DIMET - Dr. Valeria RANZANI - A.A. 2013-14

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

PhD

PROGRAM IN TRANSLATIONAL AND MOLECULAR MEDICINE

DIMET

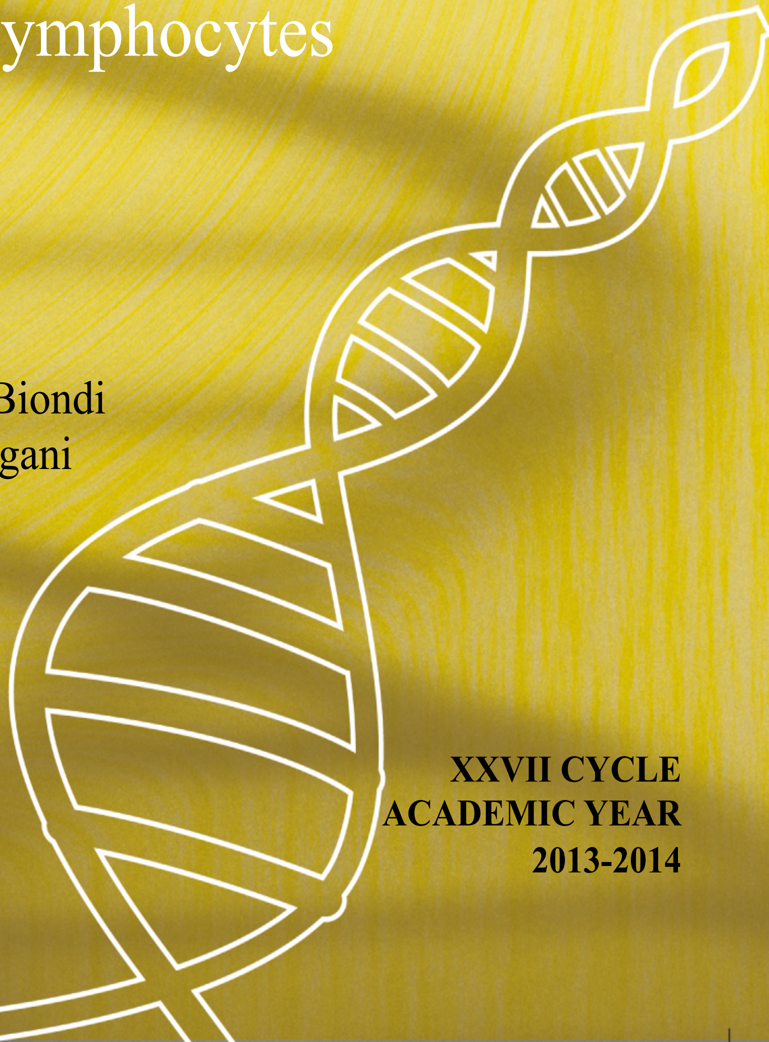
UNIVERSITY OF MILANO-BICOCCA
SCHOOL OF MEDICINE AND SCHOOL OF SCIENCE

Identification of regulatory non coding RNAs in human lymphocytes

Coordinator: Prof. Andrea Biondi
Tutor: Dr. Massimiliano Pagani

Dr. Valeria RANZANI

Matr. No. 067669



XXVII CYCLE
ACADEMIC YEAR
2013-2014

..nothing in life is to be feared, it is only to be understood..
..now is the time to understand more, so that we may fear less..

Marie Curie

Table of contents

Chapter 1

Introduction

The landscape of human transcription

After more than a century from the definition of the basic rules of hereditary, the gene is heading an identity crisis. A lot of debates about gene description are going in recent years [99, 126, 38, 40; 127; 11,103] thereby definition of the "gene" evolved from *"the site of hereditable trait*" to "*the genomic region from where the mRNA that encodes a protein is transcribed*", which defined the molecular biology's central dogma [23].

In the last century, with the advent of the new technologies, the transcriptional-centered view is becoming more complicated. In particular, the alternative splicing process and the discovery of non coding RNAs (ncRNAs) suggest that most of the human transcripts may not encode for proteins [53, 133]. Also with the addiction of precious informations coming from the ENCODE Project Consortium, further highlighted that the bulk of the genome is pervasively transcribed although the functional relevance of these transcripts remains a source of debate [2, 29; 45] (Figure 1).

