the pair as US-S if coding and non-coding were on the same strand, otherwise US-AS if the TSS-to-TSS distance was >2000bp or BT if $\leq$ 2000.
- If there was no overlap and the non-coding was downstream of the coding (relative to the strand of the coding), we defined the pair as DS-S S if coding and non-coding were on the same strand, otherwise DS-AS.

We then matched each human and mouse coding transcript to their corresponding Ensembl Gene Ids, and for each non-coding/coding gene pair in a given orientation we only retained the closest coding transcript.

Human-mouse positional comparison

To identify mouse ncRNAs arising from conserved human ncRNA promoters we extended each region in the mouse genome that resulted from blasting human ncRNA promoter (see 9.5) of 500nt in each direction, and then we intersected these regions with the 5' exon of each mouse ncRNA.

This step allowed us to obtain pairs of human/mouse ncRNAs that have a conserved promoter. We then selected those pairs for which at least one coding neighbour of the human non-coding (where neighbouring means the closest upstream, downstream and overlapping as defined in the previous step) was orthologous to at least one coding neighbour of the mouse non-coding

To identify orthologous genes between mouse and human we programmatically downloaded from Ensembl Biomart (v80) a table that associates each human gene_id to the gene_id of the orthologous gene in mouse.

The resulting annotation contains human and mouse ncRNAs whose promoter is conserved and whose neighbouring gene(s) is (are) orthologous.

We further filtered this annotation by removing all human/mouse ncRNA pairs that were in opposite orientations relative to the coding genes in the two species (i.e. DS-S, US-S or OLAP in one species and AS, BT, DS-AS, US-AS in the other).

In numerous cases we could not univocally associate each ncRNA to a single coding gene, since the same ncRNA can have multiple neighbouring coding genes orthologous and in the same orientation in mouse and human. To resolve these ambiguities and univocally assign each ncRNA to a unique coding gene we applied the following criteria:

1) In case any of the possible coding genes were either AS or OLAP in human we retained the closest (TSS-to-TSS) of those.

2) In all other cases we retained the coding with shortest TSS-to-TSS distance in human.

Annotation of pcRNA genomic characteristics

To annotate pcRNAs that overlapped Gencode lncRNAs we intersected the human pcRNA annotation with the Gencode annotation of lncRNAs considering all exonic sense overlaps.

To annotate pcRNAs that overlapped miRNAs we queried the UCSC genome browser MySQL server for all transcripts containing the string "miR" in the geneName field.

To annotate pcRNA promoters we extended each pcRNA TSS by 2000bp in each direction and merged the resulting promoter regions that overlapped (bedtools mergeBed).

9.6 Characterisation of pcRNA features and expression analysis

To produce human and mouse expression matrices we matched the Ensemble Transcript IDs of human and mouse pcRNAs with the "oID" identifiers reported by Cuffmerge; we then used the corresponding Cuffmerge IDs to track pcRNAs in the isoforms FPKM tracking files reported by Cuffnorm.

For human and mouse coding genes we used a similar approach to extract the FPKM transcript information for all transcripts of each coding gene, and

then summed the FPKMs to obtain a single expression measure at the gene level.

For the expression analysis all FPKM values below 10^{-3} were set 0 and all transcripts with 0 FPKMs in all samples were excluded.

pcRNA expression heatmaps

The expression heatmaps for human and mouse pcRNAs were produced with the function heatmap.2 of the gplots package. The rows and columns were clustered with the default methods. For visualisation purposes in order to calculate the $\log_2$ of the FPKMs the smallest FPKM value was added to each value. The vertical sidebar reports the tissue specificity score calculated as indicate below.

pcRNA expression distance heatmaps

The heatmaps showing the Euclidean distance between pcRNA expression profiles have been realized by calculating the matrix of pairwise Euclidean distances between all pcRNAs using the *dist()* function in R. The heatmap was produced with the heatmap.2 function of the gplots package using the default methods for clustering rows and columns. The horizontal sidebar reports the tissue specificity score calculated as indicate below. The vertical sidebar reports the tissue where a given pcRNA has maximal expression.

GO enrichment of pcRNA-associated coding genes

The GO enrichment of pcRNA-associated genes was obtained using the TopGO package of Bioconductor. The ontology mapping used was provided by the package org.Hs.eg.db. The background set of coding genes consisted of all human protein coding genes with an annotated mouse ortholog. GO nodes with less than 10 annotated terms were excluded from the analysis. The p-values were calculated using the "default" method of TopGO and Fisher's Exact test. P-values were corrected using the Benjamini-Hochberg method as implemented in the *p.adjust(method="BH")* function in R. For the GO enrichments of pcRNA-associated genes divided by relative orientation, we used as background the set of all pcRNA-associated coding genes. P-values were calculated as described above but were not corrected for multiple hypothesis testing.

Correlation of expression between pcRNAs and coding genes and between human and mouse pcRNAs

To calculate the Spearman's rank correlation coefficients between human pcRNAs and coding genes we first calculated a matrix of coefficients between each pcRNA and each coding gene, where the diagonal represented the coefficients between each pcRNA and its associated coding gene.

To test whether the mean correlation coefficient was higher than expected by chance we performed a permutation test: we selected 10^6 samples of random coefficients from the entire matrix, and calculated how many times the mean of the random sample was higher or equal to the mean of the diagonal of the matrix. We reported a p-value$<10^6$ when none of the random samples' means was higher or equal to the mean of the diagonal.

The correlation coefficients were calculated in R with the function *cor()* using the Spearman method.

The correlation of expression between human and mouse pcRNAs was calculated in the same way. When the same human pcRNA was associated to multiple mouse pcRNAs we calculated the correlation between all pairs.

Tissue specificity score and GO enrichment by tissue

The tissue specificity score for human and mouse pcRNAs and coding genes was based on the square root of the Jensen-Shannon divergence as in (Cabili et al., 2011). The p-value for the difference between pcRNAs and coding genes was calculated with the Wilcoxon test as implemented in the wilcox.test function in R. To verify whether the difference of tissue specificity score between pcRNAs and coding genes was due to their different expression levels, we used the MatchIt R package (*method="subclass", subclass=5, sub.by="control"*) to subdivide pcRNAs and coding genes in 5 classes so that each class had similar distributions of maximal FPKM. We then calculated the tissue specificity score distribution for each of the 5 classes.

The GO enrichment of pcRNA by tissue of maximal expression was done by selecting the coding genes associated with pcRNAs with maximal expression in the given tissue and with a tissue specificity score above the mean of all specificity scores. The GO enrichment was performed in R using the TopGO package. The background set of coding genes consisted of all pcRNA-associated coding genes. GO nodes with less than 20 annotated terms were excluded from the analysis. The p-values were calculated using the "default" method of TopGO and Fisher's Exact test. P-values were not corrected for multiple hypothesis testing.

Human-mouse conservation analysis

To calculate the sequence conservation of human and mouse pcRNAs we employed the Needleman–Wunsch algorithm to align the human and mouse pcRNA sequences (Needleman and Wunsch, 1970). In case multiple mouse pcRNA isoforms were associated with the same human pcRNAs we performed all possible pairwise alignments and only retained those with the highest sequence identity. Similarly, to calculate the sequence conservation of pcRNA-associated protein coding genes we performed pairwise alignments (Needleman–Wunsch algorithm) between all transcripts of the human gene and all transcripts of the mouse gene, and retained the alignment with the highest sequence identity.

9.7 Nanostring analysis

For the nanostring experiment we designed probes to detect 50 pairs of pcRNAs and corresponding coding genes in human and mouse. The probes were designed according to the Nanostring guidelines and to maximize their specificity and included 9 house-keeping genes for normalization (*ALAS1, B2M, CLTC, GAPDH, GUSB, HPRT, PGK1, TDB, TUBB*).

The raw count data were first normalized by Nanostring Technologies with the nSolver software using a two-step protocol. First, data were normalized to internal positive controls, then to the geometric mean of house-keeping genes. The normalised data was then imported into R for further analysis. The correlation of expression between pcRNAs and coding genes was calculated with the *cor()* function in R after averaging replicate samples. To test whether the mean correlation coefficient between pcRNAs and coding gene as well as between human and mouse pcRNA pairs was higher than expected by chance, we used a permutation test as described for the RNA-Seq analysis.

To cluster pcRNAs and coding genes based on their expression profiles with first used the mcxarray tool of MCL (van Dongen, 2008) to produce a graph where nodes represented human pcRNAs and corresponding coding genes, and edges connected nodes with a Pearson correlation coefficients higher than 0.5. We then run MCL on such graph with the inflation parameter set to 3 to identify clusters of pcRNAs and coding genes.

9.8 FOXA2-DS-S knock-down microarray analysis

RNA samples were amplified using the TotalPrep 96-RNA amplification kit from Ambion (Applied Biosystems). Briefly, the RNA was converted into cDNA, and amplified by In Vitro Transcription (IVT) to generate biotin-labeled cRNA.

The cRNA was then hybridized to the *HumanHT-12 Expression Chips, version 4* following the Direct Hybridization assay.

The data obtained was imported into R and analyzed with the beadarray Bioconductor package (Dunning et al., 2007) and the illuminaHumanv4.db annotation package. We summarized the data for each array using the *summarize()* function of beadarray with default parameters ($\log_2$ transformation, removal of outliers with a 3 median absolute deviation cutoff) and removed all probes without a quality score or with a "Bad" quality score in the annotation package. We then normalized the data with the *normaliseIllumina()* function with the quantile method and retrieved Ensembl IDs for each array probe using biomaRt. We then performed the differential expression analysis using limma with the model formula ~0+Condition, where Condition identifies the Control samples, FOXA2-KD samples and FOXA2-DS-S samples. We also supplied to the *lmFit()* function a weight for each array estimated using the *arrayWeights()* function. The GO enrichment analysis was performed separately on significantly up-regulated (adjusted p-value<0.05 and $\log_2$ fold change>0) and down-regulated (adjusted p-value<0.05 and $\log_2$ fold change<0) genes using the TopGO package. As background set we used all probes in the array with an Ensembl gene id. The GO enrichment was performed with the classic algorithm and p-values calculated with the fishes exact test. P-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg method as implemented in the *p.adjust()* function. All GO terms with less than 20 annotated genes were excluded from the analysis.

9.9 Microarray meta-analysis

This methodology pertains to work done in the laboratory of Dr Maracaja-Coutinho

The probe set sequences for the GPL570 platform (Affymetrix Human Genome U133 Plus 2.0 Array) were retrieved from the Affymetrix website, and aligned against the human genome (hg38) using Blat (Kent, 2002). We next removed probe sets with less than 90% identity and coverage, and cross-referenced the remaining ones against the pcRNAs genomic coordinates (BED12 format) and the protein coding genes (Gencode version 23) using BEDtools (Quinlan, 2014). Probe sets were annotated as pcRNA or as coding gene if at least 70% of their sequence mapped to the reference transcript sequence.

We then download from the GEO database 63 microarray studies on the GPL570 platform, which contained tumour and non-tumour tissue samples. For each study, raw data (CEL files) were processed and normalized using the RMA algorithm, and samples were manually classified as "normal" or "tumour" according to the description provided by authors. We used Student's

t-test (fold-change > 1.25 and p-value < 0.005) to identify transcripts differentially expressed in either tumour or normal samples. Spearman correlation was used to compare the expression between the pcRNA and its associated coding gene. Plots were generated in R with gplots and ggplot2.

9.10 pcRNA histone modification profiles

To produce histone modification maps of pcRNA promoters we downloaded the normalised bigWig files from the EBI ENCODE respository for 14 ChIP-Seq experiments (Control, Ctcf, H2az, H3k27ac, H3k27me3, H3k36me3, H3k4-me1, H3k4me2, H3k4me3, H3k79me2, H3k9ac, H3k9me1, H3k9me3, H4k20-me1) in GM12878, H1-hESC, HSMM and K562 cells. We then converted the data to bedGraph format (with bigWigToBedGraph) and used liftOver to convert the coordinates from hg19 to hg38. Then, we used bedtools mergeBed to merge the overlapping intervals and converted the resulting files back to bigWig format (bedGraphToBigWig).

For each cell line we then used the computeMatrix of the Deeptools package (reference point TSS) to calculate the coverage in each ChIP-Seq experiment of each pcRNA as well as of 100 random Gencode lncRNAs and 100 random Gencode coding Genes (random coding genes and lncRNAs selected with Bedtools sample with seed 383847). The resulting matrix was then loaded in R to produced the profile plots.

9.11 Analysis of H3K27me3 in ESCs

To study the H3K27me3 profiles of pcRNA promoters, the bedGraph files (in hg38 coordinates, see above for details of conversion from hg19) of H3K27me3 and H3K4me3 in GM12878, H1-hESC, HSMM and K562 have been mapped to the promoters of pcRNAs (defined as TSS +/- 1Kb) using the mapBed tool of bedtools to calculate the mean coverage of each promoter in each cell line. The data was then loaded into R and the data for H1-hESCs was subjected to hierarchical clustering using the hclust function (default parameters) on the Euclidean distances matrix (dist function) between the base 10 logarithms of the mean promoter coverage for H3K27me3 and H3K4me3 (the $\log_{10}$ was calculated after adding 0.01 to each value). The GO enrichment for the coding genes associated to the pcRNAs in each cluster was performed using TopGO (classic algorithm) using the Fisher Exact Test to compute p-values. P-values correction and background set were the same as described for "*GO enrichment of pcRNA-associated coding genes*".

9.12 ENCODE ChIP-seq data analysis

These methods refer to work done by Dr Han

The ENCODE ChIP-Seq peak data for 2,216 experiments were downloaded in hg19 coordinates and converted to hg38 coordinates using liftOver. To identify transcription factors that bind pcRNAs or pcRNA-associated coding gene promoters we overlapped with ChIP-Seq peaks with the promoter regions (500bp upstream of the TSS) of pcRNAs and pcRNA-associated coding genes. The Pearson correlation in the binding profiles between pcRNA/coding promoters was calculated applying the *corr2* function of Matlab to the binary matrices of TF binding. The significance was calculated by Monte Carlo simulation.

9.13 Known TF-binding motif data analysis

These methods refer to work done by Dr Han

The known motifs of TF-binding were downloaded from JASPAR (freeze 2014-12-10, 263 motifs), Kheradpour and Kellis (2014) (2,065 motifs) and Jolma et al. (2013) (843 motifs). We then applied the same analytical procedures as described for the ENCODE ChIP-seq data analysis.

9.14 Identification of CTCF binding sites in pcRNA promoters

To identify CTCF binding sites within pcRNA promoters we downloaded the TFBS clusters (V3) from the ENCODE portal at the UCSC Genome Browser (wgEncodeRegTfbsClusteredWithCellsV3.bed.gz) and filtered the file for CTCF sites. We then converted the CTCF binding sites to hg38 coordinates using the liftOver tool and calculated how many pcRNA promoters overlapped HiC loops by (1) extending the TSS of each pcRNA by 2kb in both directions, (2) merging overlapping promoters (bedtools merge) and (3) intersecting the promoters with the CTCF sites. We repeated the same procedure for pcRNA-associated genes, Gencode coding genes and Gencode lncRNAs. To test whether pcRNA promoter were significantly enriched in CTCF binding sites compared to Gencode lncRNAs we performed a hypergeometric test as implemented in the phyper function in R.

9.15 Identification of HiC loops that overlap pcRNAs

We obtained the annotation of HiC loops by downloading the loops list files for HMEC, HUVEC, NHEK, K562, HeLa, KBM7, IMR90 and GM12878 cells

deposited on GEO (GSE63525) and converted the intervals into Hg38 coordinates using liftOver. To calculate how many pcRNA promoters overlapped HiC loops we first extended the TSS of each pcRNA by 2kb in both directions, merged overlapping promoters (bedtools merge) and intersected the promoters with the loop coordinates. We also repeated the same procedure for all pcRNA-associated coding genes, Gencode coding genes and Gencode lncRNAs. To test whether pcRNA promoters are significantly enriched in HiC peaks compared to Gencode lncRNAs we performed a hypergeometric test as implemented in the *phyper()* function in R.

We applied the same strategy to identify TAD boundaries overlapping pcRNAs promoters. However, because TAD boundaries are single nucleotides rather then intervals, we extended each boundary of 10kb in each direction.

To analyse the end-points of the loops that overlap pcRNA promoters we downloaded the ENCODE Broad HMM data (Ernst et al., 2011) from the UCSC repository for GM12878, H1-hESC, HEPG2, HMEC, HSMM, HUVEC, K562, NHEK and NHLF cells. After converting the coordinated to Hg38 with liftOver we intersected each HMM dataset with the coordinates of the end points of the loops that overlapped pcRNAs or – as a control – all Gencode lncRNAs.

The data were then loaded into R for further analysis. First, we simplified the data by reducing the number of HMM categories in the following way: Strong and Weak Enhancer categories were grouped as Enhancer; Active, Weak and Poised promoter were grouped as Promoter; Txn_Elongation, Txn_Transition and Weak_Txn were grouped as Transcript; everything else was grouped as Other.

Then, for each end-point in each cell line we calculated the fraction covered by each HMM category and plotted these data as a heatmap using the heatmap.2 function of the gplots package. To determine whether pcRNA-loop end-points were enriched in any specific HMM category we calculated for each HMM category x the fraction of end-points annotated as x for at least $y\%$ of their length, where y ranged from 1 to 0 in steps of 0.1. Finally, we compared this distribution to the distribution obtained for all Gencode lncRNAs using the Kolmogorov-Smirnov test (as implemented in the *ks.test()* function in R).

9.16 TAD/Loop Boundary Enrichment Analysis

These methods refer to work done by Dr Han

To identify pcRNAs localized at the boundary of TADs/Loops, we generated a density plot that shows the cumulative count of pcRNAs appearing across TAD/Loop regions (for each TAD including 10% proximity regions outside

the TADs). We extended the TSS of each pcRNA by 2kb in both directions, then merged overlapping regions (BEDTools merge). We then intersected the extended promoters with the loop and TAD coordinates (BEDTools intersect).

To visualize the cumulative counts as a density plot, we only cumulated 10-bp window centered by TSS of each overlapping pcRNA to show precise localization of pcRNA TSS. We used all lncRNAs in the Gencode database (excluding pcRNAs) as a control. We then performed Kolmogorov-Smirnov test (as implemented in the kstest2 function in MatLab) to check how the enrichment is significant.