The evolving definition of the gene and the growing transcriptional complexity complicate the scientific works and the common language in which scientists discuss. For this reason the modern effort to redefine the gene is attempting to retrofit biological complexity into an existing vocabulary that is understable to scientists across a range of disciplines. To this end, Gerstein and colleagues recently proposed a new concept of gene as "*a union of genomic sequences encoding a coherent set of potentially overlapping functional products*" [38], but also this definition is limited to unit of functionality and does not include the collective term for a group of transcripts. There are other important open questions concealed within the "what is a gene?" debate, for example the size of the transcriptome and the relative functional portion, what is the real meaning of "functional" gene and how functionality can be captured in gene annotation.

A key point in the definition of the total number of human transcripts is the significant difference of transcripts produced and expressed in cells of distinct tissues and developmental stages [55, 121, 161, 9, 65]. Moreover, splicing abnormalities as well as the presence of polymorfism are commonly observed in both cancer and immortalized cells [10, 162, 16, 26; The ENCODE Project Consortium 2012] as well as the presence of polymorfism [106; 128; 42]. No single human transcriptome can be considered as a reference, so a "consensus" set of transcripts that combines all known transcripts into one "gene set" is necessary. Several large-scale gene annotation projects on human genome are evolving, including GENCODE [53], RefSeq [131] and UCSC Genes [30]. These projects are mainly based on transcriptomics data and represent a merge between models of different data sources: GENCODE includes Ensembl models and manually annotated HAVANA transcripts, RefSeq combines manual and automated processed (most human annotation takes place on full-length cDNAs that are subsequently linked to the chromosome) and finally UCSC joins RefSeq models and other sources as GenBank ESTs.

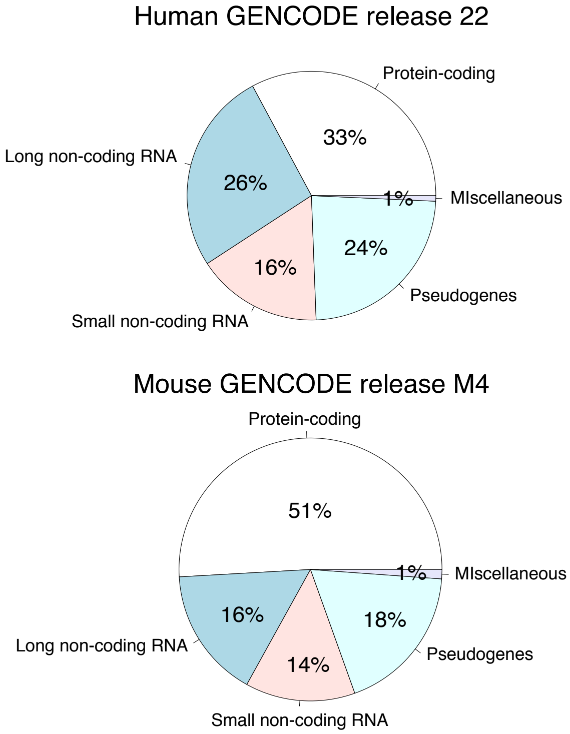
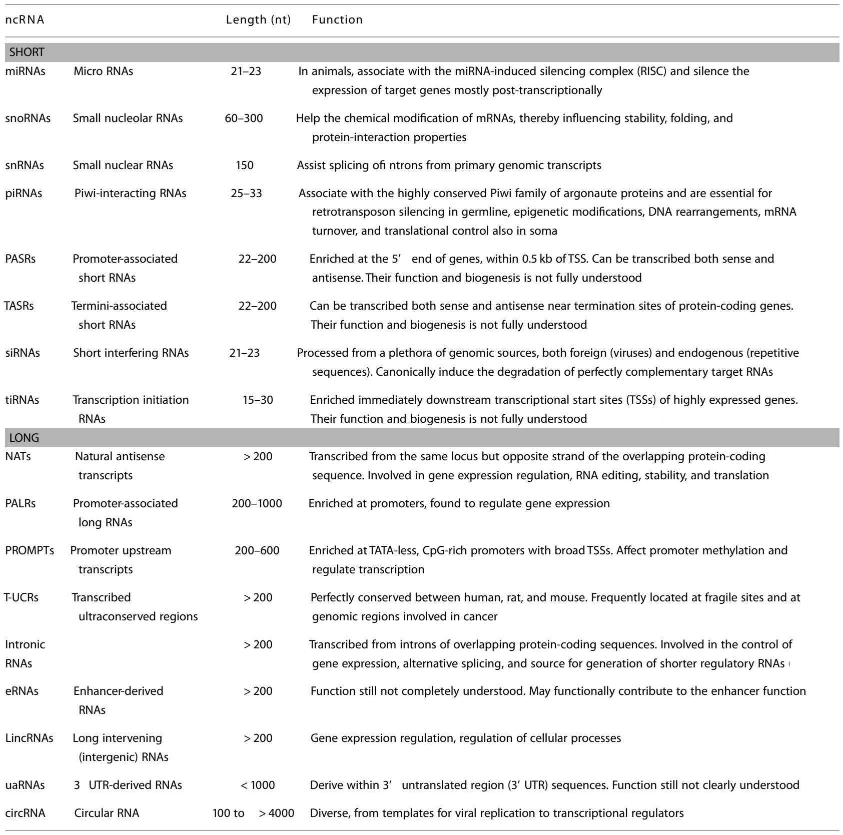


Figure 1. Long non-coding RNAs are the most abundant ncRNA species in the mammalian genome. Pie charts showing the genome-wide distribution of protein- and non-coding genes in the human and mouse genomes. Percentages shown are calculated from the GENCODE version 22 (http:// www.gencodegenes.org).

In particular the GENCODE gene sets are used by the entire ENCODE consortium and by many other projects (eg. 1000 Genomes) as reference gene sets. These datasets revealed that a total of 62% and 75% of the human genome is covered by either processed or primary transcripts, respectively, with an overall estimation of 80% potentially functional sequence in human DNA. Coding and non coding transcripts are predominantly localized in the cytosol and nucleus, respectively, and protein coding genes expression is higher than non coding counterparts [132]. Approximately, 6% of all annotated coding and non coding transcripts overlap with small RNAs (sRNAs) and are probably precursor of these RNAs. Several RNA-seq analyses have shown that splicing events occur predominantly during transcription and are fully completed in cytosolic polyA$^+$ RNA [154]. Moreover, lncRNAs have canonical gene structures and histone modification, appear to be subjected to weaker evolutionary constraint than coding genes and are preferentially enriched in nucleus of the cells [25]. All these growing projects aimed at identifying all functional elements in the human genome, point out that the transcriptional regulation is controlled by complex interactions between DNA sequence, transcription factors, histone tail modifications, DNA methylation as well as by a new layer of regulation: the non coding RNAs [132].

Non coding RNA genes and the modern RNA world

The class of non coding RNAs catalized the interest in the field of functional genomics in the last decade [61]. They are mainly classified based on their size into small and long non coding RNAs (Tabella 1). Some of these RNAs have general housekeeping functions and include ribosome-associated RNA (rRNA), transfer RNA (tRNA) and small nuclear/nucleolar RNA (sn/snoRNAs). Different classes of short RNAs (miRNAs, siRNAs and piRNAs) posses regulatory functions in several cellular processes including cell [6, 20], cencer progression [37, 176] and immunity [49, 91]. The class of long non coding RNAs (lncRNA) encompasses a different set of functional transcripts characterized by a length over 200 nucleotides with no potential to encode for functional proteins of more than 30 aminoacids [86, 102].



ncRNA		Length (nt)	Function
SHORT			
miRNAs	Micro RNAs	21–23	In animals, associate with the miRNA-induced silencing complex (RISC) and silence the expression of target genes mostly post-transcriptionally
snoRNAs	Small nucleolar RNAs	60–300	Help the chemical modification of mRNAs, thereby influencing stability, folding, and protein-interaction properties
snRNAs	Small nuclear RNAs	150	Assist splicing of introns from primary genomic transcripts
piRNAs	Piwi-interacting RNAs	25–33	Associate with the highly conserved Piwi family of argonaute proteins and are essential for retrotransposon silencing in germline, epigenetic modifications, DNA rearrangements, mRNA turnover, and translational control also in soma
PASRs	Promoter-associated short RNAs	22–200	Enriched at the 5′ end of genes, within 0.5 kb of TSS. Can be transcribed both sense and antisense. Their function and biogenesis is not fully understood
TASRs	Termini-associated short RNAs	22–200	Can be transcribed both sense and antisense near termination sites of protein-coding genes. Their function and biogenesis is not fully understood
siRNAs	Short interfering RNAs	21–23	Processed from a plethora of genomic sources, both foreign (viruses) and endogenous (repetitive sequences). Canonically induce the degradation of perfectly complementary target RNAs
tiRNAs	Transcription initiation RNAs	15–30	Enriched immediately downstream transcriptional start sites (TSSs) of highly expressed genes. Their function and biogenesis is not fully understood
LONG			
NATs	Natural antisense transcripts	> 200	Transcribed from the same locus but opposite strand of the overlapping protein-coding sequence. Involved in gene expression regulation, RNA editing, stability, and translation
PALRs	Promoter-associated long RNAs	200–1000	Enriched at promoters, found to regulate gene expression
PROMPTs	Promoter upstream transcripts	200–600	Enriched at TATA-less, CpG-rich promoters with broad TSSs. Affect promoter methylation and regulate transcription
T-UCRs	Transcribed ultraconserved regions	> 200	Perfectly conserved between human, rat, and mouse. Frequently located at fragile sites and at genomic regions involved in cancer
Intronic RNAs		> 200	Transcribed from introns of overlapping protein-coding sequences. Involved in the control of gene expression, alternative splicing, and source for generation of shorter regulatory RNAs (
eRNAs	Enhancer-derived RNAs	> 200	Function still not completely understood. May functionally contribute to the enhancer function
LincRNAs	Long intervening (intergenic) RNAs	> 200	Gene expression regulation, regulation of cellular processes
uaRNAs	3 UTR-derived RNAs	< 1000	Derive within 3′ untranslated region (3′ UTR) sequences. Function still not clearly understood
circRNA	Circular RNA	100 to > 4000	Diverse, from templates for viral replication to transcriptional regulators