9.17 PhastCons Conservation Analysis

These methods refer to work done by Dr Han

To understand the general conservation level of pcRNAs, we used phastCons scores resulting from the multiple alignments of 99 vertebrate genomes to the human genome. The files were downloaded in wigFix format from the UCSC database (hg38.100way.phastCons). We extracted the phastCons scores corresponding to pcRNA exons and calculated the average score for each pcRNA. The same calculation was also done for Gencode coding genes and lncRNAs. The density of the normalized phastCons scores per pcRNA was plotted using a Kernel smoothing function estimate (as implemented in the ksdensity function in MatLab). To test whether pcRNA are significantlya more conserved than Gencode lncRNAs, we applied the Kolmogorov-Smirnov test as implemented in the kstest2 function in MatLab.

9.18 Conserved domain search

These methods refer to work done by Dr Han

To identify conserved domains, we aligned transcribed sequences of human pcRNAs against their corresponding mouse pcRNAs. We took two alignment approaches: sliding-window and exon-by-exon. For the first approach, we made a 200nt-long window on each human pcRNA sequence and shifted the window by 40nt to align against the whole length of the transcribed mouse pcRNA sequence. For the second approach, we took each exon of human pcRNAs and aligned them against the whole length of the transcribed mouse pcRNA sequences. In both approaches, we used the Matlab function localalign, which returns local optimal and suboptimal alignments between two sequence. We found that both approaches gave highly concordant results. We then applied the following filtering steps: (1) we retained alignments only if the alignment score was greater than 100 or the ratio of identical matches

was greater than 80%, (2) we removed duplicate alignments among isoforms of pcRNAs based on an annotation of merged pcRNA isoforms (3) we removed alignments if the aligned regions in the human and mouse pcRNAs were not in same order of exons on their transcribed sequences, and (4) we retained the best alignment if there were multiple alignments for the same region. To annotate the merged isoforms we extracted the exonic regions of each pcRNA and merged them using the BEDTools merge function. This process allowed us to generated a heatmap of conserved regions in human pcRNAs, which we clustered using the MATLAB function kmeans on the squared Euclidean distance. We found 16 clusters and merged them into four larger clusters.

Motif search in conserved domains

These methods refer to work done by Dr Han

To determine which regulatory motifs are over-represented in conserved domains with respect to background non-conserved regions, we identified all possible ungapped 8-mers in the conserved domains and computed their frequency. An 8-mer is considered over-represented if its frequency in the conserved domain is significantly higher than the frequency in background non-conserved region. In the list of over-represented motifs, we found the presence of repeats that consisted of a single nucleotide or dimer repeated for the entire 8-mer. This phenomenon is common in genomic sequences and generally is associated with non-functional components, and thus, these were filtered out.

To assess the statistical significance of the computed frequency for the over-represented motifs, we generated random sequences according to the nucleotide composition of the original sets of sequences. The frequencies for the random 8-mers were computed, and the distribution of the frequencies was approximated by the extreme value distribution. We used the MATLAB function gevfit to compute the maximum likelihood estimation of the extreme value distribution. We then overlaid a scaled version of its probability density function, computed using the Matlab function gevpdf, with the histogram of the frequency of the random 8-mer sequences. We repeated this process 100 times for bootstrapping and calculated the p-value. We concluded that the over-representation of the 8-mer motifs in conserved domain is statistically significant if the p-value estimate is less than $1x10^{-4}$.

Consensus motifs and De novo motif discovery

These methods refer to work done by Dr Han

To identify consensus motifs, the 32 enriched 8-mers were phylogenetically clustered into 10 groups. We used the Matlab function seqlinkage to construct a phylogenetic tree from pairwise distances. We then used the function seqlogo to identify consensus motif and the weight matrix for the clustered 8-mer(s) in each group. We then used the MEME suite to find known transcription factors with motifs matching the 10 identified consensus motifs.

Enriched motif search in enhancer region of the other end of loop anchor points

These methods refer to work done by Dr Han

We checked whether the 32 enriched motifs found in theconserved pcRNA domains are also over-represented in enhancer regions on the other end of the loop anchor points. The definitions for enhancer region and loop anchor points are described in previous Method section, "Identification of HiC loops that overlap pcRNAs". The 32 enriched motifs were searched in both pcRNA transcribed sequences as well as enhancer regions on the other end of overlapping loop anchor points. We counted a given motif only if the motif was found in both pcRNA and enhancer region. We also searched non-enhancer regions on the other end of the loop anchor points as a control set. The counts were normalized by the total length of enhancer or non-enhancer region accordingly.

10

Secretion mechanisms of extracellular microRNAs in Neural Stem Cells

10.1 Expression analysis of miRNAs

The small RNA-Seq experiments on NPCs, EVs and EXOs in Basal, Th1 and Th2 conditions consisted of four different experiments as outlined in Table 10.1. The culturing of NPCs, purification of EXOs and EVs and RNA extraction were realised by Dr Iraci and Dr Cossetti in the laboratory of Dr Pluchino as described in Cossetti et al. (2014a). Library preparation and sequencing for the first experiments were realised by GeneWorks (Adelaide, Australia) on the Illumina GAII platform. The library preparation followed the standard Illumina protocol with a modified set of size markers optimised for sequencing of miRNAs. Library preparation and sequencing for the successive experiments were realised by the Beijing Genomic Institute (Shenzhen, Guangdong, China) on the Illumina GAII platform following the standard Illumina protocol.

The data was analysed using the Kraken pipeline (Davis et al., 2013). Briefly, the reads were trimmed with Reaper to remove the 3' adapter (`TCG TAT GCC GTC TTC TGC TTG`) and then filtered with seqImp (v13-274) to only retain those of length between 18 and 26nt after trimming. Reads were then mapped to the reference mouse genome (NCBI37/mm9) with seqImp, discarding reads with more than 2 mismatches or mapping to more than 20 genomic locations. The abundance of miRNAs was quantified with seqImp against

Experiment	NPC	EXO	EV
Exp 1 (GW)	Line A	Line A	Line A
Exp 2 (BGI)	Line B		
Exp 3 (BGI)	Line C		
Exp 4 (BGI)		Lines B,C	

Table 10.1 *Table reporting the four sRNA-Seq experiments performed on NPCs and vesicles. The NPC lines A,B and C refers to three independent NPC preparations produced in the laboratory of Stefano Pluchino as reported in Cossetti et al., 2014a. The sequencing was performed using the Illumina technology by GeneWorks Adelaide, Australia (GW) and the Beijing Genomic Institute, Shenzhen, Guangdong, China (BGI).*

miRBase v18 (Kozomara and Griffiths-Jones, 2011). Reads were counted toward the expression of a certain miRNA if they overlapped for at least 15nt. In case reads mapped to multiple loci their depth was equally split across all loci. The data was then imported into R/Bioconductor (Huber et al., 2015) for statistical analysis. Any miRNA with 0 counts in all samples was removed, and the counts matrix was then normalised using DESeq2 (Love et al., 2014). To evaluate the effects of cytokine stimulation we performed the differential expression analysis using the design formula `Experiment + Condition`, where `Experiment` was a factor representing the sequencing batch and `Condition` was a factor with 9 levels indicating the biological sample (NPC, EV or EXO in Basal, Th1 or Th2). After fitting the model we extracted the contrasts of interest using the `results()` function of DESeq2 with default parameters. To identify significantly secreted miRNAs we fitted a separate model with the design formula `Experiment + Type`, where `Type` was a factor of 3 levels representing the biological sample (NPC, EXO or EV). We then extracted the contrast EXO vs NPC and EV vs NPC to identify miRNAs significantly secreted in EXOs and EVs respectively. In all cases, the p-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

The heatmap of sample to sample distanced was produced by calculating the Euclidean distance between samples after applying the variance stabilising transformation implemented in the `varianceStabilizingTransformation()` function of DESeq2. The heatmap of scaled counts for significant miRNAs was realised by selecting any miRNA with an adjusted p-value <0.01 in any contrast and transforming the counts into z-score using the `scale()` function in R.

10.2 Dual luciferase assay for miRNA secretion

This experiment has been performed by Dr Nunzio Iraci in the Pluchino laboratory.

The target sequences of mmu-miR-103 and mmu-miR-365 were cloned in four copies downstream of the psiCHECK-2 Luciferase vector (Promega). The vectors were then transfected in 2×10^4 NIH 3T3 cells using 1 ng of vector for the Dharmacon® DharmaFECT® Duo Transfection reagent (Thermo Scientific). After 4h in culture EVs from NPCs in Basal or Th1 conditions were added to the 3T3 cell media (NPC/EV:3T3 ratio 50:1). After 6h the Firefly and Renilla luciferase was measures using the Dual Luciferase Assay kit (Promega) in a GloMax-96 Microplate Luminometer (Promega) manufacturer's instructions.

EVs from NPCs cultured either in Basal or Th1 condition were added after 4h (NPC/EV:3T3 ratio 50:1) and the activity of Firefly or Renilla luciferase was measured 6h after the addition of EVs with the Dual Luciferase Assay kit (Promega) in a GloMax-96 Microplate Luminometer (Promega) according to the manufacturer's instructions. The data were analysed with an Anova one-way test with Dunnett's Multiple Comparison post-hoc test.

10.3 Identification of miRNA promoters

To identify the promoters of miRNAs in the mouse genome we first compiled an annotation of intronic and intergenic miRNA precursors using the annotations provided by miRBase (v19).

Then, we assembled a collection of features indicative of promoters:

1. CpG islands were downloaded in BED format from the CpG track of the UCSC genome browser (Gardiner-Garden and Frommer, 1987).

2. DNaseI Hypersensitivity Sites (DHSs) data were produced by ENCODE (University of Washington) on 48 cell lines. Overlapping DHS peaks from multiple cell lines were merged using BEDToolsmerge (Quinlan, 2014) and then filtered to only retain the portion of the peaks that had signal in at least 10 cell lines.

3. ChIP-Seq data for trimethylation of histone 3 lysine 4 (H3K4me3) on 32 cell lines from the ENCODE project (PSU, LICR and Caltech, ENCODE Project Consortium et al., 2012) and on Neural Stem Cells from Mikkelsen et al. (2007). The data were processed in the same way as the DHS data.

4. Eponine predictions of Transcriptional Start Sites (TSSs). Eponine (Down and Hubbard, 2002) was executed (threshold 0.990) on 10kb of DNA sequence upstream of each pre-miRNA. The predictions were then expanded to a fixed size of 200nt centred around the predicted TSSs.

5. Annotated TSS of host gene for intronic miRNAs. The annotation of miRNA host genes was obtained from miRBase, from which we extracted the coordinates of the TSS. In order to avoid considering the promoters of other upstream genes as miRNA promoters, we annotated the TSSs of upstream, non-host transcripts as 'Insulator" TSSs (see below).

6. Genome-wide chromatin segmentation using chromHMM based on data produced by the ENCODE consortium ENCODE Project Consortium et al., 2012; Ernst and Kellis, 2012. This data was produced in human,

and we extracted promoter annotations and converted them to mouse coordinates (i.e. syntenic promoters) using the liftOver tool.

For each of the annotation datasets reported above we extracted all features residing in a window of 100kb upstream of miRNA loci using the windowBed tool of BEDTools.

To produce an annotation of candidate promoters we first split features longer than a predetermined threshold into multiple, adjacent features of size equal to the threshold. The threshold was chosen using Tuckey's method for outliers detection, therefore setting it to the third quartile of the feature size distribution + 1.5 times the inter quantile range (Tukey, 1977). Then, the features were clustered into candidate promoters using the merge tool of BEDToolswith a distance threshold of 200nt (corresponding approximately to 25 % of the mean size of a feature as well as the DNA size in one nucleosome). Finally, we identified all clusters upstream of pre-miRNAs and assigned them a score by calculating how many features supported them (BEDToolsannotate). Each feature was given the same weight except DHS that was considered 20 % more than the others, due the high quality of the data and the strength of DNase Hypersensitivity as a promoter mark. The presence of an Insulator TSS overlapping a cluster was given an arbitrarily large negative score, in order to discard the promoters of upstream genes.

This score was then weighted according to the distance of the cluster to the pre-miRNA, in order to favour closer candidate promoters in case of ties in the score. The weighting was proportional to the squared root of the distance according to the following formula:

$$\mathrm{Score} = S_p \cdot (1 - \sqrt{d})$$

where S_p is the sum of the scores of the features that support the cluster and d is the distance in megabases of the cluster from the pre-miRNA. For each miRNA the cluster with the highest score was then considered the best cluster. To calculate the z-score of each cluster we used the `scale()` function in R on the vector of scores of all the clusters of a given miRNA.

Pomoter conservation was measured by intersecting the coordinates of each cluster with the PhasCons30wayPlacental track of the UCSC genome browser (Siepel et al., 2005) using the R package rtracklayer (Lawrence et al., 2009). For comparison the same procedure was applied to 1000 random genomic intervals. The significance of the different conservation between miRNA promoters and random regions was calculated with the Welch's t-test (Welch, 1947).

10.4 Analysis of TF binding in secreted miRNA promoters

In order to identify Transcription Factor Binding Sites (TFBSs) enriched in the promoters of secreted miRNAs, we first extracted the promoter sequence of murine miRNAs with a significant p-value (≤ 0.05) in either EXOs vs NPCs or EVs vs NPCs. We then used the tool `findMotifsGenome` of the Homer suit to identify short motifs enriched in this set of promoters compared to a background set consisting of the promoter sequences of all miRNAs expressed in NPCs. Homer was executed to identify motifs of 6,8,10 and 12 nucleotides and to correct all p-values with an empirical false discovery rate based on 1000 random shuffles of the promoter sets.

10.5 Identification of miRNA secretion motif

10.5.1 *Motif enrichment*

The normalised small RNA-Seq data (see 10.1) was filtered to remove miRNAs with mean counts below the 50th percentile of mean counts (i.e. mean normalised counts>137.2332). Then, only miRNAs significantly enriched or depleted in Exosomes (EXOs) vs Neural Progenitor Cells (NPCs) (p-value<0.01) were retained. BCRANK was then started on the filtered sequences ordered by $\log_2$fold change (EXOs vs NPCs) in decreasing order with the following settings:

- length=3 (starting motif length)
- restarts=100 (number of random restarts)
- use.P1=FALSE (no penalty for bases other than ACGT)
- use.P2=FALSE (no penalty for repeats)

The identified enriched motifs were then mapped to individual miRNAs using the `matchingSites` function of BCRANK and visualised in a volcano plot with ggplot2. The sequence logos of the motifs were generated with the seqLogo package.

For the motif refinement analysis BCRANK was executed as decribed above, but the input ordered list of miRNAs was produced using the thresholds outlined in table 5.4 (page 136).

10.5.2 *Arm switch analysis*

To select miRNAs undergoing arm switch we selected mature miRNAs with a significantly positive enrichment in EXO vs NPCs for one arm and a significantly negative enrichment for the other arm (adjusted p-values<0.05). BCRANK was then executed on the arm switching miRNAs as previously described.

10.5.3 *GC content bias analysis*

The GC content of miRNAs was calculated with the `GC` function of the Bioconductor package seqinr. The correlations between GC content and expression or secretion fold change were computer with the R function `lm`.

11

Extracellular vesicles from Neural Stem Cells transfer IFN-ɣ via Ifngr1 to activate Stat1 signalling in target cells

The methods for all the wet-lab experiments performed for this project by Dr Cossetti, Dr Iraci and collaborators are reported in full[1] in Cossetti et al. (2014a).

11.1 RNA Sequencing

Total RNA was purified from one preparation of NPCs, Extracellular Vesicles (EVs) and exosomes in Basal, Th1 and Th2 conditions using Trizol.

The purity and integrity of the extracted RNA were then measured by BioAnalyser (Agilent). The Poly-A selection and the construction of a paired end unstranded library were done by EASIH (The Eastern Sequence and Informatics Hub, University of Cam, Cambridge) according to the Illumina standard protocol. The library was then sequenced by EASIH on the Illumina Genome Analyser II. The average sequencing depth was ~17.2×10^6 read pairs of 72nt each. The sequenced reads were then pre-processed using the Kraken suite of tools (Davis et al., 2013) in order to trim 3' adaptor contaminations and discard reads with 5' adaptor contamination. Further trimming was done to remove low quality stretches and poly-N stretches.

The redundant reads in each sample were then collapsed keeping the highest quality score in each position. The reads were mapped to the reference mouse genome (NCBI37/mm9) with Tophat (Trapnell et al., 2012b) using the Ensembl transcriptome (v66) as a reference. Tophat was executed providing the mean fragment length and standard deviation, which were estimated by aligning a random sample of reads from each sample to a set of Ensembl unspliced exonic transcripts longer than 1000bp. Gene expression estimates were obtained using Cuffdiff (v2.0.2, Trapnell et al., 2012a). The data was then filtered in the following way:

- genes with a status other than <OK> in any sample were discarded;
- genes with redundant short names were collapsed retaining only the one having highest inter-quantile range;
- genes with mean expression below the median of the mean gene expression of all genes were discarded;

[1] `http://dx.doi.org/10.1016/j.molcel.2014.08.020`

- genes whose longest isoform was shorter than the mean fragment length were removed;
- genes with an expression of 0 in all samples were removed.

The list of IFN-ɣ pathway genes for histogram plots was obtained from the literature.

The GO enrichment analysis on genes differentially expressed in NPC Th1 was performed with the topGO package (v. 2.12.0) in R/Bioconductor using the classic algorithm and Fisher's exact test. The gene list for GO enrichment was obtained by selecting genes with fold-change >5 or <0.2 from the background list obtained from the filtering outlined above. The p-values obtained were corrected using the Benjamini-Hochberg method as implemented in the p.adjust function in R.

11.2 Microarray analysis on target cells exposed to EVs

Total RNA was extracted with Trizol from 3T3 cells not exposed or exposed to EVs from Basal, Th1 or Th2 NPCs.

The RNA was amplified and labelled using the SuperScript Indirect RNA Amplification System (Invitrogen) and Alexa Fluor 555 Decapack Set (Invitrogen) according to the manufacturer's instructions. After labelling, the RNA was hybridised to NCode microarrays (Invitrogen) using a MAUI Hybridization System (BioMicro Systems), according to the manufacturer's protocol. The hybridisation solution was loaded on the array at 42 °C and hybridisation was carried out overnight. Scanning of the arrays was done with an Agilent DNA microarray scanner at 5 µm resolution. Feature extraction was done with the NimbleScan software using manual grid adjustment and auto spot finding and segmentation. Data analysis was done in R using the limma package (Smyth et al., 2005). After background correction and normalisation (Smyth and Speed, 2003), differential expression testing was done by fitting a linear model to the data and calculating Bayesian statistics as implemented in limma. The resulting p-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg method.