Table 1. Different classes of short and long regulatory non-coding RNAs [124]**.**

lncRNAs are commonly classified in association with annotated protein-coding mRNAs (Figure 2) and comprise the long intergenic ncRNAs (lincRNA), intronic lncRNA, sense or antisense lncRNA, competitive endogenous RNAs (ceRNAs) and enhancer RNA (eRNA). The prominent category of ncRNAs are sense ncRNAs that overlap coding mRNAs on the same strand and share some sequence with the latter, yet do not encode proteins [67, 24, 26, 94]. This class includes unspliced sense partially intronic RNAs (PINs) and spliced transcripts that combine exons from coding and non coding region of a gene [26, 94]. The intronic lncRNAs (TINs) are produced starting from the intron of a protein coding gene and these transcripts compose the majority (about 70%) of all non coding (non-rRNA) nuclear-encoded RNA. Some of these transcripts represent intronic circular ncRNAs (ciRNAs) that are produced from introns that escape debranching and are involved in the regulation of the expression of their parent genes [174]. Another class of lncRNAs, defined as natural antisense transcripts (NATs), overlap the opposite DNA strand of the associated protein coding genes and occur in 50-70% of all protein-coding genes [17, 69]. LncRNAs are defined bidirectional when their expression and the expression of a neighboring coding transcript is initiated in close proximity and intergenic lncRNAs if they lie between two different coding regions with no overlap [25]. Intergenic long ncRNAs are the most studied lncRNAs because they are independent transcritpional units and are more likely associated with an intrinsic function than being transcriptional noise. LincRNAs are the main subject of this thesis (Figure 2).

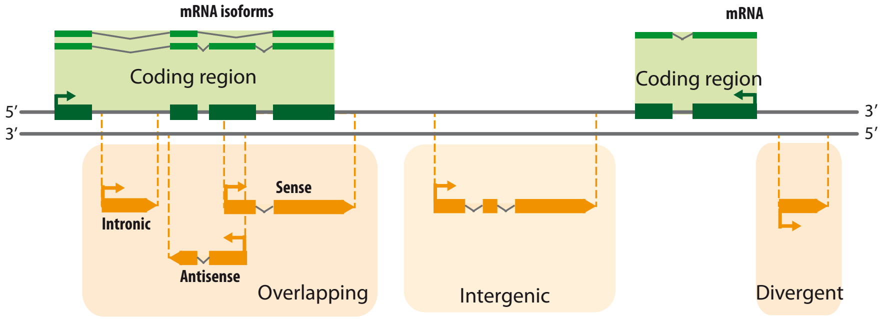


Figure 2. Long non-coding RNAs classification. Long non-coding RNAs are generally defined after their genomic and transcriptional context. The figure shows the possible relationships between long ncRNAs (in orange), coding regions (in dark green), and transcribed mRNA (in light green) [120].

LncRNAs are also classified according to localization in the cells: nuclear lncRNAs and citoplasm lncRNAs. Nuclear lncRNAs have been found so far to be mainly implicated in the recruitment of chromatin modifiers in specific genomic loci [133, 4, 48, 72]. In particular they recruit DNA methyltransferase (DNMT3) and histone modifiers as polycomb repressive complex PRC2 [135, 175] and H3K9 methyltransferase [111, 122]. Not only the recrutiment of repressive complex trigger transcriptional repression by formation of repressive heterocromatin [34], but the act of lncRNA transcription itself can negatively affects gene expression [98, 78]. Moreover, lncRNAs can work also as transcriptional activators through the recruitment of chromatin-modifying complexes, such as H3K4 methyltransferase MLL1 complexes [164, 7] and by changes in three-dimensional chromatin conformation mediated by the activation of specific enhancer regions [164, 119, 85].

It is possible to distinguish nuclear lncRNAs by their course of action: cis- and trans-acting lncRNAs (Figure 3). The first class is involved in the control of the expression of genes located in the proximity of their transcription sites and sometimes can spread their action to long distances on the same chromosome [34]. trans-acting lncRNAs can both activate or repress the expression of genes located in indipendent loci [135, 4, 48]. Targeting mechanisms of lncRNAs and retention to their binding sites are still largely unknown, but different hypothesis have been proposed, including formation of RNA-DNA triplex [143], DNA recognition by RNA structures [19] and bridging protein recruitment [62].

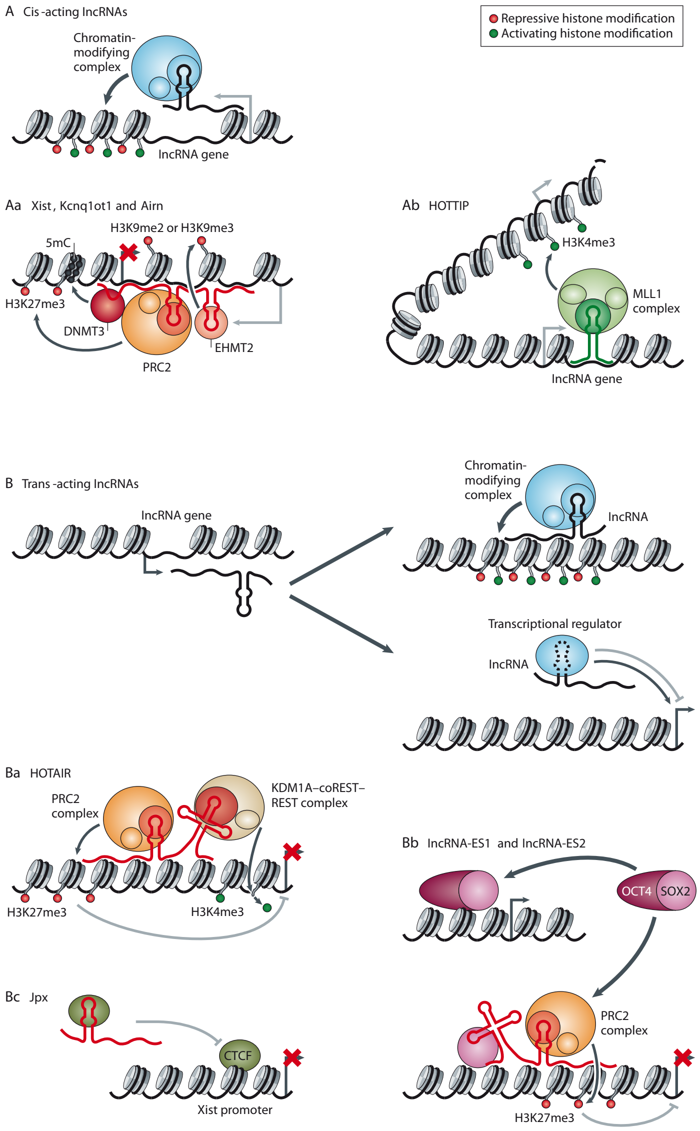


Figure 3. Models of nuclear lncRNA function. Examples of long non-codingRNAs (lncRNAs) that regulate transcription in cis (part A) and in trans (part B), by recruiting specific transcriptional regulators onto specific chromosomal loci. Aa) *Xist*, *Kcnq1ot1* and *Airn* are lncRNAs involved in the formation of repressive chromatin through recruitment

of DNMT3 which induces DNA methylation, PRC2 related to H3K27me3 production and EHMT2 that produces H3K9me2 abd H3K9me3 [79]. Ab) *HOTTIP* recruits MLL1 complex inducing the activation of transcription through H3K4me3 activation [164] Ba) *HOTAIR* is a trans-acting regulator of HOXD genes [135] that recruits in the same locus PRC2 and H3K4 demethylation complexing KDM1A-coREST-REST (lysine-specific histone demethylase 1A-REST corepressor 1-RE1-silencing transcriptional factor) Bb) *lncRNA-ES1* and *lncRNA-ES2* associate with both PRC2 and SOX2 inducing the silencing of SOX2-bound developmental genes to control embrionic stem cell pluripotency [116] Bc) *Jpx* lncRNA interacts with the transcriptional repressor CTCF to block their binding to Xist promoter activating Xist transcription [151] [34].