11.3 Functional networks using GeneMANIA

The microarray and SILAC experiments were used to produce a list of genes and proteins differentially expressed in 3T3 cells exposed to Th1 EVs vs 3T3 cells exposed to basal EVs. This genes and proteins list was then used as an input to create interaction networks with GeneMANIA (Mostafavi et al., 2008)

using the default weighting method. GO enrichment analysis of genes and proteins in the network was also done with GeneMANIA. The GO terms with corrected p-values (Benjamini-Hochberg method) were considered significant.

Part VII

Appendix

a

Publications

- Cossetti C, Smith JA, Iraci N, **Leonardi T**, Alfaro-Cervello C, Pluchino S, Extracellular membrane vesicles and immune regulation in the brain. Front Physiol, 2012.
- Hill AF, Pegtel DM, Lambertz U, **Leonardi T**, O'Driscoll L, Pluchino S, Ter-Ovanesyan D, Nolte-'t Hoen EN. ISEV position paper: extracellular vesicle RNA analysis and bioinformatics. J Extracellular Vesicles, 2013.
- Smith JA, **Leonardi T**, Huang B, Iraci N, Vega B, Pluchino S. Extracellular vesicles and their synthetic analogues in aging and age-associated brain diseases. Biogerontology, 2014
- Cossetti C*, Iraci N*, Mercer TR, **Leonardi T**, Alpi E, Drago D, Alfaro-Cervello C, Saini HK, Davis MP, Schaeffer J, Vega B, Stefanini M, Zhao C, Muller W, Garcia-Verdugo JM, Mathivanan S, Bachi A, Enright AJ, Mattick JS, Pluchino S. Extracellular Vesicles from Neural Stem Cells Transfer IFN-γ via Ifngr1 to Activate Stat1 Signaling in Target Cells. Molecular Cell, 2014
- Clark MB, Mercer TR, Bussotti G, **Leonardi T**, Haynes KR, Crawford J, Brunck ME, Cao KA, Thomas GP, Chen WY, Taft RJ, Nielsen LK, Enright AJ, Mattick JS, Dinger ME. Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. Nature Methods, 2015
- Fuster-Matanzo A, Gessler F, **Leonardi T**, Iraci N, Pluchino S. Acellular approaches for regenerative medicine: on the verge of clinical trials with extracellular membrane vesicles? Stem Cell Res Ther, 2015
- Lener T, …, **Leonardi T**, Pluchino S, …, Giebel B. Applying extracellular vesicles based therapeutics in clinical trials - an ISEV position paper. J Extracell Vesicles. 2015
- Iraci N*, **Leonardi T***, Gessler F, Vega B, Pluchino S. Focus on Extracellular Vesicles: Physiological Role and Signalling Properties of Extracellular Membrane Vesicles. Int J Mol Sci. 2016
- Bussotti G, **Leonardi T**, Clark MB, Mercer TR, Crawford J, Malquori L, Notredame C, Dinger ME, Mattick JS and Enright AJ Improved defini-

tion of the mouse transcriptome via targeted RNA sequencing. Genome Research 2016

- Amaral PP*, **Leonardi T***, Han N*, Viré E, Gascoigne D.K., Arias-Carrasco R, Büscher M, Zhang A, Pluchino S, Maracaja-Coutinho V, Nakaya HI, Hemberg M, Shiekhattar R, Enright AJ, Kouzarides T. Genomic positional conservation identifies topological anchor point (tap)RNAs linked to developmental loci. bioRxiv 2016

b

Acknowledgements

This work would have been impossible without the help and support of numerous colleagues and friends. I would like to thank my supervisors, Dr Anton Enright and Dr Stefano Pluchino for their continued guidance and for providing the ideal environment and the resources for completing my studies. I also want to thank the invaluable contribution of all the collaborators and co-authors who contributed to the work reported in this thesis. In particular, I thank Dr Paulo Amaral, who conceived the pcRNAs project and with whom I had many hours of fruitful discussions. I am also very grateful to Dr Nunzio Iraci, for his guidance and for his invaluable contributions to all the work pertaining to extracellular vesicles, and to Dr Chiara Cossetti, who led the Stat1 project. I also want to thank Dr Namshik Han for his contribution to the pcRNAs project.

In addition, I would like to acknowledge the members of my Thesis Advisory Committee for their continued feedback and guidance: Dr Paul Bertone, Prof Donal O'Carrol and Dr George Kassiotis.

I am also very grateful to all members of the Enright lab, past and present, for their scientific guidance and friendship: Giovanni Bussotti, Stijn van Dongen, Adrien Leger, Jack Monahan, Leonor Quintais, Harpreet Saini, Dimitrios Vitsios and Matthew Davis, toward whom I am particularly grateful for the invaluable scientific input that he provided throughout my PhD. I am equally thankful to all the members of the Pluchino lab: Elena, Matteo, Alice, Almudena, Jayden, Dai, Nunzio, Chiara, Inma, Josh, Florian, Luca, Jeroen, Bea, Giulio, Iacopo, Giulia and Joan.

I would also like to thank Prof Tony Kouzarides, and all the members of his lab, who were always happy to adopt me as one of them for meetings and parties. In particular, I want to thank Emmanuelle Viré and Magdalena Büscher for their roles in the pcRNA project.

I am also thankful to several members of the Mattick lab with whom I have collaborated in the last few years: Marcel Dinger, Tim Mercer, Mike Clark and Dennis Gascoigne, who wrote the first implementation of the pcRNAs pipeline.

Many thanks also to all the friends and colleagues with whom I have enjoyed many scientific discussions in front of a beer: Matteo Donegà, Giulia Tyzack, Marco Paoli, Edoardo Gaude, Seth Cheetham and Emanuele Osimo.

And a special thanks to those who proof read this thesis: Stijn, Mat, Nunzio and Paulo.

I also want to thank the support staff at the European Bioinformatics Institute and the European Molecular Biology Laboratory for their continued and efficient help.

Finally, I want to thank my parents, for their support in my studies and for the fantastic environment in which I grew up, my brothers Federico and Giovanni and my sister Giulia, who never refused to give me a drop of blood when I was looking for the Barr bodies.

And the biggest thank to Cecilia, who has been at my side in these years and who has always given me unconditional love and support.

Part VIII

Bibliography

Adalsteinsson, B. T. and A. C. Ferguson-Smith (2014). "Epigenetic control of the genome-lessons from genomic imprinting." In: *Genes (Basel)* 5.3, pp. 635–655. DOI: 10.3390/genes5030635.

Alexiou, P. et al. (2009). "miRGen 2.0: a database of microRNA genomic information and regulation". In: *Nucleic Acids Res* 38.Database, pp. D137–D141. DOI: 10.1093/nar/gkp888.

Altman, J. (1963). "Autoradiographic investigation of cell proliferation in the brains of rats and cats." In: *The Anatomical record* 145, pp. 573–591. DOI: 10.1002/ar.1091450409.

Amaral, P. P. and J. S. Mattick (2008). "Noncoding RNA in development". In: *Mamm Genome* 19.7-8, pp. 454–492. DOI: 10.1007/s00335-008-9136-7.

Amaral, P. P. et al. (2008). "The eukaryotic genome as an RNA machine". In: *Science (New York, NY)* 319.5871, pp. 1787–1789. DOI: 10.1126/science.1155472.

Amaral, P. P. et al. (2009). "Complex architecture and regulated expression of the Sox2ot locus during vertebrate development". In: *RNA (New York, NY)* 15.11, pp. 2013–2027. DOI: 10.1261/rna.1705309.

Amaral, P. P. et al. (2011). "lncRNAdb: a reference database for long noncoding RNAs". In: *Nucleic Acids Res* 39.Database issue, pp. D146–51. DOI: 10.1093/nar/gkq1138.

Amaral, P. P. et al. (2013). "Non-coding RNAs in homeostasis, disease and stress responses: an evolutionary perspective". In: *Brief Funct Genomics* 12.3, pp. 254–278. DOI: 10.1093/bfgp/elt016.

Ameur, A. et al. (2009). "Identification of candidate regulatory SNPs by combination of transcription-factor-binding site prediction, SNP genotyping and haploChIP." In: *Nucleic Acids Res* 37.12, e85. DOI: 10.1093/nar/gkp381.

An, T. et al. (2015). "Exosomes serve as tumour markers for personalized diagnostics owing to their important role in cancer metastasis." In: *Journal of extracellular vesicles* 4, p. 27522. DOI: 10.3402/jev.v4.27522.

Anderson, H. C. (1969). "Vesicles associated with calcification in the matrix of epiphyseal cartilage." In: *The Journal of cell biology* 41 (1), pp. 59–72. DOI: 10.1083/jcb.41.1.59.

Andreola, G. (2002). "Induction of Lymphocyte Apoptosis by Tumor Cell Secretion of FasL-bearing Microvesicles". In: *J Exp Med* 195.10, pp. 1303–1316. DOI: 10.1084/jem.20011624.

Arner, E. et al. (2015). "Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells". In: *Science (New York, NY)* 347.6225, pp. 1010–1014. DOI: 10.1126/science.1259418.

Arnold, P. Y. and M. D. Mannie (1999). "Vesicles bearing MHC class II molecules mediate transfer of antigen from antigen-presenting cells to CD4+ T cells." In: *Eur J Immunol* 29.4, pp. 1363–1373. DOI: 10.1002/(sici)1521-4141(199904)29:04<1363::aid-immu1363>3.3.co;2-s.

Arroyo, J. D. et al. (2011). "Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma." In: *Proceedings of the National Academy of Sciences* 108.12, pp. 5003–5008. DOI: 10.1073/pnas.1019055108.

Askarian-Amiri, M. E. et al. (2014). "Emerging role of long non-coding RNA SOX2OT in SOX2 regulation in breast cancer." In: *PloS one* 9 (7), e102140. DOI: 10.1371/journal.pone.0102140.

Augui, S. et al. (2011). "Regulation of X-chromosome inactivation by the X-inactivation centre." In: *Nat Rev Genet* 12.6, pp. 429–442. DOI: 10.1038/nrg2987.

Auyeung, V. C. et al. (2013). "Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing." In: *Cell* 152.4, pp. 844–858. DOI: 10.1016/j.cell.2013.01.031.

Bacigaluppi, M. et al. (2008). "Neural stem/precursor cells for the treatment of ischemic stroke". In: *J Neurol Sci* 265.1-2, pp. 73–77. DOI: 10.1016/j.jns.2007.06.012.

Baek, D. et al. (2008). "The impact of microRNAs on protein output". In: *Nature* 455.7209, pp. 64–71. DOI: 10.1038/nature07242.

Baer, B. and R. Kornberg (1983). "The protein responsible for the repeating structure of cytoplasmic poly(A)-ribonucleoprotein." In: *J Cell Biol* 96.3, pp. 717–721. DOI: 10.1083/jcb.96.3.717.

Baietti, M. F. et al. (2012). "Syndecan–syntenin–ALIX regulates the biogenesis of exosomes". In: *Nat Cell Biol* 14.7, pp. 677–685. DOI: 10.1038/ncb2502.

Balaj, L. et al. (2011). "Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences." In: *Nat Commun* 2, p. 180. DOI: 10.1038/ncomms1180.

Balbin, O. A. et al. (2015). "The landscape of antisense gene expression in human cancers". In: *Genome Res* 25.7, pp. 1068–1079. DOI: 10.1101/gr.180596.114.

Ballester, B. et al. (2014). "Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways". In: *eLife* 3, e02626. DOI: 10.7554/eLife.02626.

Barr, I. et al. (2012). "Ferric, not ferrous, heme activates RNA-binding protein DGCR8 for primary microRNA processing." In: *Proc Natl Acad Sci U S A* 109.6, pp. 1919–1924. DOI: 10.1073/pnas.1114514109.

Barski, A. et al. (2009). "Chromatin poises miRNA- and protein-coding genes for expression." In: *Genome research* 19 (10), pp. 1742–1751. DOI: 10.1101/gr.090951.109.

Bartel, D. P. (2009). "MicroRNAs: Target Recognition and Regulatory Functions". In: *Cell* 136.2, pp. 215–233. DOI: 10.1016/j.cell.2009.01.002.

Batagov, A. O. et al. (2011). "Identification of nucleotide patterns enriched in secreted RNAs as putative cis-acting elements targeting them to exosome nano-vesicles". In: *BMC genomics* 12.Suppl 3, S18. DOI: 10.1186/1471-2164-12-S3-S18.

Bedford, P. (1999). "MHC class II molecules transferred between allogeneic dendritic cells stimulate primary mixed leukocyte reactions". In: *Int Immunol* 11.11, pp. 1739–1744. DOI: 10.1093/intimm/11.11.1739.

Bejerano, G. et al. (2004). "Ultraconserved elements in the human genome". In: *Science (New York, NY)* 304.5675, pp. 1321–1325. DOI: 10.1126/science.1098119.

Bell, C. C. et al. (2016). "The Evx1/Evx1as gene locus regulates anterior-posterior patterning during gastrulation". In: *Sci Rep* 6, p. 26657. DOI: 10.1038/srep26657.

Benetatos, L. et al. (2011). "MEG3 imprinted gene contribution in tumorigenesis." In: *Int J Cancer* 129.4, pp. 773–779. DOI: 10.1002/ijc.26052.

Beninson, L. A. and M. Fleshner (2014). "Exosomes: An emerging factor in stress-induced immunomodulation". In: *Semin Immunol* 26.5, pp. 394–401. DOI: 10.1016/j.smim.2013.12.001.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal Statistical Society. Series B.* 57, pp. 289–300.

Bernstein, B. E. et al. (2006). "A bivalent chromatin structure marks key developmental genes in embryonic stem cells". In: *Cell* 125.2, pp. 315–326. DOI: 10.1016/j.cell.2006.02.041.

Bernstein, E. et al. (2003). "Dicer is essential for mouse development". In: *Nat Genet* 35.3, pp. 215–217. DOI: 10.1038/ng1253.

Bertone, P. et al. (2004). "Global identification of human transcribed sequences with genome tiling arrays." In: *Science (New York, NY)* 306.5705, pp. 2242–2246. DOI: 10.1126/science.1103388.

Bhatnagar, S. and J. S. Schorey (2007). "Exosomes released from infected macrophages contain Mycobacterium avium glycopeptidolipids and are proinflammatory." In: *The Journal of biological chemistry* 282.35, pp. 25779–25789. DOI: 10.1074/jbc.M702277200.

Bobrie, A. et al. (2011). "Exosome Secretion: Molecular Mechanisms and Roles in Immune Responses." In: *Traffic (Copenhagen, Denmark)*. DOI: 10.1111/j.1600-0854.2011.01225.x.

Bolukbasi, M. F. et al. (2012). "miR-1289 and |[ldquo]|Zipcode|[rdquo]|-like Sequence Enrich mRNAs in Microvesicles". In: *Mol Ther Nucleic Acids* 1.2, e10. DOI: 10.1038/mtna.2011.2.

Bootz, A. et al. (2004). "Comparison of scanning electron microscopy, dynamic light scattering and analytical ultracentrifugation for the sizing of poly(butyl cyanoacrylate) nanoparticles." In: *European journal of pharmaceutics and biopharmaceutics : official journal of Arbeitsgemeinschaft für Pharmazeutische Verfahrenstechnik e.V* 57.2, pp. 369–375. DOI: 10.1016/S0939-6411(03)00193-0.

Borchert, G. M. et al. (2006). "RNA polymerase III transcribes human microRNAs". In: *Nature Structural & Molecular Biology* 13.12, pp. 1097–1101. DOI: 10.1038/nsmb1167.

Borsani, G. et al. (1991). "Characterization of a murine gene expressed from the inactive X chromosome". In: *Nature* 351.6324, pp. 325–329. DOI: 10.1038/351325a0.

Braun, J. E. et al. (2012). "A direct interaction between DCP1 and XRN1 couples mRNA decapping to 5' exonucleolytic degradation." In: *Nat Struct Mol Biol* 19.12, pp. 1324–1331. DOI: 10.1038/nsmb.2413.

Brawand, D. et al. (2011). "The evolution of gene expression levels in mammalian organs." In: *Nature* 478.7369, pp. 343–348. DOI: 10.1038/nature10532.

Britten, R. J. and E. H. Davidson (1969). "Gene regulation for higher cells: a theory". In: *Science (New York, NY)* 165.891, pp. 349–357. DOI: 10.1126/science.165.3891.349.

Brockdorff, N. et al. (1991). "Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome." In: *Nature* 351 (6324), pp. 329–331. DOI: 10.1038/351329a0.

Brockdorff, N. et al. (1992). "The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus." In: *Cell* 71 (3), pp. 515–526.

Brown, J. A. et al. (2012). "Formation of triple-helical structures by the 3'-end sequences of MALAT1 and MENβ noncoding RNAs." In: *Proceedings of the National Academy of Sciences* 109.47, pp. 19202–19207. DOI: 10.1073/pnas.1217338109.

Brown, R. S. (2005). "Zinc finger proteins: getting a grip on RNA". In: *Curr Opin Struct Biol* 15.1, pp. 94–98.

Buck, A. H. et al. (2014). "Exosomes secreted by nematode parasites transfer small RNAs to mammalian cells and modulate innate immunity." In: *Nature communications* 5, p. 5488. DOI: 10.1038/ncomms6488.

Bukong, T. N. et al. (2014). "Exosomes from Hepatitis C Infected Patients Transmit HCV Infection and Contain Replication Competent Viral RNA in Complex with Ago2-miR122-HSP90". In: *PLoS Pathog* 10.10. Ed. by G. Luo, e1004424. DOI: 10.1371/journal.ppat.1004424.

Bussotti, G. et al. (2016). "Improved definition of the mouse transcriptome via targeted RNA sequencing." In: *Genome Res* 26.5, pp. 705–716. DOI: 10.1101/gr.199760.115.

Butovsky, O. et al. (2006). "Microglia activated by IL-4 or IFN-γ differentially induce neurogenesis and oligodendrogenesis from adult stem/progenitor cells". In: *Mol Cell Neurosci* 31.1, pp. 149–160. DOI: 10.1016/j.mcn.2005.10.006.

Cabili, M. N. et al. (2011). "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses". In: *Genes & Development* 25.18, pp. 1915–1927. DOI: 10.1101/gad.17446611.