Cytoplasmic lncRNAs mediated gene regulation mechanisms [4] (Figure 4). These lncRNAs often show sequence complementarity with transcripts generated from either the same chromosomal locus or indipendent loci and this implies that they can modulate the control of translation by base pairing recognition of the target. In the same way, lncRNAs can modulate mRNA stability: both β-site APP-cleaving enzyme 1-antisense (BACE1-AS) [31] and tissue differentiation-inducing non- protein-coding RNA (TINCR) [75] increase the stability of their target mRNAs, whereas half-STAU1 (staufen double-stranded RNA-binding protein 1)-binding site RNAs (1/2sbsRNAs) [41, 163] decrease target mRNA stability (Figure 4).

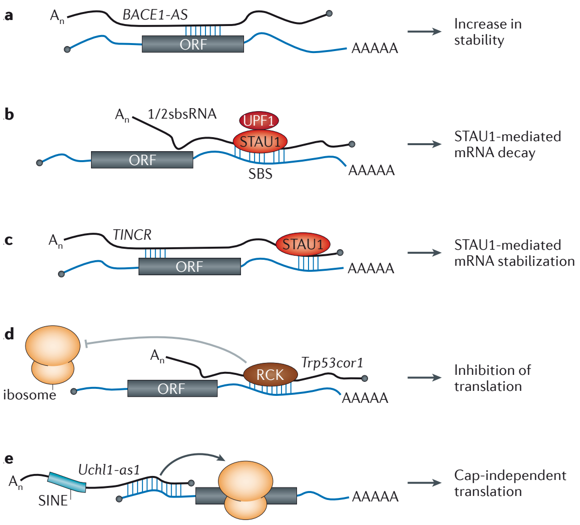


Figure 4. Models of cytoplasmic lncRNA function. The mechanism involved a base-pairing complementarity between lncRNAs and target RNA sequence. a) complementarity between BACE1 mRNA and ist antisense transcripts *BACE1-AS*, increasing BACE1 stability and protein expression [31]. b) STAU1-mediated mRNA decay mechanism based on intermolecular base pairing between Alu element in 3'UTR of STAU1 mRNA and Alu element within *1/2sbsRNA* lncRNA with recruitment of UPF1 RNA helicase. c) Contrariwise, STAU1-mediated mRNA stabilization with the implication of *TINCR* that recognizes its target mRNA through a 25base-long motif. [75]. d) Mechanism of inhibition of translation that implicates *Trp53cor1* lncRNA together with RNA helicase RCK [171]. e) *Uchl1-as1* lncRNA induced translation upon stress induction [14] [34].

A peculiar mode of action of lncRNAs is represented by competitive endogenous RNAs (ceRNAs) class. They are transcripts independently originated by protein coding genes and are previously described as transcribed retropseudogenes that retain the miRNA-binding function of their parent miRNAs [141]. The main function proposed for ceRNAs is de-repressing the level of protein coding gene that share with ceRNA the same miRNA response elements (MREs), also defined as miRNA "decoys" or "sponges" [141]. The transcriptonal regulation mediated by ceRNAs represents and elegant mechanism by which lncRNAs control the function of protein coding genes through miRNA mediators [58]. Recently, an additional example of ceRNA was found in a newly identified class of circular RNAs (circRNAs) [52, 51, 101], which function as sponges for miRNAs in neuronal cells (Figure 5). Whereas the linear ceRNAs have a short half-life that allows a rapid control of sponge activity, circRNAs have much greater stability and their turnover can be controlled by the presence of a perfectly matched miRNA target site [52, 51, 101].

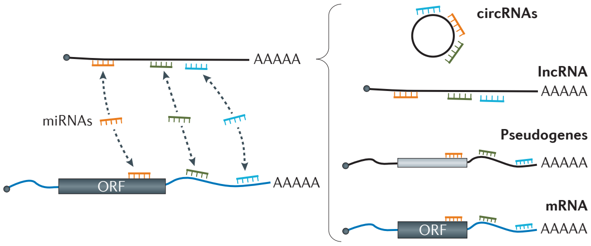


Figure 5. Model of ceRNA action. This mechanism is base on complementarity between miRNAs and different targets (circRNAs, lncRNAs, pseudogenes and mRNAs). This represents a new type of regulatory circuitry in which different types of RNAs (both coding and non-coding) can crosstalk to each other by competing for shared miRNAs [52, 51, 129, 141, 15] [34].