Cabili, M. N. et al. (2015). "Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution." In: *Genome Biol* 16, p. 20. DOI: 10.1186/s13059-015-0586-4.

Cai, X. et al. (2004). "Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs". In: *RNA (New York, NY)* 10.12, pp. 1957–1966. DOI: 10.1261/rna.7135204.

Cao, Q.-L. et al. (2002). "Differentiation of engrafted neuronal-restricted precursor cells is inhibited in the traumatically injured spinal cord." In: *Exp Neurol* 177.2, pp. 349–359. DOI: 10.1006/exnr.2002.7981.

Carlevaro-Fita, J. et al. (2016). "Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells." In: *RNA (New York, N.Y.)* 22 (6), pp. 867–882. DOI: 10.1261/rna.053561.115.

Carninci, P. et al. (2005). "The transcriptional landscape of the mammalian genome." In: *Science (New York, NY)* 309.5740, pp. 1559–1563. DOI: 10.1126/science.1112014.

Carvalho, J. V. de et al. (2014). "Nef Neutralizes the Ability of Exosomes from CD4+ T Cells to Act as Decoys during HIV-1 Infection." In: *PLoS One* 9.11, e113691. DOI: 10.1371/journal.pone.0113691.

Cattanach, B. and M. Kirk (1985). "Differential activity of maternally and paternally derived chromosome regions in mice." In: *Nature* 315.6019, pp. 496–498. DOI: 10.1038/315496a0.

Cavalli, G. and T. Misteli (2013). "Functional implications of genome topology". In: *Nat Struct Mol Biol* 20.3, pp. 290–299. DOI: 10.1038/nsmb.2474.

Cerase, A. et al. (2015). "Xist localization and function: new insights from multiple levels." In: *Genome Biol* 16, p. 166. DOI: 10.1186/s13059-015-0733-y.

Chairoungdua, A. et al. (2010). "Exosome release of β-catenin: a novel mechanism that antagonizes Wnt signaling." In: *The Journal of cell biology* 190 (6), pp. 1079–1091. DOI: 10.1083/jcb.201002049.

Chaumeil, J. et al. (2006). "A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced". In: *Genes & Development* 20.16, pp. 2223–2237. DOI: 10.1101/gad.380906.

Chawla, G. and N. S. Sokol (2014). "ADAR mediates differential expression of polycistronic microRNAs." In: *Nucleic Acids Res* 42.8, pp. 5245–5255. DOI: 10.1093/nar/gku145.

Chen, Q. et al. (2016). "Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder." In: *Science (New York, N.Y.)* 351 (6271), pp. 397–400. DOI: 10.1126/science.aad7977.

Chen, Y. et al. (2014). "A DDX6-CNOT1 complex and W-binding pockets in CNOT9 reveal direct links between miRNA target recognition and silencing." In: *Mol Cell* 54.5, pp. 737–750. DOI: 10.1016/j.molcel.2014.03.034.

Chevillet, J. R. et al. (2014). "Quantitative and stoichiometric analysis of the microRNA content of... - PubMed - NCBI". In: *Proc Natl Acad Sci U S A* 111.41, pp. 14888–14893. DOI: 10.1073/pnas.1408301111.

Chodroff, R. A. et al. (2010). "Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes." In: *Genome Biol* 11.7, R72. DOI: 10.1186/gb-2010-11-7-r72.

Chooniedass-Kothari, S. et al. (2004). "The steroid receptor RNA activator is the first functional RNA encoding a protein". In: *FEBS Lett* 566.1-3, pp. 43–47. DOI: 10.1016/j.febslet.2004.03.104.

Christie, M. et al. (2013). "Structure of the PAN3 pseudokinase reveals the basis for interactions with the PAN2 deadenylase and the GW182 proteins." In: *Mol Cell* 51.3, pp. 360–373. DOI: 10.1016/j.molcel.2013.07.011.

Chu, C. et al. (2015). "Systematic discovery of Xist RNA binding proteins." In: *Cell* 161.2, pp. 404–416. DOI: 10.1016/j.cell.2015.03.025.

Chu, K. et al. (2004). "Human neural stem cells improve sensorimotor deficits in the adult rat brain with experimental focal ischemia." In: *Brain Res* 1016.2, pp. 145–153. DOI: 10.1016/j.brainres.2004.04.038.

Chureau, C. et al. (2002). "Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine". In: *Genome Res* 12.6, pp. 894–908.

Clark, M. B. et al. (2015). "Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing." In: *Nat Methods*. DOI: 10.1038/nmeth.3321.

Clayton, A. (2005). "Induction of heat shock proteins in B-cell exosomes". In: *J Cell Sci* 118.16, pp. 3631–3638. DOI: 10.1242/jcs.02494.

Clayton, A. et al. (2007). "Human Tumor-Derived Exosomes Selectively Impair Lymphocyte Responses to Interleukin-2". In: *Cancer Res* 67.15, pp. 7458–7466. DOI: 10.1158/0008-5472.CAN-06-3456.

Colombo, M. et al. (2014). "Biogenesis, Secretion, and Intercellular Interactions of Exosomes and Other Extracellular Vesicles". In: *Annual Review of Cell and Developmental Biology* 30.1, pp. 255–289. DOI: 10.1146/annurev-cellbio-101512-122326.

Comings, D. E. (1972). "The Structure and Function of Chromatin". In: *Advances in Human Genetics*. Boston, MA: Springer US, pp. 237–431. DOI: 10.1007/978-1-4757-4429-3_5.

Corcoran, D. L. et al. (2009). "Features of Mammalian microRNA Promoters Emerge from Polymerase II Chromatin Immunoprecipitation Data". In: *PLoS One* 4.4. Ed. by C. K. Patil, e5279–10. DOI: 10.1371/journal.pone.0005279.

Cornish-Bowden, A. (1985). "Nomenclature for incompletely specified bases in nucleic acid sequences: rcommendations 1984". In: *Nucl Acids Res* 13.9, pp. 3021–3030. DOI: 10.1093/nar/13.9.3021.

Cossetti, C. et al. (2012). "Extracellular membrane vesicles and immune regulation in the brain." In: *Front Physiol* 3, p. 117. DOI: 10.3389/fphys.2012.00117.

Cossetti, C. et al. (2014a). "Extracellular vesicles from neural stem cells transfer IFN-γ via Ifngr1 to activate Stat1 signaling in target cells." In: *Mol Cell* 56.2, pp. 193–204. DOI: 10.1016/j.molcel.2014.08.020.

Cossetti, C. et al. (2014b). "Soma-to-germline transmission of RNA in mice xenografted with human... - PubMed - NCBI". In: *PLoS One* 9.7, e101629. DOI: 10.1371/journal.pone.0101629.

Couchman, J. R. (2010). "Transmembrane signaling proteoglycans." In: *Annual review of cell and developmental biology* 26, pp. 89–114. DOI: 10.1146/annurev-cellbio-100109-104126.

Dai, L. et al. (2016). "Cytoplasmic Drosha activity generated by alternative splicing." In: *Nucleic Acids Res*. DOI: 10.1093/nar/gkw668.

Dallosso, A. R. et al. (2007). "Alternately spliced WT1 antisense transcripts interact with WT1 sense RNA and show epigenetic and splicing defects in cancer." In: *RNA (New York, NY)* 13.12, pp. 2287–2299. DOI: 10.1261/rna.562907.

Davis, M. P. A. et al. (2013). "Kraken: a set of tools for quality control and analysis of high-throughput sequence data." In: *Methods (San Diego, Calif.)* 63.1, pp. 41–49. DOI: 10.1016/j.ymeth.2013.06.027.

De Santa, F. et al. (2010). "A large fraction of extragenic RNA pol II transcription sites overlap enhancers". In: *PLoS Biol* 8.5, e1000384. DOI: 10.1371/journal.pbio.1000384.

Deatherage, B. L. and B. T. Cookson (2012). "Membrane vesicle release in bacteria, eukaryotes, and archaea: a conserved yet underappreciated aspect of microbial life." In: *Infect Immun* 80.6, pp. 1948–1957. DOI: 10.1128/IAI.06014-11.

Denli, A. M. et al. (2004). "Processing of primary microRNAs by the Microprocessor complex." In: *Nature* 432.7014, pp. 231–235. DOI: 10.1038/nature03049.

Deregibus, M. C. et al. (2007). "Endothelial progenitor cell derived microvesicles activate an angiogenic program in endothelial cells by a horizontal transfer of mRNA". In: *Blood* 110.7, pp. 2440–2448. DOI: 10.1182/blood-2007-03-078709.

Derrien, T. et al. (2012). "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression." In: *Genome Res* 22.9, pp. 1775–1789. DOI: 10.1101/gr.132159.111.

Dhir, A. et al. (2015). "Microprocessor mediates transcriptional termination of long noncoding RNA transcripts hosting microRNAs." In: *Nature structural & molecular biology* 22 (4), pp. 319–327. DOI: 10.1038/nsmb.2982.

Dieci, G. et al. (2007). "The expanding RNA polymerase III transcriptome". In: *Trends in genetics : TIG* 23.12, pp. 614–622. DOI: 10.1016/j.tig.2007.09.001.

Dinger, M. E. et al. (2008a). "Differentiating protein-coding and noncoding RNA: challenges and ambiguities". In: *PLoS Comput Biol* 4.11. Ed. by J. McEntyre, e1000176. DOI: 10.1371/journal.pcbi.1000176.

Dinger, M. E. et al. (2008b). "Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation". In: *Genome Res* 18.9, pp. 1433–1445. DOI: 10.1101/gr.078378.108.

Dinger, M. E. et al. (2009). "Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications". In: *Briefings in functional genomics & proteomics* 8.6, pp. 407–423. DOI: 10.1093/bfgp/elp038.

Dinger, M. E. et al. (2011). "The evolution of RNAs with multiple functions". In: *Biochimie*. DOI: 10.1016/j.biochi.2011.07.018.

Djebali, S. et al. (2012). "Landscape of transcription in human cells". In: *Nature* 489.7414, pp. 101–108. DOI: 10.1038/nature11233.

Doench, J. G. and P. A. Sharp (2004). "Specificity of microRNA target selection in translational repression". In: *Genes & Development* 18.5, pp. 504–511.

Doetsch, F. et al. (1999). "Subventricular Zone Astrocytes Are Neural Stem Cells in the Adult Mammalian Brain". In: *Cell* 97.6, pp. 703–716. DOI: 10.1016/S0092-8674(00)80783-7.

Down, T. A. et al. (2011). "Dalliance: interactive genome viewing on the web". In: *Bioinformatics* 27.6, pp. 889–890. DOI: 10.1093/bioinformatics/btr020.

Down, T. A. and T. J. P. Hubbard (2002). "Computational detection and location of transcription start sites in mammalian genomic DNA." In: *Genome Res* 12.3, pp. 458–461. DOI: 10.1101/gr.216102.

Dragovic, R. A. et al. (2011). "Sizing and phenotyping of cellular vesicles using Nanoparticle Tracking Analysis." In: *Nanomedicine : nanotechnology, biology, and medicine* 7.6, pp. 780–788. DOI: 10.1016/j.nano.2011.04.003.

Dunning, M. J. et al. (2007). "beadarray: R classes and methods for Illumina bead-based data". In: *Bioinformatics* 23.16, pp. 2183–2184. DOI: 10.1093/bioinformatics/btm311.

Durbin, J. E. et al. (1996). "Targeted disruption of the mouse Stat1 gene results in compromised innate immunity to viral disease". In: *Cell*. DOI: 10.1016/s0092-8674(00)81289-1.

Easow, G. et al. (2007). "Isolation of microRNA targets by miRNP immunopurification." In: *RNA (New York, N.Y.)* 13 (8), pp. 1198–1204. DOI: 10.1261/rna.563707.

Edelman, G. M. and J. A. Gally (1970). *Arrangement and evolution of eukaryotic genes*. The neurosciences: Second study program.

Eder, C. (2009). "Mechanisms of interleukin-1β release". In: *Immunobiology* 214.7, pp. 543–553. DOI: 10.1016/j.imbio.2008.11.007.

Ekdahl, C. T. et al. (2011). "Inflammation is detrimental for neurogenesis in adult brain". In: *Proc Natl Acad Sci U S A* 100.23, pp. 13632–13637. DOI: 10.1073/pnas.2234031100.

Ekström, E. J. et al. (2014). "WNT5A induces release of exosomes containing pro-angiogenic and immunosuppressive factors from malignant melanoma cells". In: *Mol Cancer* 13.1, p. 88. DOI: 10.1186/1476-4598-13-88.

Elbashir, S. M. et al. (2001). "Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells". In: *Nature* 411, pp. 494–498.

Eldh, M. et al. (2010). "Exosomes Communicate Protective Messages during Oxidative Stress; Possible Role of Exosomal Shuttle RNA". In: *PLoS One* 5.12, e15353. DOI: 10.1371/journal.pone.0015353.

Emmanouilidou, E. et al. (2010). "Cell-produced alpha-synuclein is secreted in a calcium-dependent manner by exosomes and impacts neuronal survival." In: *J Neurosci* 30.20, pp. 6838–6851. DOI: 10.1523/JNEUROSCI.5699-09.2010.

ENCODE Project Consortium et al. (2012). "An integrated encyclopedia of DNA elements in the human genome." In: *Nature* 489.7414, pp. 57–74. DOI: 10.1038/nature11247.

Engreitz, J. M. et al. (2013). "The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome". In: *Science (New York, NY)* 341.6147, p. 1237973. DOI: 10.1126/science.1237973.

Engreitz, J. M. et al. (2016). "Neighborhood regulation by lncRNA promoters, transcription, and splicing". In: *bioRxiv*. DOI: 10.1101/050948. eprint: http://biorxiv.org/content/early/2016/04/28/050948.full.pdf.

Engstrom, P. G. et al. (2006). "Complex loci in human and mouse genomes". In: *PLos Genet* 2.4, e47. DOI: 10.1371/journal.pgen.0020047.

Enright, A. J. et al. (2003). "MicroRNA targets in Drosophila." In: *Genome Biol* 5.1, R1. DOI: 10.1186/gb-2003-5-1-r1.

Ernst, J. and M. Kellis (2010). "Discovery and characterization of chromatin states for systematic annotation of the human genome". In: *Nat Biotechnol* 28.8, pp. 817–825. DOI: 10.1038/nbt.1662.

– (2012). "ChromHMM: automating chromatin-state discovery and characterization". In: *Nat Methods* 9.3, pp. 215–216. DOI: 10.1038/nmeth.1906.

Ernst, J. et al. (2011). "Mapping and analysis of chromatin state dynamics in nine human cell types". In: *Nature* 473.7345, pp. 43–49. DOI: 10.1038/nature09906.

Fabbri, M. et al. (2012). "MicroRNAs bind to Toll-like receptors to induce prometastatic inflammatory response." In: *Proceedings of the National Academy of Sciences*. DOI: 10.1073/pnas.1209414109.

Fabian, M. R. et al. (2011). "miRNA-mediated deadenylation is orchestrated by GW182 through two conserved motifs that interact with CCR4-NOT." In: *Nat Struct Mol Biol* 18.11, pp. 1211–1217. DOI: 10.1038/nsmb.2149.

Fang, W. and D. P. Bartel (2015). "The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes." In: *Molecular cell* 60 (1), pp. 131–145. DOI: 10.1016/j.molcel.2015.08.015.

FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. (2014). "A promoter-level mammalian expression atlas." In: *Nature* 507.7493, pp. 462–470. DOI: 10.1038/nature13182.

Fellous, M. et al. (1981). "Interferon enhances the amount of membrane-bound beta2-microglobulin and its release from human Burkitt cells." In: *Eur J Immunol* 11.6, pp. 524–526. DOI: 10.1002/eji.1830110616.

Feng, J. et al. (2006). "The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator." In: *Genes & Development* 20.11, pp. 1470–1484. DOI: 10.1101/gad.1416106.

Fevrier, B. et al. (2004). "Cells release prions in association with exosomes." In: *Proc Natl Acad Sci U S A* 101.26, pp. 9683–9688. DOI: 10.1073/pnas.0308413101.

Filipe, V. et al. (2010). "Critical evaluation of Nanoparticle Tracking Analysis (NTA) by NanoSight for the measurement of nanoparticles and protein aggregates." In: *Pharm Res* 27.5, pp. 796–810. DOI: 10.1007/s11095-010-0073-2.

Filipowicz, W. and N. Sonenberg (2015). "The long unfinished march towards understanding microRNA-mediated repression." In: *RNA* 21.4, pp. 519–524. DOI: 10.1261/rna.051219.115.

Forman, J. J. et al. (2008). "A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence." In: *Proceedings of the National Academy of Sciences of the United States of America* 105 (39), pp. 14879–14884. DOI: 10.1073/pnas.0803230105.

Friedman, R. C. et al. (2009). "Most mammalian mRNAs are conserved targets of microRNAs". In: *Genome Res* 19.1, pp. 92–105. DOI: 10.1101/gr.082701.108.

Fromm, S. A. et al. (2012). "The structural basis of Edc3- and Scd6-mediated activation of the Dcp1:Dcp2 mRNA decapping complex." In: *The EMBO journal* 31 (2), pp. 279–290. DOI: 10.1038/emboj.2011.408.

Fujimoto, S. (1973). "On the Golgi-derived vesicles in the rabbit taste bud cells: an electron microscopy and related cytochemistry". In: *The Kurume Medical Journal* 20.3, pp. 133–148. DOI: 10.2739/kurumemedj.20.133.

Fujiwara, Y. et al. (2004). "Intravenously injected neural progenitor cells of transgenic rats can migrate to the injured spinal cord and differentiate into neurons, astrocytes and oligodendrocytes." In: *Neurosci Lett* 366.3, pp. 287–291. DOI: 10.1016/j.neulet.2004.05.080.

Fuster-Matanzo, A. et al. (2015). "Acellular approaches for regenerative medicine: on the verge of clinical trials with extracellular membrane vesicles?" In: *Stem Cell Res Ther* 6.1, p. 227. DOI: 10.1186/s13287-015-0232-9.

Gage, F. H. (2000). "Mammalian neural stem cells". In: *Science (New York, NY)* 287.5457, pp. 1433–1438. DOI: 10.1126/science.287.5457.1433.

Galante, P. A. F. et al. (2007). "Sense-antisense pairs in mammals: functional and evolutionary considerations." In: *Genome Biol* 8.3, R40. DOI: 10.1186/gb-2007-8-3-r40.

Gardiner-Garden, M. and M. Frommer (1987). "CpG islands in vertebrate genomes". In: *J Mol Biol* 196.2, pp. 261–282. DOI: 10.1016/0022-2836(87)90689-9.

Gascoigne, D. K. et al. (2012). "Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes." In: *Bioinformatics* 28.23, pp. 3042–3050. DOI: 10.1093/bioinformatics/bts582.

Gastpar, R. (2005). "Heat Shock Protein 70 Surface-Positive Tumor Exosomes Stimulate Migratory and Cytolytic Activity of Natural Killer Cells". In: *Cancer Res* 65.12, pp. 5238–5247. DOI: 10.1158/0008-5472.CAN-04-3804.

Geiss, G. K. et al. (2008). "Direct multiplexed measurement of gene expression with color-coded probe pairs". In: *Nat Biotechnol* 26.3, pp. 317–325. DOI: 10.1038/nbt1385.

Gene Ontology Consortium (2015). "Gene Ontology Consortium: going forward." In: *Nucleic acids research* 43 (Database issue), pp. D1049–D1056. DOI: 10.1093/nar/gku1179.

Georgakilas, G. et al. (2014). "microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs." In: *Nat Commun* 5, p. 5700. DOI: 10.1038/ncomms6700.

Gerstein, M. B. et al. (2007). "What is a gene, post-ENCODE? History and updated definition". In: *Genome Res* 17.6, pp. 669–681. DOI: 10.1101/gr.6339607.

Gibbings, D. J. et al. (2009). "Multivesicular bodies associate with components of miRNA effector complexes and modulate miRNA activity". In: *Nat Cell Biol* 11.9, pp. 1143–1149. DOI: 10.1038/ncb1929.

Giraldez, A. J. et al. (2005). "MicroRNAs regulate brain morphogenesis in zebrafish". In: *Science (New York, NY)* 308.5723, pp. 833–838.

Giraldez, A. J. et al. (2006). "Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs." In: *Science (New York, NY)* 312.5770, pp. 75–79. DOI: 10.1126/science.1122689.

Gordon, H. et al. (2014). "Depletion of hnRNP A2/B1 overrides the nuclear retention of the HIV... - PubMed - NCBI". In: *RNA Biol* 10.11, pp. 1714–1725. DOI: 10.4161/rna.26542.

Gould, S. J. and G. Raposo (2013). "As we wait: coping with an imperfect nomenclature for extracellular vesicles". In: *Journal of extracellular vesicles* 2, p. 2892. DOI: 10.3402/jev.v2i0.20389.

Gouwy, M. et al. (2005). "Synergy in cytokine and chemokine networks amplifies the inflammatory response." In: *Cytokine & growth factor reviews* 16.6, pp. 561–580. DOI: 10.1016/j.cytogfr.2005.03.005.

Gregory, R. I. et al. (2004). "The Microprocessor complex mediates the genesis of microRNAs." In: *Nature* 432.7014, pp. 235–240. DOI: 10.1038/nature03120.

Griffiths-Jones, S. et al. (2008). "miRBase: tools for microRNA genomics." In: *Nucleic Acids Res* 36.Database issue, pp. D154–8. DOI: 10.1093/nar/gkm952.

Grishok, A. et al. (2001). "Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing". In: *Cell* 106.1, pp. 23–34. DOI: 10.1016/s0092-8674(01)00431-7.

Guerrier-Takada, C. et al. (1983). "The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme." In: *Cell* 35.3 Pt 2, pp. 849–857. DOI: 10.1016/0092-8674(83)90117-4.

Guil, S. and M. Esteller (2012). "Cis-acting noncoding RNAs: friends and foes." In: *Nature Structural & Molecular Biology* 19.11, pp. 1068–1075. DOI: 10.1038/nsmb.2428.

Guo, H. et al. (2010). "Mammalian microRNAs predominantly act to decrease target mRNA levels". In: *Nature* 466.7308, pp. 835–840. DOI: 10.1038/nature09267.

Gupta, R. A. et al. (2010). "Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis". In: *Nature* 464.7291, pp. 1071–1076. DOI: 10.1038/nature08975.

Gupta, S. and A. A. Knowlton (2007). "HSP60 trafficking in adult cardiac myocytes: role of the exosomal pathway". In: *AJP: Heart and Circulatory Physiology* 292.6, H3052–H3056. DOI: 10.1152/ajpheart.01355.2006.

Guttman, M. et al. (2009). "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals." In: *Nature* 458.7235, pp. 223–227. DOI: 10.1038/nature07672.

Guttman, M. et al. (2013). "Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins." In: *Cell* 154 (1), pp. 240–251. DOI: 10.1016/j.cell.2013.06.009.

Guttman, M. et al. (2010). "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs". In: *Nat Biotechnol* 28.5, pp. 503–510. DOI: 10.1038/nbt.1633.

Ha, M. and V. N. Kim (2014). "Regulation of microRNA biogenesis." In: *Nat Rev Mol Cell Biol* 15.8, pp. 509–524. DOI: 10.1038/nrm3838.

Hacisuleyman, E. et al. (2014). "Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre". In: *Nat Struct Mol Biol* 21.2, pp. 198–206. DOI: 10.1038/nsmb.2764.

Haerty, W. and C. P. Ponting (2015). "Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci". In: *RNA (New York, NY)*. DOI: 10.1261/rna.047324.114.

Hafner, M. et al. (2010). "Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP." In: *Cell* 141 (1), pp. 129–141. DOI: 10.1016/j.cell.2010.03.009.

Hagan, J. P. et al. (2009). "Lin28 recruits the TUTase Zcchc11 to inhibit let-7 maturation in mouse embryonic stem cells." In: *Nature structural & molecular biology* 16 (10), pp. 1021–1025. DOI: 10.1038/nsmb.1676.

Han, J. et al. (2004). "The Drosha-DGCR8 complex in primary microRNA processing." In: *Genes Dev* 18.24, pp. 3016–3027. DOI: 10.1101/gad.1262504.

Han, J. et al. (2006). "Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex". In: *Cell* 125.5, pp. 887–901. DOI: 10.1016/j.cell.2006.03.043.

Hansen, T. B. et al. (2013). "Natural RNA circles function as efficient microRNA sponges." In: *Nature* 495.7441, pp. 384–388. DOI: 10.1038/nature11993.

Harmston, N. et al. (2016). "Topologically associated domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation". In: *bioRxiv*. DOI: 10.1101/042952. eprint: http://biorxiv.org/content/early/2016/08/28/042952.full.pdf.

Harrow, J. et al. (2006). "GENCODE: producing a reference annotation for ENCODE". In: *Genome Biol* 7 Suppl 1, S4 1–9. DOI: 10.1186/gb-2006-7-s1-s4.

Harrow, J. et al. (2012). "GENCODE: the reference human genome annotation for The ENCODE Project". In: *Genome Res* 22.9, pp. 1760–1774. DOI: 10.1101/gr.135350.111.

Hausser, J. and M. Zavolan (2014). "Identification and consequences of miRNA-target interactions–beyond repression of gene expression." In: *Nat Rev Genet* 15.9, pp. 599–612. DOI: 10.1038/nrg3765.

Hawari, F. I. et al. (2004). "Release of full-length 55-kDa TNF receptor 1 in exosome-like vesicles: A mechanism for generation of soluble cytokine receptors". In: *Proc Natl Acad Sci U S A* 101.5, pp. 1297–1302. DOI: 10.1073/pnas.0307981100.

Hawkes, E. J. et al. (2016). "COOLAIR Antisense RNAs Form Evolutionarily Conserved Elaborate Secondary Structures." In: *Cell reports* 16 (12), pp. 3087–3096. DOI: 10.1016/j.celrep.2016.08.045.

Heidari, N. et al. (2014). "Genome-wide map of regulatory interactions in the human genome". In: *Genome Res* 24.12, pp. 1905–1917. DOI: 10.1101/gr.176586.114.

Heijnen, H. F. et al. (1999). "Activated platelets release two types of membrane vesicles: microvesicles by surface shedding and exosomes derived from exocytosis of multivesicular bodies and alpha-granules." In: *Blood* 94.11, pp. 3791–3799.

Heinz, S. et al. (2010). "Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities". In: *Mol Cell* 38.4, pp. 576–589. DOI: 10.1016/j.molcel.2010.05.004.

Hendrickson, D. G. et al. (2009). "Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA." In: *PLoS Biol* 7.11, e1000238. DOI: 10.1371/journal.pbio.1000238.

Heo, I. et al. (2008). "Lin28 mediates the terminal uridylation of let-7 precursor MicroRNA." In: *Molecular cell* 32 (2), pp. 276–284. DOI: 10.1016/j.molcel.2008.09.014.

Hezroni, H. et al. (2015). "Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species." In: *Cell Reports* 11.7, pp. 1110–1122. DOI: 10.1016/j.celrep.2015.04.023.

Hobson, D. J. et al. (2012). "RNA polymerase II collision interrupts convergent transcription." In: *Mol Cell* 48.3, pp. 365–374. DOI: 10.1016/j.molcel.2012.08.027.

Holliday, R. (1970). *The organization of DNA in eukaryotic chromosomes*. Symp. Soc. Gen. Microbiol.

Hsieh, C.-L. et al. (2014). "Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation." In: *Proc Natl Acad Sci U S A* 111.20, pp. 7319–7324. DOI: 10.1073/pnas.1324151111.

Huang, S. et al. (1993). "Immune response in mice that lack the interferon-gamma receptor". In: … 259.5102, pp. 1742–1745. DOI: 10.1126/science.8456301.

Huang, W. et al. (2015). "DDX5 and its associated lncRNA Rmrp modulate TH17 cell effector functions". In: *Nature* 528.7583, pp. 517–522. DOI: 10.1038/nature16193.

Huber, W. et al. (2015). "Orchestrating high-throughput genomic analysis with Bioconductor." In: *Nat Methods* 12.2, pp. 115–121. DOI: 10.1038/nmeth.3252.

Humphreys, D. T. et al. (2005). "MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function." In: *Proc Natl Acad Sci U S A* 102.47, pp. 16961–16966. DOI: 10.1073/pnas.0506482102.

Hung, T. et al. (2011). "Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters". In: *Nat Genet* 43.7, pp. 621–629. DOI: 10.1038/ng.848.

Hurley, J. H. and P. I. Hanson (2010). "Membrane budding and scission by the ESCRT machinery: it's all in the neck." In: *Nature reviews. Molecular cell biology* 11 (8), pp. 556–566. DOI: 10.1038/nrm2937.

Hutvagner, G. et al. (2001). "A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA". In: *Science (New York, NY)* 12, p. 12. DOI: 10.1126/science.1062961.

Ingolia, N. T. et al. (2014). "Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes." In: *Cell reports* 8 (5), pp. 1365–1379. DOI: 10.1016/j.celrep.2014.07.045.

Iraci, N. et al. (2016). "Focus on Extracellular Vesicles: Physiological Role and Signalling Properties of Extracellular Membrane Vesicles." In: *Int J Mol Sci* 17.2, p. 171. DOI: 10.3390/ijms17020171.

Ismail, N. et al. (2013). "Macrophage microvesicles induce macrophage differentiation and miR-223 transfer." In: *Blood* 121.6, pp. 984–995. DOI: 10.1182/blood-2011-08-374793.

Iyer, M. K. et al. (2015). "The landscape of long noncoding RNAs in the human transcriptome." In: *Nature Publishing Group* 47.3, pp. 199–208. DOI: 10.1038/ng.3192.

Jeong, S.-W. et al. (2003). "Human neural stem cell transplantation promotes functional recovery in rats with experimental intracerebral hemorrhage." In: *Stroke; a journal of cerebral circulation* 34.9, pp. 2258–2263. DOI: 10.1161/01.STR.0000083698.20199.1F.

Jeong, O.-S. et al. (2016). "Long noncoding RNA linc00598 regulates CCND2 transcription and modulates the G1 checkpoint." In: *Scientific reports* 6, p. 32172. DOI: 10.1038/srep32172.

Jia, H. et al. (2010). "Genome-wide computational identification and manual annotation of human long noncoding RNA genes". In: *RNA* 16.8, pp. 1478–1487. DOI: 10.1261/rna.1951310.

Jiang, W. et al. (2015). "The lncRNA DEANR1 facilitates human endoderm differentiation by activating FOXA2 expression." In: *Cell Reports* 11.1, pp. 137–148. DOI: 10.1016/j.celrep.2015.03.008.

Johnstone, R. M. et al. (1987). "Vesicle formation during reticulocyte maturation. Association of plasma membrane activities with released vesicles (exosomes)." In: *The Journal of biological chemistry* 262.19, pp. 9412–9420.

Jolma, A. et al. (2013). "DNA-binding specificities of human transcription factors". In: *Cell* 152.1-2, pp. 327–339. DOI: 10.1016/j.cell.2012.12.009.

Jonas, S. and E. Izaurralde (2015). "Towards a molecular understanding of microRNA-mediated gene silencing." In: *Nat Rev Genet* 16.7, pp. 421–433. DOI: 10.1038/nrg3965.

Jong, O. G. de et al. (2012). "Cellular stress conditions are reflected in the protein and RNA content of endothelial cell-derived exosomes". In: *Journal of extracellular vesicles* 1, p. 569. DOI: 10.3402/jev.v1i0.18396.

Kalamvoki, M. et al. (2014). "Cells infected with herpes simplex virus 1 export to uninfected cel... - PubMed - NCBI". In: *Proc Natl Acad Sci U S A* 111.46, E4991–E4996. DOI: 10.1073/pnas.1419338111.

Kalra, H. et al. (2012). "Vesiclepedia: a compendium for extracellular vesicles with continuous community annotation." In: *PLoS Biol* 10.12, e1001450. DOI: 10.1371/journal.pbio.1001450.

Kaneko, S. et al. (2010). "Phosphorylation of the PRC2 component Ezh2 is cell cycle-regulated and up-regulates its binding to ncRNA." In: *Genes Dev* 24.23, pp. 2615–2620. DOI: 10.1101/gad.1983810.

Kanhere, A. et al. (2010). "Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2". In: *Mol Cell* 38.5, pp. 675–688. DOI: 10.1016/j.molcel.2010.03.019.

Kapranov, P. et al. (2007). "RNA maps reveal new RNA classes and a possible function for pervasive transcription." In: *Science (New York, NY)* 316.5830, pp. 1484–1488. DOI: 10.1126/science.1138341.

Kapranov, P. et al. (2005). "Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays". In: *Genome Res* 15.7, pp. 987–997. DOI: 10.1101/gr.3455305.

Karolchik, D. et al. (2004). "The UCSC Table Browser data retrieval tool". In: *Nucleic Acids Res* 32.Database issue, pp. D493–6. DOI: 10.1093/nar/gkh103.

Katayama, S. et al. (2005). "Antisense transcription in the mammalian transcriptome". In: *Science (New York, NY)* 309.5740, pp. 1564–1566. DOI: 10.1126/science.1112009.

Kato, T. et al. (2014). "Exosomes from IL-1β stimulated synovial fibroblasts induce osteoarthritic changes in articular chondrocytes". In: *Arthritis research & therapy* 16.4, R163. DOI: 10.1186/ar4679.

Kawahara, Y. et al. (2007). "RNA editing of the microRNA-151 precursor blocks cleavage by the Dicer-TRBP complex". In: *EMBO Rep* 8.8, pp. 763–769. DOI: 10.1038/sj.embor.7401011.

Kawamata, T. et al. (2009). "Structural determinants of miRNAs for RISC loading and slicer-independent unwinding." In: *Nat Struct Mol Biol* 16.9, pp. 953–960. DOI: 10.1038/nsmb.1630.

Kent, W. J. (2002). "BLAT–the BLAST-like alignment tool". In: *Genome Res* 12.4, pp. 656–664. DOI: 10.1101/gr.229202..

Kertesz, M. et al. (2010). "Genome-wide measurement of RNA secondary structure in yeast". In: *Nature* 467.7311, pp. 103–107. DOI: 10.1038/nature09322.

Khalil, A. M. et al. (2009). "Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression." In: *Proceedings of the National Academy of Sciences* 106.28, pp. 11667–11672. DOI: 10.1073/pnas.0904715106.

Kheradpour, P. and M. Kellis (2014). "Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments". In: *Nucleic Acids Res* 42.5, pp. 2976–2987. DOI: 10.1093/nar/gkt1249.

Khvorova, A. et al. (2003). "Functional siRNAs and miRNAs exhibit strand bias". In: *Cell* 115.2, pp. 209–216. DOI: 10.1016/s0092-8674(03)00801-8.

Kim, D.-K. et al. (2015a). "EVpedia: A community web resource for prokaryotic and eukaryotic extracellular vesicles research." In: *Semin Cell Dev Biol* 40, pp. 4–7. DOI: 10.1016/j.semcdb.2015.02.005.

Kim, D. et al. (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". In: *Genome Biol* 14.4, R36. DOI: 10.1186/gb-2013-14-4-r36.

Kim, T. H. et al. (2005). "A high-resolution map of active promoters in the human genome". In: *Nature* 436.7052, pp. 876–880. DOI: 10.1038/nature03877.

Kim, T. K. et al. (2010). "Widespread transcription at neuronal activity-regulated enhancers". In: *Nature* 465.7295, pp. 182–187. DOI: 10.1038/nature09033.

Kim, Y. W. et al. (2015b). "Chromatin looping and eRNA transcription precede the transcriptional activation of gene in the beta-globin locus". In: *Biosci Rep* 35.2. DOI: 10.1042/BSR20140126.

Kim, Y.-K. et al. (2016). "Re-evaluation of the roles of DROSHA, Export in 5, and DICER in microRNA biogenesis." In: *Proc Natl Acad Sci U S A* 113.13, E1881–E1889. DOI: 10.1073/pnas.1602532113.

Kino, T. et al. (2010). "Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor". In: *Sci Signal* 3.107, ra8. DOI: 10.1126/scisignal.2000568.

Klumperman, J. and G. Raposo (2014). "The complex ultrastructure of the endolysosomal system." In: *Cold Spring Harbor perspectives in biology* 6 (10), a016857. DOI: 10.1101/cshperspect.a016857.

Knuesel, M. T. et al. (2009). "The human CDK8 subcomplex is a histone kinase that requires Med12 for activity and can function independently of mediator." In: *Mol Cell Biol* 29.3, pp. 650–661. DOI: 10.1128/MCB.00993-08.

Koch, F. et al. (2011). "Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters." In: *Nat Struct Mol Biol* 18.8, pp. 956–963. DOI: 10.1038/nsmb.2085.