Recent evidences ascribed a potential roles of long non coding RNAs at enhancer regions [141]. In particular, RNA-seq, ChIP-seq, and chromatin-conformation-capture studies have defined that these genomic regions transcribe for new transcripts defined enhancer RNAs (eRNAs). The putative enhancer regions are marked by high levels of H3K4me1 and H3K4me2 relative to the H3K4me3 [154]. eRNAs are 5' capped [160, 63] but don't exhibit 3' polyadenilation or splicing events. They have a rate of transcription frequency comparable to protein coding genes [63] but the half-life is shorter compared to lncRNAs and other mRNAs. Whether eRNAs are merely a correlation or a functional component of enhancers activity is actually a source of debate. Three possibilities have been considered with respect to the physiological roles of enhancer transcription. The first hypotesis considers enhancer transcription as "noise" from the spurious engagement of RNA PolII complexes to the open chromatin environment of enhancers. The second possibility hypothesizes that it is the process of transcription, not the features of the eRNA transcript itself, that is necessary for the activating functions of enhancers. The third possibility is that the RNA transcripts per se functionally contribute to enhancer activity [85, 58]. These possibilities are not mutually exclusive and are under investigations.

Predicting long non coding RNAs using RNA sequencing

The majority of non coding RNAs was initially discovered through earlier studies on expressed sequence tags (ESTs) and tilling microarray [61]. These studies identified a large and previously unknown repertoire of transcripts that have been subsequently validated by deep sequencing approaches. As opposed to conventional microarray-based technology that is used to profile only known transcripts, deep-sequencing approaches, in particular RNA-seq, facilitate a genome-wide expression profiling including the identification of novel and rare transcripts like long non coding RNAs and novel alternative splicing isoforms [96, 106, 107, 109, 110, 148].

To better understand the potential functional role of newly described lncRNAs, the first step is their accura te identification and annotation. Non coding RNAs are now more easily and more accurately identified by transcriptome sequencing (RNA-seq). This technology allows the reconstruction of virtually whole transcriptome and the discovery of novel transcripts and of their relative abundance [58]. Many RNA-seq studies have now been performed aimed at the characterization of long non coding RNAs in different organisms, cell types and tissues [12, 47, 172, 150, 125, 112, 84, 26, 5]; at the same time a lot of computational approaches for the reconstruction of the transcriptomes have been developed that revealed several features of lncRNAs. For example the classification of lncRNA loci with respect to protein coding genes can be achieved only through the analysis of RNA-seq data. Moreover defintion of lncRNAs transcript abundance, exon composition and splicing efficiency could not be obtained without RNA-sequencing approaches [58]. Nonetheless other NGS strategies are necessary for lncRNAs prediction: full-lenght cDNA sequencing, chromatin state maps and RNA polymerase II occupancy (Figure 6).

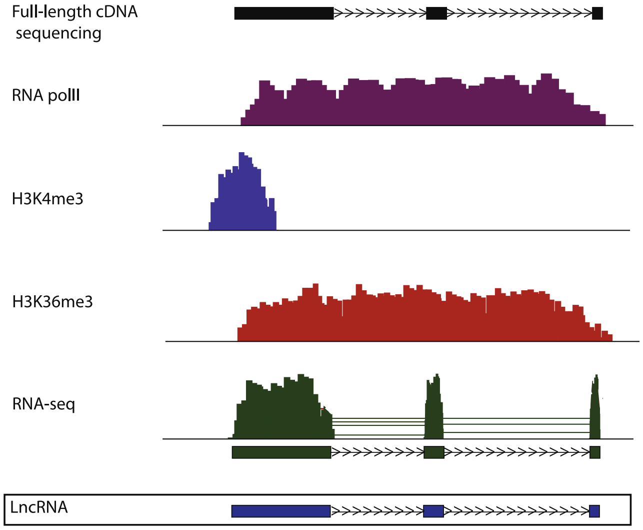


Figure 6. Common methods for predicting lncRNA loci using NGS. The bottom panel represents a lncRNA that is transcribed from the genome. The remaining panels represent the expected distribution of sequenced reads across this lncRNA locus [58].