Kohlmaier, A. et al. (2004). "A chromosomal memory triggered by Xist regulates histone methylation in X inactivation." In: *PLoS Biol* 2.7, E171. DOI: 10.1371/journal.pbio.0020171.

Koppers-Lalic, D. et al. (2014). "Nontemplated Nucleotide Additions Distinguish the Small RNA Composition in Cells from Exosomes". In: *Cell Reports* 8.6, pp. 1649–1658. DOI: 10.1016/j.celrep.2014.08.027.

Kore, R. A. and E. C. Abraham (2014). "Inflammatory cytokines, interleukin-1 beta and tumor necrosis facto... - PubMed - NCBI". In: *Biochem Biophys Res Commun* 453.3, pp. 326–331. DOI: 10.1016/j.bbrc.2014.09.068.

Kozomara, A. and S. Griffiths-Jones (2011). "miRBase: integrating microRNA annotation and deep-sequencing data." In: *Nucleic acids research* 39 (Database issue), pp. D152–D157. DOI: 10.1093/nar/gkq1027.

Kuehn, M. J. and N. C. Kesty (2005). "Bacterial outer membrane vesicles and the host-pathogen interaction." In: *Genes & Development* 19.22, pp. 2645–2655. DOI: 10.1101/gad.1299905.

Kutter, C. et al. (2012). "Rapid turnover of long noncoding RNAs and the evolution of gene expression." In: *PLos Genet* 8.7, e1002841. DOI: 10.1371/journal.pgen.1002841.

Kwon, S. C. et al. (2016). "Structure of Human DROSHA." In: *Cell* 164.1-2, pp. 81–90. DOI: 10.1016/j.cell.2015.12.019.

Kwon, S. H. et al. (2014). "Intercellular transfer of GPRC5B via exosomes drives HGF-mediated outward growth". In: *Curr Biol*. DOI: 10.1016/j.cub.2013.12.010.

Lagos-Quintana, M. et al. (2001). "Identification of novel genes coding for small expressed RNAs". In: *Science (New York, NY)* 294.5543, pp. 853–858. DOI: 10.1126/science.1064921.

Lam, M. T. Y. et al. (2014). "Enhancer RNAs and regulated transcriptional programs." In: *Trends Biochem Sci* 39.4, pp. 170–182. DOI: 10.1016/j.tibs.2014.02.007.

Lancaster, G. I. and M. A. Febbraio (2005). "Exosome-dependent Trafficking of HSP70: A Novel Secretory Pathway for Cellular Stress Proteins". In: *The*

Journal of biological chemistry 280.24, pp. 23349–23355. DOI: 10.1074/jbc.M502017200.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2". In: *Nat Methods* 9.4, pp. 357–359. DOI: 10.1038/nmeth.1923.

Larsen, F. et al. (1992). "CpG islands as gene markers in the human genome". In: *Genomics* 13.4, pp. 1095–1107. DOI: 10.1016/0888-7543(92)90024-M.

Latos, P. A. et al. (2012). "Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing". In: *Science (New York, NY)* 338.6113, pp. 1469–1472. DOI: 10.1126/science.1228110.

Lau, N. C. et al. (2001). "An abundant class of tiny RNAs with probable regulatory roles inCaenorhabditis elegans". In: *Science (New York, NY)* 294.5543, pp. 858–862.

Lawrence, M. et al. (2009). "rtracklayer: an R package for interfacing with genome browsers." In: *Bioinformatics (Oxford, England)* 25.14, pp. 1841–1842. DOI: 10.1093/bioinformatics/btp328.

Lee, R. C. et al. (1993). "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14". In: *Cell* 75.5, pp. 843–854. DOI: 10.1016/0092-8674(93)90529-y.

Lee, S. et al. (2016). "Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins." In: *Cell* 164 (1-2), pp. 69–80. DOI: 10.1016/j.cell.2015.12.017.

Lee, Y. S. et al. (2009). "Silencing by small RNAs is linked to endosomal trafficking." In: *Nat Cell Biol* 11.9, pp. 1150–1156. DOI: 10.1038/ncb1930.

Lee, Y. et al. (2002). "MicroRNA maturation: stepwise processing and subcellular localization". In: *The EMBO journal* 21.17, pp. 4663–4670. DOI: 10.1093/emboj/cdf476.

Lee, Y. et al. (2003). "The nuclear RNase III Drosha initiates microRNA processing". In: *Nature* 425.6956, pp. 415–419. DOI: 10.1038/nature01957.

Lee, Y. et al. (2004). "MicroRNA genes are transcribed by RNA polymerase II". In: *The EMBO journal* 23.20, pp. 4051–4060. DOI: 10.1038/sj.emboj.7600385.

Lehrbach, N. J. et al. (2009). "LIN-28 and the poly(U) polymerase PUP-2 regulate let-7 microRNA processing in Caenorhabditis elegans." In: *Nature structural & molecular biology* 16 (10), pp. 1016–1020. DOI: 10.1038/nsmb.1675.

Lener, T. et al. (2015). "Applying extracellular vesicles based therapeutics in clinical trials - an ISEV position paper." In: *J Extracell Vesicles* 4, p. 30087. DOI: 10.3402/jev.v4.30087.

Lévesque, K. et al. (2006). "Trafficking of HIV-1 RNA is mediated by heterogeneous nuclear ribonucleoprotein A2 expression and impacts on viral as-

sembly." In: *Traffic (Copenhagen, Denmark)* 7.9, pp. 1177–1193. DOI: 10.1111/j.1600-0854.2006.00461.x.

Lewis, B. P. et al. (2003). "Prediction of mammalian microRNA targets". In: *Cell* 115.7, pp. 787–798. DOI: 10.1016/s0092-8674(03)01018-3.

Lewis, B. P. et al. (2005). "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets". In: *Cell* 120.1, pp. 15–20. DOI: 10.1016/j.cell.2004.12.035.

Li, J. et al. (2013a). "Exosomes mediate the cell-to-cell transmission of IFN-α-induced antiviral activity". In: *Nat Immunol* 14.8, pp. 793–803. DOI: 10.1038/ni.2647.

Li, W. et al. (2013b). "Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation." In: *Nature* 498.7455, pp. 516–520. DOI: 10.1038/nature12210.

Lim, L. P. et al. (2005). "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs." In: *Nature* 433.7027, pp. 769–773. DOI: 10.1038/nature03315.

Lin, M. F. et al. (2011). "PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions." In: *Bioinformatics* 27.13, pp. i275–82. DOI: 10.1093/bioinformatics/btr209.

Link, S. et al. (2016). "Alternative splicing affects the subcellular localization of Drosha." In: *Nucleic Acids Res* 44.11, pp. 5330–5343. DOI: 10.1093/nar/gkw400.

Lipovich, L. et al. (2006). "Primate-specific endogenous cis-antisense transcription in the human 5q31 protocadherin gene cluster". In: *J Mol Evol* 62.1, pp. 73–88. DOI: 10.1007/s00239-005-0041-3.

Liu, S. J. et al. (2016). "Single-cell analysis of long non-coding RNAs in the developing human neocortex." In: *Genome Biol* 17, p. 67. DOI: 10.1186/s13059-016-0932-1.

Longatti, A. et al. (2014a). "Virion-independent transfer of replication competent HCV RNA between permissive cells." In: *J Virol*. DOI: 10.1128/JVI.02721-14.

Longatti, A. et al. (2014b). "Virion-Independent Transfer of Replication-Competent Hepatitis C Virus RNA between Permissive Cells". In: *Journal of Virology* 89.5 (5). Ed. by J.-H. J. Ou, pp. 2956–2961. DOI: 10.1128/jvi.02721-14.

Love, M. I. et al. (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." In: *Genome Biol* 15.12, p. 550. DOI: 10.1186/s13059-014-0550-8.

Lu, P. et al. (2003). "Neural stem cells constitutively secrete neurotrophic factors and promote extensive host axonal growth after spinal cord injury." In: *Exp Neurol* 181.2, pp. 115–129. DOI: 10.1016/s0014-4886(03)00037-2.

Lund, E. et al. (2004). "Nuclear export of microRNA precursors". In: *Science (New York, NY)* 303.5654, pp. 95–98. DOI: 10.1126/science.1090599.

Luo, S. et al. (2016). "Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells". In: *Cell stem cell* 18.5, pp. 637–652. DOI: 10.1016/j.stem.2016.01.024.

Luo, Z. et al. (2015). "Zic2 is an enhancer-binding factor required for embryonic stem cell specification". In: *Mol Cell* 57.4, pp. 685–694. DOI: 10.1016/j.molcel.2015.01.007.

Lykke-Andersen, K. et al. (2008). "Maternal Argonaute 2 is essential for early mouse development at the maternal-zygotic transition." In: *Molecular biology of the cell* 19 (10), pp. 4383–4392. DOI: 10.1091/mbc.E08-02-0219.

Lyle, R. et al. (2000). "The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1". In: *Nat Genet* 25.1, pp. 19–21. DOI: 10.1038/75546.

Lyon, M. F. (1961). "Gene action in the X-chromosome of the mouse (Mus musculus L.)." In: *Nature* 190, pp. 372–373. DOI: 10.1038/190372a0.

Ma, J.-B. et al. (2005). "Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein." In: *Nature* 434.7033, pp. 666–670. DOI: 10.1038/nature03514.

Macias, S. et al. (2015). "DGCR8 Acts as an Adaptor for the Exosome Complex to Degrade Double-Stranded Structured RNAs." In: *Molecular cell* 60 (6), pp. 873–885. DOI: 10.1016/j.molcel.2015.11.011.

Mack, M. et al. (2000). "Transfer of the chemokine receptor CCR5 between cells by membrane-derived microparticles: a mechanism for cellular human immunodeficiency virus 1 infection." In: *Nature medicine* 6.7 (7), pp. 769–775. DOI: 10.1038/77498.

Macrae, I. J. et al. (2006). "Structural basis for double-stranded RNA processing by Dicer." In: *Science* 311.5758, pp. 195–198. DOI: 10.1126/science.1121638.

Maenner, S. et al. (2010). "2-D structure of the A region of Xist RNA and its implication for PRC2 association". In: *PLoS Biol* 8.1, e1000276. DOI: 10.1371/journal.pbio.1000276.

Mahmoudi, S. et al. (2010). "WRAP53 is essential for Cajal body formation and for targeting the survival of motor neuron complex to Cajal bodies." In: *PLoS biology* 8 (11), e1000521. DOI: 10.1371/journal.pbio.1000521.

Mahmoudi, S. et al. (2009). "Wrap53, a natural p53 antisense transcript required for p53 induction upon DNA damage". In: *Mol Cell* 33.4, pp. 462–471. DOI: 10.1016/j.molcel.2009.01.028.

Maida, Y. et al. (2009). "An RNA-dependent RNA polymerase formed by TERT and the RMRP RNA". In: *Nature* 461.7261, pp. 230–235. DOI: 10.1038/nature08283.

Mak, W. et al. (2004). "Reactivation of the paternal X chromosome in early mouse embryos." In: *Science* 303.5658, pp. 666–669. DOI: 10.1126/science.1092674.

Malik, S. and R. G. Roeder (2010). "The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation." In: *Nat Rev Genet* 11.11, pp. 761–772. DOI: 10.1038/nrg2901.

Mariner, P. D. et al. (2008). "Human Alu RNA Is a Modular Transacting Repressor of mRNA Transcription during Heat Shock". In: *Mol Cell* 29.4, pp. 499–509.

Maroney, P. A. et al. (2006). "Evidence that microRNAs are associated with translating messenger RNAs in human cells." In: *Nature structural & molecular biology* 13 (12), pp. 1102–1107.

Marques, A. C. and C. P. Ponting (2009). "Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness." In: *Genome Biol* 10.11, R124. DOI: 10.1186/gb-2009-10-11-r124.

Marsico, A. et al. (2013). "PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs". In: *Genome Biol* 14.8, R84. DOI: 10.1186/gb-2013-14-8-r84.

Marson, A. et al. (2008). "Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells". In: *Cell* 134.3, pp. 521–533. DOI: 10.1016/j.cell.2008.07.020.

Martino, G. et al. (2011). "Brain Regeneration in Physiology and Pathology: The Immune Signature Driving Therapeutic Plasticity of Neural Stem Cells". In: *Physiol Rev* 91.4, pp. 1281–1304. DOI: 10.1152/physrev.00032.2010.

Mathonnet, G. et al. (2007). "MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F." In: *Science* 317.5845, pp. 1764–1767. DOI: 10.1126/science.1146067.

Mathys, H. et al. (2014). "Structural and biochemical insights to the role of the CCR4-NOT complex and DDX6 ATPase in microRNA repression." In: *Mol Cell* 54.5, pp. 751–765. DOI: 10.1016/j.molcel.2014.03.036.

Mattick, J. S. (2009). "Deconstructing the dogma: a new view of the evolution and genetic programming of complex organisms". In: *Ann N Y Acad Sci* 1178, pp. 29–46. DOI: 10.1111/j.1749-6632.2009.04991.x.

McGrath, J. and D. Solter (1984). "Completion of mouse embryogenesis requires both the maternal and paternal genomes." In: *Cell* 37.1, pp. 179–183. DOI: 10.1016/0092-8674(84)90313-1.

McHugh, C. A. et al. (2015). "The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3." In: *Nature* 521.7551, pp. 232–236. DOI: 10.1038/nature14443.

Megraw, M. (2006). "MicroRNA promoter element discovery in Arabidopsis". In: *RNA (New York, NY)* 12.9, pp. 1612–1619. DOI: 10.1261/rna.130506.

Melo, S. A. et al. (2014). "Cancer exosomes perform cell-independent microRNA biogenesis and promote tumorigenesis." In: *Cancer Cell* 26.5, pp. 707–721. DOI: 10.1016/j.ccell.2014.09.005.

Memczak, S. et al. (2013). "Circular RNAs are a large class of animal RNAs with regulatory potency". In: *Nature* 495.7441, pp. 333–338. DOI: 10.1038/nature11928.

Mercer, T. R. et al. (2008). "Specific expression of long noncoding RNAs in the mouse brain". In: *Proc Natl Acad Sci U S A* 105.2, pp. 716–721. DOI: 10.1073/pnas.0706729105.

Mercer, T. R. and J. S. Mattick (2013). "Structure and function of long noncoding RNAs in epigenetic regulation". In: *Nat Struct Mol Biol* 20.3, pp. 300–307. DOI: 10.1038/nsmb.2480.

Mercer, T. R. et al. (2009). "Long non-coding RNAs: insights into functions." In: *Nat Rev Genet* 10.3, pp. 155–159. DOI: 10.1038/nrg2521.

Mikkelsen, T. S. et al. (2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells". In: *Nature* 448.7153, pp. 553–560. DOI: 10.1038/nature06008.

Minajigi, A. et al. (2015). "Chromosomes. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation". In: *Science (New York, NY)* 349.6245. DOI: 10.1126/science.aab2276.

Ming, G.-l. and H. Song (2005). "Adult neurogenesis in the mammalian central nervous system". In: *Annu Rev Neurosci* 28, pp. 223–250. DOI: 10.1146/annurev.neuro.28.051804.101459.

Mittelbrunn, M. et al. (2011). "Unidirectional transfer of microRNA-loaded exosomes from T cells to antigen-presenting cells". In: *Nat Commun* 2, p. 282. DOI: 10.1038/ncomms1285.

Mondal, T. et al. (2015). "MEG3 long noncoding RNA regulates the TGF-beta pathway genes through formation of RNA-DNA triplex structures". In: *Nat Commun* 6, p. 7743. DOI: 10.1038/ncomms8743.

Monje, M. L. et al. (2003). "Inflammatory blockade restores adult hippocampal neurogenesis." In: *Science (New York, NY)* 302.5651, pp. 1760–1765. DOI: 10.1126/science.1088417.

Montecalvo, A. et al. (2012). "Mechanism of transfer of functional microRNAs between mouse dendritic cells via exosomes." In: *Blood* 119.3, pp. 756–766. DOI: 10.1182/blood-2011-02-338004.

Monteys, A. M. et al. (2010). "Structure and activity of putative intronic miRNA promoters". In: *RNA (New York, NY)* 16.3, pp. 495–505. DOI: 10.1261/rna.1731910.

Morlando, M. et al. (2008). "Primary microRNA transcripts are processed co-transcriptionally." In: *Nat Struct Mol Biol* 15.9, pp. 902–909. DOI: 10.1038/nsmb.1475.

Mostafavi, S. et al. (2008). "GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function". In: *Genome Biol* 9.Suppl 1, S4. DOI: 10.1186/gb-2008-9-s1-s4.

Mueller, F.-J. et al. (2006). "Adhesive interactions between human neural stem cells and inflamed human vascular endothelium are mediated by integrins." In: *Stem Cells* 24.11, pp. 2367–2372. DOI: 10.1634/stemcells.2005-0568.

Mugridge, J. S. et al. (2016). "Structural basis of mRNA-cap recognition by Dcp1-Dcp2." In: *Nature structural & molecular biology* 23 (11), pp. 987–994. DOI: 10.1038/nsmb.3301.

Munro, T. P. et al. (1999). "Mutational analysis of a heterogeneous nuclear ribonucleoprotein A2 response element for RNA trafficking." In: *The Journal of biological chemistry* 274.48, pp. 34389–34395. DOI: 10.1074/jbc.274.48.34389.

Nachbaur, K. et al. (1988). "Cytokines in the control of beta-2 microglobulin release. I. In vitro studies on various haemopoietic cells." In: *Immunobiology* 177.1, pp. 55–65. DOI: 10.1016/s0171-2985(88)80091-3.

Nagano, T. et al. (2008). "The Air Noncoding RNA Epigenetically Silences Transcription by Targeting G9a to Chromatin". In: *Science (New York, NY)* 322.5908, pp. 1717–1720. DOI: 10.1126/science.1163802.

Necsulea, A. et al. (2014). "The evolution of lncRNA repertoires and expression patterns in tetrapods". In: *Nature* 505.7485, pp. 635–640. DOI: 10.1038/nature12943.

Al-Nedawi, K. et al. (2009). "Microvesicles: messengers and mediators of tumor progression." In: *Cell cycle (Georgetown, Tex)* 8.13, pp. 2014–2018. DOI: 10.4161/cc.8.13.8988.

Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". In: *J Mol Biol* 48.3, pp. 443–453. DOI: 10.1016/0022-2836(70)90057-4.

Nguyen, T. A. et al. (2015). "Functional Anatomy of the Human Microprocessor." In: *Cell* 161.6, pp. 1374–1387. DOI: 10.1016/j.cell.2015.05.010.

Nicholls, R. et al. (1989). "Genetic imprinting suggested by maternal heterodisomy in nondeletion Prader-Willi syndrome." In: *Nature* 342.6247, pp. 281–285. DOI: 10.1038/342281a0.

Noland, C. L. and J. A. Doudna (2013). "Multiple sensors ensure guide strand selection in human RNAi pathways." In: *RNA* 19.5, pp. 639–648. DOI: 10.1261/rna.037424.112.

Nolte-'t Hoen, E. N. M. et al. (2009). "Activated T cells recruit exosomes secreted by dendritic cells via LFA-1". In: *Blood* 113.9, pp. 1977–1981. DOI: 10.1182/blood-2008-08-174094.

Nolte-'t Hoen, E. N. M. et al. (2012). "Deep sequencing of RNA from immune cell-derived vesicles uncovers the selective incorporation of small non-coding RNA biotypes with potential regulatory functions." In: *Nucleic acids research* 40 (18), pp. 9272–9285. DOI: 10.1093/nar/gks658.

Nottrott, S. et al. (2006). "Human let-7a miRNA blocks protein production on actively translating polyribosomes." In: *Nature structural & molecular biology* 13 (12), pp. 1108–1114.

O'Brien, S. J. (1973). "On Estimating Functional Gene Number in Eukaryotes". In: *Nature* 242.115, pp. 52–54. DOI: 10.1038/10.1038/newbio242052a0.

Odom, D. T. et al. (2006). "Core transcriptional regulatory circuitry in human hepatocytes." In: *Mol Syst Biol* 2, p. 20060017. DOI: 10.1038/msb4100059.

Okada, C. et al. (2009). "A high-resolution structure of the pre-microRNA nuclear export machinery." In: *Science* 326.5957, pp. 1275–1279. DOI: 10.1126/science.1178705.

Okamoto, I. et al. (2004). "Epigenetic dynamics of imprinted X inactivation during early mouse development." In: *Science* 303.5658, pp. 644–649. DOI: 10.1126/science.1092727.

Okamura, K. et al. (2007). "The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila". In: *Cell* 130.1, pp. 89–100. DOI: 10.1016/j.cell.2007.06.028.

Okazaki, Y. et al. (2002). "Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs". In: *Nature* 420.6915, pp. 563–73. DOI: 10.1038/nature01266.

Okoye, I. S. et al. (2014). "MicroRNA-containing T-regulatory-cell-derived exosomes suppress pat... - PubMed - NCBI". In: *Immunity* 41.1, pp. 89–103. DOI: 10.1016/j.immuni.2014.05.019.

Oliver, S. G. et al. (1992). "The complete DNA sequence of yeast chromosome III." In: *Nature* 357 (6373), pp. 38–46.

Orgel, L. E. and F. H. Crick (1980). "Selfish DNA: the ultimate parasite." In: *Nature* 284.5757, pp. 604–607.

Ørom, U. A. et al. (2010). "Long noncoding RNAs with enhancer-like function in human cells". In: *Cell* 143.1, pp. 46–58. DOI: 10.1016/j.cell.2010.09.001.

Ozsolak, F. et al. (2008). "Chromatin structure analyses identify miRNA promoters". In: *Genes & Development* 22.22, pp. 3172–3183. DOI: 10.1101/gad.1706508.

Palmer, T. (1997). "The Adult Rat Hippocampus Contains Primordial Neural Stem Cells". In: *Mol Cell Neurosci* 8.6, pp. 389–404. DOI: 10.1006/mcne.1996.0595.

Paralkar, V. R. et al. (2016). "Unlinking an lncRNA from Its Associated cis Element." In: *Molecular cell* 62 (1), pp. 104–110. DOI: 10.1016/j.molcel.2016.02.029.

Parker, J. S. et al. (2005). "Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex." In: *Nature* 434.7033, pp. 663–666. DOI: 10.1038/nature03462.

Parker, J. S. et al. (2009). "Enhancement of the seed-target recognition step in RNA silencing by a PIWI/MID domain protein." In: *Mol Cell* 33.2, pp. 204–214. DOI: 10.1016/j.molcel.2008.12.012.

Pasquinelli, A. E. et al. (2000). "Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA". In: *Nature* 408.6808, pp. 86–89. DOI: 10.1038/35040556.

Patel, D. M. et al. (1999). "Class II MHC/peptide complexes are released from APC and are acquired by T cell responders during specific antigen recognition." In: *The Journal of Immunology* 163.10, pp. 5201–5210.

Pawlicki, J. M. and J. A. Steitz (2008). "Primary microRNA transcript retention at sites of transcription leads to enhanced microRNA production". In: *The Journal of cell biology* 182.1, pp. 61–76. DOI: 10.1083/jcb.200803111.

Pegtel, D. M. et al. (2010). "Functional delivery of viral miRNAs via exosomes". In: *Proc Natl Acad Sci U S A* 107.14, pp. 6328–6333. DOI: 10.1073/pnas.0914843107.

Penny, G. D. et al. (1996). "Requirement for Xist in X chromosome inactivation". In: *Nature* 379.6561, pp. 131–137.

Perez-Hernandez, D. et al. (2013). "The intracellular interactome of tetraspanin-enriched microdomains reveals their function as sorting machineries toward exosomes." In: *The Journal of biological chemistry* 288 (17), pp. 11649–11661. DOI: 10.1074/jbc.M112.445304.

Peters, J. (2014). "The role of genomic imprinting in biology and disease: an expanding view." In: *Nat Rev Genet* 15.8, pp. 517–530. DOI: 10.1038/nrg3766.

Petersen, C. P. et al. (2006). "Short RNAs repress translation after initiation in mammalian cells." In: *Molecular cell* 21 (4), pp. 533–542.

Pickard, M. R. and G. T. Williams (2015). "Molecular and Cellular Mechanisms of Action of Tumour Suppressor GAS5 LncRNA". In: *Genes (Basel)* 6.3, pp. 484–499. DOI: 10.3390/genes6030484.

Pillai, R. S. et al. (2005). "Inhibition of translational initiation by Let-7 MicroRNA in human cells". In: *Science (New York, NY)* 309.5740, pp. 1573–1576. DOI: 10.1126/science.1115079.

Pillai, R. S. et al. (2004). "Tethering of human Ago proteins to mRNA mimics the miRNA-mediated repression of protein synthesis." In: *RNA* 10.10, pp. 1518–1525. DOI: 10.1261/rna.7131604.

Pluchino, S. and C. Cossetti (2013). "How stem cells speak with host immune cells in inflammatory brain diseases." In: *Glia* 61.9, pp. 1379–1401. DOI: 10.1002/glia.22500.

Pluchino, S. et al. (2003). "Injection of adult neurospheres induces recovery in a chronic model of multiple sclerosis." In: *Nature* 422.6933, pp. 688–694. DOI: 10.1038/nature01552.

Pluchino, S. et al. (2005). "Neurosphere-derived multipotent precursors promote neuroprotection by an immunomodulatory mechanism." In: *Nature* 436.7048, pp. 266–271. DOI: 10.1038/nature03889.

Pluchino, S. et al. (2008). "Persistent inflammation alters the function of the endogenous brain stem cell compartment." In: *Brain* 131.Pt 10, pp. 2564–2578. DOI: 10.1093/brain/awn198.

Pluchino, S. et al. (2009). "Human neural stem cells ameliorate autoimmune encephalomyelitis in non-human primates." In: *Ann Neurol* 66.3, pp. 343–354. DOI: 10.1002/ana.21745.

Pnueli, L. et al. (2015). "RNA transcribed from a distal enhancer is required for activating the chromatin at the promoter of the gonadotropin alpha-subunit gene". In: *Proceedings of the National Academy of Sciences* 112.14, pp. 4369–4374. DOI: 10.1073/pnas.1414841112.

Pocock, J. M. and A. C. Liddle (2001). "Microglial signalling cascades in neurodegenerative disease". In: *Glial cell function*. Elsevier, pp. 555–565. DOI: 10.1016/S0079-6123(01)32103-9.

Ponjavic, J. et al. (2007). "Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs." In: *Genome Res* 17.5, pp. 556–565. DOI: 10.1101/gr.6036807.

Qu, Y. et al. (2007). "Nonclassical IL-1 beta secretion stimulated by P2X7 receptors is dependent on inflammasome activation and correlated with exosome release in murine macrophages." In: *Journal of immunology (Baltimore,*

Md. : 1950) 179 (3), pp. 1913–1925. DOI: 10.4049/ jimmunol.179.3.1913.

Quick-Cleveland, J. et al. (2014). "The DGCR8 RNA-binding heme domain recognizes primary microRNAs by clamping the hairpin." In: *Cell Rep* 7.6, pp. 1994–2005. DOI: 10.1016/j.celrep.2014.05.013.

Quinlan, A. R. (2014). "BEDTools: The Swiss-Army Tool for Genome Feature Analysis". In: *Curr Protoc Bioinformatics* 47, pp. 11121–111234. DOI: 10.1002/0471250953.bi1112s47.

Quinn, J. J. and H. Y. Chang (2016). "Unique features of long non-coding RNA biogenesis and function." In: *Nat Rev Genet* 17.1, pp. 47–62. DOI: 10.1038/nrg.2015.10.

Quinn, J. J. et al. (2014). "Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification." In: *Nat Biotechnol* 32.9, pp. 933–940. DOI: 10.1038/nbt.2943.

Quinn, J. J. et al. (2016). "Rapid evolutionary turnover underlies conserved lncRNA-genome interactions." In: *Genes & development* 30 (2), pp. 191–207. DOI: 10.1101/gad.272187.115.

Rahman, S. et al. (2016). "Single-cell profiling reveals that eRNA accumulation at enhancer-promoter loops is not required to sustain transcription." In: *Nucleic acids research*. DOI: 10.1093/nar/gkw1220.

Rajendran, L. et al. (2006). "Alzheimer's disease beta-amyloid peptides are released in association with exosomes." In: *Proc Natl Acad Sci U S A* 103.30, pp. 11172–11177. DOI: 10.1073/pnas.0603838103.

Ramalingam, P. et al. (2014). "Biogenesis of intronic miRNAs located in clusters by independent transcription and alternative splicing." In: *RNA* 20.1, pp. 76–87. DOI: 10.1261/rna.041814.113.

Rampon, C. et al. (2008). "Molecular mechanism of systemic delivery of neural precursor cells to the brain: assembly of brain endothelial apical cups and control of transmigration by CD44." In: *Stem Cells* 26.7, pp. 1673–1682. DOI: 10.1634/stemcells.2008-0122.

Rao, S. et al. (2014). "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping". In: *Cell* 159.7, pp. 1665–1680. DOI: 10.1016/j.cell.2014.11.021.

Raposo, G. and W. Stoorvogel (2013). "Extracellular vesicles: exosomes, microvesicles, and friends." In: *The Journal of cell biology* 200.4, pp. 373–383. DOI: 10.1083/jcb.201211138.

Raposo, G. et al. (1996). "B lymphocytes secrete antigen-presenting vesicles." In: *The Journal of experimental medicine* 183.3, pp. 1161–1172. DOI: 10.1084/jem.183.3.1161.

Ratajczak, J. et al. (2006). "Embryonic stem cell-derived microvesicles reprogram hematopoietic progenitors: evidence for horizontal transfer of mRNA and protein delivery." In: *Leukemia* 20.5, pp. 847–856. DOI: 10.1038/sj.leu.2404132.

Ravasi, T. et al. (2006). "Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome". In: *Genome Res* 16.1, pp. 11–19. DOI: 10.1101/gr.4200206.

Ray, D. et al. (2013). "A compendium of RNA-binding motifs for decoding gene regulation." In: *Nature* 499.7457, pp. 172–177. DOI: 10.1038/nature12311.

Rinn, J. L. and H. Y. Chang (2012). "Genome regulation by long noncoding RNAs". In: *Annu Rev Biochem* 81.1, pp. 145–166. DOI: 10.1146/annurev-biochem-051410-092902.

Rinn, J. L. et al. (2007). "Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs". In: *Cell* 129.7, pp. 1311–1323. DOI: 10.1016/j.cell.2007.05.022.

Robbins, P. D. and A. E. Morelli (2014). "Regulation of immune responses by extracellular vesicles." In: *Nat Rev Immunol* 14.3, pp. 195–208. DOI: 10.1038/nri3622.

Rodriguez, A. (2004). "Identification of Mammalian microRNA Host Genes and Transcription Units". In: *Genome Res* 14.10a, pp. 1902–1910. DOI: 10.1101/gr.2722704.

Rodriguez, A. et al. (2007). "Requirement of bic/microRNA-155 for Normal Immune Function". In: *Science (New York, NY)* 316.5824, pp. 608–611. DOI: 10.1126/science.1139253.

Rolls, A. et al. (2007). "Toll-like receptors modulate adult hippocampal neurogenesis". In: *Nat Cell Biol* 9.9, pp. 1081–1088. DOI: 10.1038/ncb1629.

Rosenbloom, K. R. et al. (2015). "The UCSC Genome Browser database: 2015 update". In: *Nucleic Acids Res* 43.Database issue, pp. D670–81. DOI: 10.1093/nar/gku1177.

Roth, B. M. et al. (2013). "The core microprocessor component DiGeorge syndrome critical region 8 (DGCR8) is a nonspecific RNA-binding protein." In: *J Biol Chem* 288.37, pp. 26785–26799. DOI: 10.1074/jbc.M112.446880.

Roucourt, B. et al. (2015). "Heparanase activates the syndecan-syntenin-ALIX exosome pathway." In: *Cell research* 25 (4), pp. 412–428. DOI: 10.1038/cr.2015.29.

Ruby, J. G. et al. (2007). "Intronic microRNA precursors that bypass Drosha processing". In: *Nature* 448.7149, pp. 83–86.

Sadir, R. et al. (1998). "The heparan sulfate binding sequence of interferon-gamma increased the on rate of the interferon-gamma-interferon-gamma

receptor complex formation." In: *The Journal of biological chemistry* 273.18, pp. 10919–10925.

Sáenz-Cuesta, M. et al. (2014). "Extracellular Vesicles in Multiple Sclerosis: What are They Telling Us?" In: *Front Cell Neurosci* 8. DOI: 10.3389/fncel.2014.00100.

Saini, H. K. et al. (2007). "Genomic analysis of human microRNA transcripts." In: *Proc Natl Acad Sci U S A* 104.45, pp. 17719–17724. DOI: 10.1073/pnas.0703890104.

Salzman, J. et al. (2012). "Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types." In: *PLoS One* 7.2, e30733. DOI: 10.1371/journal.pone.0030733.

Sanborn, A. L. et al. (2015). "Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes." In: *Proceedings of the National Academy of Sciences of the United States of America* 112 (47), E6456–E6465. DOI: 10.1073/pnas.1518552112.

Sasaki, T. and N. Shimizu (2007). "Evolutionary conservation of a unique amino acid sequence in human DICER protein essential for binding to Argonaute family proteins." In: *Gene* 396.2, pp. 312–320. DOI: 10.1016/j.gene.2007.04.001.

Schaukowitch, K. et al. (2014). "Enhancer RNA facilitates NELF release from immediate early genes." In: *Mol Cell* 56.1, pp. 29–42. DOI: 10.1016/j.molcel.2014.08.023.

Schirle, N. T. and I. J. MacRae (2012). "The crystal structure of human Argonaute2." In: *Science* 336.6084, pp. 1037–1040. DOI: 10.1126/science.1221551.

Schmidt, O. and D. Teis (2012). "The ESCRT machinery." In: *Current biology : CB* 22 (4), R116–R120. DOI: 10.1016/j.cub.2012.01.028.

Schüler, A. et al. (2014). "Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs." In: *Mol Biol Evol* 31.12, pp. 3164–3183. DOI: 10.1093/molbev/msu249.

Schwarz, D. S. et al. (2003). "Asymmetry in the assembly of the RNAi enzyme complex". In: *Cell* 115.2, pp. 199–208. DOI: 10.1016/s0092-8674(03)00759-1.

Selbach, M. et al. (2008). "Widespread changes in protein synthesis induced by microRNAs." In: *Nature* 455.7209, pp. 58–63. DOI: 10.1038/nature07228.

Sessa, L. et al. (2007). "Noncoding RNA synthesis and loss of Polycomb group repression accompanies the colinear activation of the human HOXA cluster." In: *RNA (New York, NY)* 13.2, pp. 223–239. DOI: 10.1261/rna.266707.

Sharma, U. et al. (2016). "Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals." In: *Science (New York, N.Y.)* 351 (6271), pp. 391–396. DOI: 10.1126/science.aad6780.

She, M. et al. (2008). "Structural basis of dcp2 recognition and activation by dcp1." In: *Molecular cell* 29 (3), pp. 337–349. DOI: 10.1016/j.molcel.2008.01.002.

Shearwin, K. E. et al. (2005). "Transcriptional interference–a crash course." In: *Trends Genet* 21.6, pp. 339–345. DOI: 10.1016/j.tig.2005.04.009.

Shen, B. et al. (2011). "Protein Targeting to Exosomes/Microvesicles by Plasma Membrane Anchors". In: *The Journal of biological chemistry* 286.16, pp. 14383–14395. DOI: 10.1074/jbc.M110.208660.

Siddiqui, N. et al. (2007). "Poly(A) nuclease interacts with the C-terminal domain of polyadenylate-binding protein domain from poly(A)-binding protein." In: *J Biol Chem* 282.34, pp. 25067–25075. DOI: 10.1074/jbc.M701256200.

Siepel, A. et al. (2005). "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes". In: *Genome Res* 15.8, pp. 1034–1050. DOI: 10.1101/gr.3715005.

Sigova, A. A. et al. (2015). "Transcription factor trapping by RNA in gene regulatory elements". In: *Science (New York, NY)*. DOI: 10.1126/science.aad3346.

Simon, M. D. et al. (2013). "High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation." In: *Nature* 504.7480, pp. 465–469. DOI: 10.1038/nature12719.

Skog, J. et al. (2008). "Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers." In: *Nat Cell Biol* 10.12, pp. 1470–1476. DOI: 10.1038/ncb1800.

Sleutels, F. et al. (2002). "The non-coding Air RNA is required for silencing autosomal imprinted genes". In: *Nature* 415.6873, pp. 810–813. DOI: 10.1038/415810a.

Smith, J. A. et al. (2014). "Extracellular vesicles and their synthetic analogues in aging and age-associated brain diseases." In: *Biogerontology* 16.2, pp. 147–185. DOI: 10.1007/s10522-014-9510-7.

Smith, M. A. et al. (2013). "Widespread purifying selection on RNA structure in mammals." In: *Nucleic Acids Res* 41.17, pp. 8220–8236. DOI: 10.1093/nar/gkt596.

Smith, Z. J. et al. (2015). "Single exosome study reveals subpopulations distributed among cell lines with variability related to membrane content." In: *Journal of extracellular vesicles* 4, p. 28533. DOI: 10.3402/jev.v4.28533.

Smyth, G. and T. Speed (2003). "Normalization of cDNA microarray data". In: *Methods* 31.4, pp. 265–273.

Smyth, G. et al. (2005). *Limma: linear models for microarray data user's guide*.

Squadrito, M. L. et al. (2014). "Endogenous RNAs Modulate MicroRNA Sorting to Exosomes and Transfer to Acceptor Cells". In: *Cell Reports* 8.5, pp. 1432–1446. DOI: 10.1016/j.celrep.2014.07.035.

Stoorvogel, W. et al. (1991). "Late endosomes derive from early endosomes by maturation." In: *Cell* 65 (3), pp. 417–427. DOI: 10.1016/0092-8674(91)90459-c.

Struyf, S. et al. (2009). "Synergistic up-regulation of MCP-2/CCL8 activity is counteracted by chemokine cleavage, limiting its inflammatory and antitumoral effects." In: *Eur J Immunol* 39.3, pp. 843–857. DOI: 10.1002/eji.200838660.

Stuffers, S. et al. (2009). "Multivesicular endosome biogenesis in the absence of ESCRTs." In: *Traffic (Copenhagen, Denmark)* 10 (7), pp. 925–937. DOI: 10.1111/j.1600-0854.2009.00920.x.

Su, H. et al. (2009). "Essential and overlapping functions for mammalian Argonautes in microRNA silencing." In: *Genes Dev* 23.3, pp. 304–317. DOI: 10.1101/gad.1749809.

Surani, M. et al. (1984). "Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis." In: *Nature* 308.5959, pp. 548–550. DOI: 10.1038/308548a0.

Szajnik, M. et al. (2010). "Tumor-Derived Microvesicles Induce, Expand and Up-Regulate Biological Activities of Human Regulatory T Cells (Treg)". In: *PLoS One* 5.7, e11469. DOI: 10.1371/journal.pone.0011469.

Szostak, N. et al. (2014). "Sorting signal targeting mRNA into hepatic extracellular vesicles". In: *RNA Biol* 11.7, pp. 836–844. DOI: 10.4161/rna.29305.

Tang, X. et al. (2010). "Phosphorylation of the RNase III enzyme Drosha at Serine300 or Serine302 is required for its nuclear localization." In: *Nucleic Acids Res* 38.19, pp. 6610–6619. DOI: 10.1093/nar/gkq547.

Tang, Y. et al. (2011). "FOXA2 functions as a suppressor of tumor metastasis by inhibition of epithelial-to-mesenchymal transition in human lung cancers". In: *Cell Res* 21.2, pp. 316–326. DOI: 10.1038/cr.2010.126.

Tang, Z. et al. (2015). "CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription". In: *Cell* 163.7, pp. 1611–1627. DOI: 10.1016/j.cell.2015.11.024.

Tauro, B. J. et al. (2013). "Oncogenic H-ras reprograms Madin-Darby canine kidney (MDCK) cell-derived exosomal proteins following epithelial-mesenchymal transition." In: *Molecular & cellular proteomics : MCP* 12.8, pp. 2148–2159. DOI: 10.1074/mcp.M112.027086.

Taylor, A. R. et al. (2007). "Regulation of heat shock protein 70 release in astrocytes: Role of signaling kinases". In: *Developmental Neurobiology* 67.13, pp. 1815–1829. DOI: 10.1002/dneu.20559.

Théry, C. et al. (2002a). "Exosomes: composition, biogenesis and function." In: *Nat Rev Immunol* 2.8, pp. 569–579. DOI: 10.1038/nri855.

Théry, C. et al. (2002b). "Indirect activation of naïve CD4+ T cells by dendritic cell-derived exosomes." In: *Nat Immunol* 3.12, pp. 1156–1162. DOI: 10.1038/ni854.

Théry, C. et al. (1999). "Molecular characterization of dendritic cell-derived exosomes. Selective accumulation of the heat shock protein hsc73." In: *The Journal of cell biology* 147.3, pp. 599–610. DOI: 10.1083/jcb.147.3.599.

Théry, C. et al. (2001). "Proteomic Analysis of Dendritic Cell-Derived Exosomes: A Secreted Subcellular Compartment Distinct from Apoptotic Vesicles". In: *The Journal of Immunology* 166.12, pp. 7309–7318. DOI: 10.4049/jimmunol.166.12.7309.

Thurman, R. E. et al. (2012). "The accessible chromatin landscape of the human genome". In: *Nature* 489.7414, pp. 75–82. DOI: 10.1038/nature11232.

Tian, Y. et al. (2014). "A phosphate-binding pocket within the platform-PAZ-connector helix cassette of human Dicer." In: *Mol Cell* 53.4, pp. 606–616. DOI: 10.1016/j.molcel.2014.01.003.

Tichon, A. et al. (2016). "A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells." In: *Nature communications* 7, p. 12209. DOI: 10.1038/ncomms12209.

Tomari, Y. et al. (2004). "A protein sensor for siRNA asymmetry." In: *Science* 306.5700, pp. 1377–1380. DOI: 10.1126/science.1102755.

Trajkovic, K. et al. (2008). "Ceramide Triggers Budding of Exosome Vesicles into Multivesicular Endosomes". In: *Science* 319.5867, pp. 1244–1247. DOI: 10.1126/science.1153124.

Trams, E. G. et al. (1981). "Exfoliation of membrane ecto-enzymes in the form of micro-vesicles." In: *Biochim Biophys Acta* 645.1, pp. 63–70. DOI: 10.1016/0005-2736(81)90512-5.

Trapnell, C. et al. (2012a). "Differential analysis of gene regulation at transcript resolution with RNA-seq". In: *Nat Biotechnol* 31.1, pp. 46–53. DOI: 10.1038/nbt.2450.

Trapnell, C. et al. (2012b). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks". In: *Nat Protoc* 7.3, pp. 562–578. DOI: 10.1038/nprot.2012.016.

Tsai, M. C. et al. (2010). "Long noncoding RNA as modular scaffold of histone modification complexes". In: *Science (New York, NY)* 329.5992, pp. 689–693. DOI: 10.1126/science.1192002.

Tsutsumi, A. et al. (2011). "Recognition of the pre-miRNA structure by Drosophila Dicer-1." In: *Nat Struct Mol Biol* 18.10, pp. 1153–1158. DOI: 10.1038/nsmb.2125.

Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley series in behavioral science. Addison-Wesley Publishing Company. DOI: 10.2307/2529486.

Turiák, L. et al. (2011). "Proteomic characterization of thymocyte-derived microvesicles and apoptotic bodies in BALB/c mice." In: *Journal of proteomics* 74 (10), pp. 2025–2033. DOI: 10.1016/j.jprot.2011.05.023.

Uchida, N. et al. (2004). "Identification of a human cytoplasmic poly(A) nuclease complex stimulated by poly(A)-binding protein." In: *J Biol Chem* 279.2, pp. 1383–1391. DOI: 10.1074/jbc.M309125200.

Ulitsky, I. et al. (2011). "Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution." In: *Cell* 147.7, pp. 1537–1550. DOI: 10.1016/j.cell.2011.11.055.

Valadi, H. et al. (2007). "Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells". In: *Nat Cell Biol* 9.6, pp. 654–659. DOI: 10.1038/ncb1596.

van Dongen, S. (2008). "Graph clustering via a discrete uncoupling process". In: *SIAM J Matrix Anal Appl* 30.1, pp. 121–141. DOI: 10.1137/040608635.

Verweij, F. J. et al. (2011). "LMP1 association with CD63 in endosomes and secretion via exosomes limits constitutive NF-κB activation". In: *The EMBO journal* 30.11, pp. 2115–2129. DOI: 10.1038/emboj.2011.123.

Villarroya-Beltri, C. et al. (2013). "Sumoylated hnRNPA2B1 controls the sorting of miRNAs into exosomes through binding to specific motifs." In: *Nat Commun* 4, p. 2980. DOI: 10.1038/ncomms3980.

Villarroya-Beltri, C. et al. (2014). "Sorting it out: Regulation of exosome loading". In: *Semin Cancer Biol* 28, pp. 3–13. DOI: 10.1016/j.semcancer.2014.04.009.

Wakiyama, M. et al. (2007). "Let-7 microRNA-mediated mRNA deadenylation and translational repression in a mammalian cell-free system." In: *Genes Dev* 21.15, pp. 1857–1862. DOI: 10.1101/gad.1566707.

Wang, J. et al. (2014). "FOXA2 suppresses the metastasis of hepatocellular carcinoma partially through matrix metalloproteinase-9 inhibition". In: *Carcinogenesis* 35.11, pp. 2576–2583. DOI: 10.1093/carcin/bgu180.

Wang, K. C. et al. (2011). "A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression". In: *Nature* 472.7341, pp. 120–124. DOI: 10.1038/nature09819.

Wang, L. et al. (2013). "CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model". In: *Nucleic Acids Res* 41.6, e74. DOI: 10.1093/nar/gkt006.

Wang, P. et al. (2009). "Global analysis of H3K4 methylation defines MLL family member targets and points to a role for MLL1-mediated H3K4 methylation in the regulation of transcriptional initiation by RNA polymerase II." In: *Mol Cell Biol* 29.22, pp. 6074–6085. DOI: 10.1128/MCB.00924-09.

Wang, Y. et al. (2008). "Structure of the guide-strand-containing argonaute silencing complex." In: *Nature* 456.7219, pp. 209–213. DOI: 10.1038/nature07315.

Washietl, S. et al. (2014). "Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals." In: *Genome Res* 24.4, pp. 616–628. DOI: 10.1101/gr.165035.113.

Washietl, S. et al. (2005). "Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome". In: *Nat Biotechnol* 23.11, pp. 1383–1390. DOI: 10.1038/nbt1144.

Weitz, S. H. et al. (2014). "Processing of microRNA primary transcripts requires heme in mammalian cells." In: *Proc Natl Acad Sci U S A* 111.5, pp. 1861–1866. DOI: 10.1073/pnas.1309915111.

Welch, B. L. (1947). "The generalization of 'Student's' problem when several different population variances are involved". In: *Biometrika* 34.1-2, pp. 28–35. DOI: 10.1093/biomet/34.1-2.28.

Wen, Z. et al. (1995). "Maximal activation of transcription by Stat1 and Stat3 requires both tyrosine and serine phosphorylation." In: *Cell* 82.2, pp. 241–250. DOI: 10.1016/0092-8674(95)90311-9.

Wightman, B. et al. (1993). "Posttranscriptional regulation of the heterochronic gene lin-14 by lin- 4 mediates temporal pattern formation inC. elegans". In: *Cell* 75.5, pp. 855–862. DOI: 10.1016/0092-8674(93)90530-4.

Wiley, R. D. and S. Gummuluru (2006). "Immature dendritic cell-derived exosomes can mediate HIV-1 trans infection". In: *Proc Natl Acad Sci U S A* 103.3, pp. 738–743. DOI: 10.1073/pnas.0507995103.

Williamson, C. et al. (2013). *Mouse Imprinting Data and References*. MRC Harwell, Oxfordshire. URL: http://www.har.mrc.ac.uk/research/genomic_imprinting (visited on 08/05/2016).

Wilson, R. C. et al. (2015). "Dicer-TRBP complex formation ensures accurate mammalian microRNA biogenesis." In: *Mol Cell* 57.3, pp. 397–407. DOI: 10.1016/j.molcel.2014.11.030.

Wilusz, J. E. et al. (2008). "3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA." In: *Cell* 135.5, pp. 919–932. DOI: 10.1016/j.cell.2008.10.012.

Wilusz, J. E. et al. (2012). "A triple helix stabilizes the 3' ends of long noncoding RNAs that lack poly(A) tails." In: *Genes & Development* 26.21, pp. 2392–2407. DOI: 10.1101/gad.204438.112.

Witwer, K. W. et al. (2013). "Standardization of sample collection, isolation and analysis methods in extracellular vesicle research." In: *Journal of extracellular vesicles* 2. DOI: 10.3402/jev.v2i0.20360.

Wolf, P. (1967). "The nature and significance of platelet products in human plasma." In: *British journal of haematology* 13 (3), pp. 269–288. DOI: 10.1111/j.1365-2141.1967.tb08741.x.

Wollert, T. and J. H. Hurley (2010). "Molecular mechanism of multivesicular body biogenesis by ESCRT complexes." In: *Nature* 464 (7290), pp. 864–869. DOI: 10.1038/nature08849.

Wong, G. H. W. et al. (1984). "Inducible expression of H–2 and Ia antigens on brain cells". In: *Nature* 310.5979, pp. 688–691. DOI: 10.1038/310688a0.

Wutz, A. and R. Jaenisch (2000). "A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation." In: *Molecular cell* 5 (4), pp. 695–705.

Xie, D. et al. (2013a). "Dynamic trans-acting factor colocalization in human cells". In: *Cell* 155.3, pp. 713–724. DOI: 10.1016/j.cell.2013.09.043.

Xie, M. et al. (2013b). "Mammalian 5'-capped microRNA precursors that generate a single microRNA." In: *Cell* 155.7, pp. 1568–1580. DOI: 10.1016/j.cell.2013.11.027.

Xue, Z. et al. (2016). "A G-Rich Motif in the lncRNA Braveheart Interacts with a Zinc-Finger Transcription Factor to Specify the Cardiovascular Lineage." In: *Molecular cell* 64 (1), pp. 37–50. DOI: 10.1016/j.molcel.2016.08.010.

Yamashita, A. et al. (2005). "Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA turnover." In: *Nat Struct Mol Biol* 12.12, pp. 1054–1063. DOI: 10.1038/nsmb1016.

Yang, Y. et al. (2016). "Enhancer RNA-driven looping enhances the transcription of the long noncoding RNA DHRS4-AS1, a controller of the DHRS4 gene cluster". In: *Sci Rep* 6, p. 20961. DOI: 10.1038/srep20961.

Yao, B. et al. (2012). "Defining a new role of GW182 in maintaining miRNA stability". In: *EMBO Rep* 13.12, pp. 1102–1108. DOI: 10.1038/embor.2012.160.

Yap, K. L. et al. (2010). "Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a". In: *Mol Cell* 38.5, pp. 662–674. DOI: 10.1016/j.molcel.2010.03.021.

Yi, R. et al. (2003). "Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs." In: *Genes Dev* 17.24, pp. 3011–3016. DOI: 10.1101/gad.1158803.

Zeng, Y. et al. (2005). "Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha". In: *The EMBO journal* 24.1, pp. 138–148. DOI: 10.1038/sj.emboj.7600491.

Zhan, R. et al. (2009). "Heat shock protein 70 is secreted from endothelial cells by a non-classical pathway involving exosomes". In: *Biochem Biophys Res Commun* 387.2, pp. 229–233. DOI: 10.1016/j.bbrc.2009.06.095.

Zhang, H. et al. (2004). "Single processing center models for human Dicer and bacterial RNase III." In: *Cell* 118.1, pp. 57–68. DOI: 10.1016/j.cell.2004.06.017.

Zhang, K. et al. (2014). "Identification and functional analysis of long non-coding RNAs in mouse cleavage stage embryonic development based on single cell transcriptome data." In: *BMC Genomics* 15, p. 845. DOI: 10.1186/1471-2164-15-845.

Zhang, X. et al. (2009). "A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster." In: *Blood* 113.11, pp. 2526–2534. DOI: 10.1182/blood-2008-06-162164.

Zhang, X. et al. (2015). "Exosomes in cancer: small particle, big player." In: *Journal of hematology & oncology* 8, p. 83. DOI: 10.1186/s13045-015-0181-x.

Zhang, X. et al. (2010). "Maternally expressed gene 3, an imprinted noncoding RNA gene, is associated with meningioma pathogenesis and progression". In: *Cancer Res* 70.6, pp. 2350–2358. DOI: 10.1158/0008-5472.CAN-09-3885.

Zhang, Y. et al. (2013). "Circular intronic long noncoding RNAs." In: *Mol Cell* 51.6, pp. 792–806. DOI: 10.1016/j.molcel.2013.08.017.

Zhao, J. et al. (2008). "Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome". In: *Science (New York, NY)* 322.5902, pp. 750–756. DOI: 10.1126/science.1163045.

Zheng, P. et al. (2015). "Quantitative Proteomics Analysis Reveals Novel Insights into Mechanisms of Action of Long Noncoding RNA Hox Transcript Antisense Intergenic RNA (HOTAIR) in HeLa Cells." In: *Mol Cell Proteomics* 14.6, pp. 1447–1463. DOI: 10.1074/mcp.M114.043984.

Zhou, X. et al. (2007). "Characterization and identification of microRNA core promoters in four model species". In: *PLoS Comput Biol* 3.3, e37. DOI: 10.1371/journal.pcbi.0030037.eor.

Zhou, Y. et al. (2012). "MEG3 noncoding RNA: a tumor suppressor." In: *J Mol Endocrinol* 48.3, R45–R53. DOI: 10.1530/JME-12-0008.

Zhuang, G. et al. (2012). "Tumour-secreted miR-9 promotes endothelial cell migration and angiogenesis by activating the JAK-STAT pathway". In: *The EMBO journal* 31.17, pp. 3513–3523. DOI: 10.1038/emboj.2012.183.

Zid, B. M. and E. K. O'Shea (2014). "Promoter sequences direct cytoplasmic localization and translation of mRNAs during starvation in yeast." In: *Nature* 514 (7520), pp. 117–121. DOI: 10.1038/nature13578.

Ziv, Y. et al. (2006). "Synergy between immune cells and adult neural stem/progenitor cells promotes functional recovery from spinal cord injury". In: *Proc Natl Acad Sci U S A* 103.35, pp. 13174–13179. DOI: 10.1073/pnas.0603747103.

Zöller, M. (2009). "Tetraspanins: push and pull in suppressing and promoting metastasis". In: *Nat Rev Cancer* 9.1, pp. 40–55. DOI: 10.1038/nrc2543.