The optimal strategy to obtain accurate transcript models is the full-lenght cDNA sequencing. This strategy was adopted by FANTOM project [70] that initially annotated more than 21000 cDNA clones, 3000 of which were defined as "unclassifiable" because showed no homology to known protein coding genes, no discernible protein motif and no open reading frame longer than 100 aa [70]. The FANTOM data were subsequently integrated with transcriptional start site (TSSs) informations using CAGE (Cap-Analysis of Gene Expression) technology and about 30% of the total number of transcripts were annotated as non coding RNAs [13]. Moreover, other studies of annotated coding loci demonstrated that many of them transcribed additional co-regulated non coding sense and antisense transcripts [69, 134]. The advent of ChIP (Chromatin ImmunoPrecipitation) coupled with NGS (ChIP-seq) provided a more comprehensive chromatin state maps to investigate the regulatory elements and region of the genome [3, 104, 139]. For example, the relative aboundance of H3K4me1 and H3K4me3 are indicative markers of active promoters and enhancer elements [54], while the "K4-K36 domains", characterized by H3K36me3 that denotes transcribed gene bodies coupled with H3K4me3 that is also a marker for the definition of TSS, outside annotated protein coding genes could potentially described novel lncRNA loci [46] (Figure 6). Finally, the information on all the fragments bound to RNApol II and subjected to NGS is important to infer transcription at these loci and this approach was widely adopted to identify non-coding transcription [58]. This method, though, has limitations: RNApol II occupancy is not indicative of transcriptional elongation, for this reason it is not possible to separate random RNApol II binding events from non-random occupancy related to transcription. Moreover, it is difficult to associate intergenic RNApol II binding to lncRNA loci when these are in close proximity to protein coding genes [66]. Despite these limitations, the studies based on these approaches have provided an initial insights into the nature of intergenic transcription.

A workflow for the discovery of lncRNAs has been outlined in the last years and followed in several studies (Figure 7).

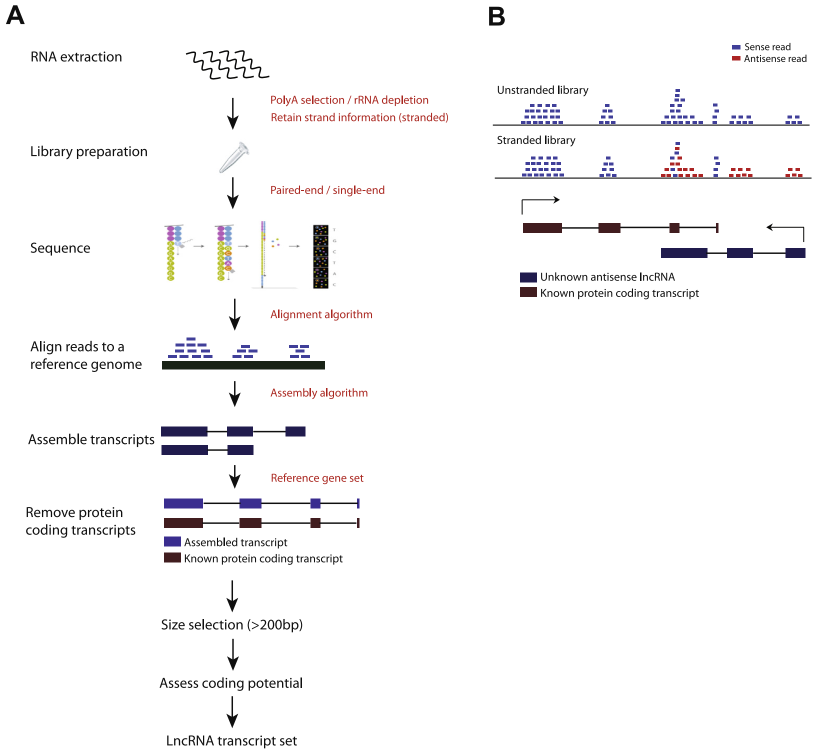


Figure 7. A) workflow for predicting lncRNAs using deep RNA sequencing. B) Representation of unstranded vs. stranded RNA seq libraries after mapping to a reference genome [58].

Ribosomal RNA removal: The first step is based on removal of ribosome-associated RNA (rRNAs) that represents the 90% of the total RNA. If this fraction is maintained in the sample, the majority of the sequenced reads would align only to rRNA, leaving relatively few reads for assembly and abundance estimation of both coding and non coding transcripts. For this reason RNA-sequencing experiments are commonly performed only on RNA molecules that contain a polyA tail or on RNA specifically depleted of rRNA [58]. Studies conducted by the ENCODE consortium revealed that only 3.3% of transcripts fell exclusively in the polyA$^-$ fraction [26], then using polyA$^+$ selection in a RNA-seq experiment not only ensures the recovery of the vast majority of annotated transcripts, but also removes pre-mRNA molecules that would reduce coverage depth across exons [58].

Paired-end versus single-end sequecing: sequencing is performed on cDNA libraries that comprise all the fragments corresponding to the set of transcripts of a given sample. The sequencing strategy to "read" these fragments could be single-end (SE) if just one end of each fragment is sequenced, paired-end if read pairs corresponding to both the start and the end of each cDNA fragment are generated [58]. The length of the reads is generally between 50-130 bp with the currently available technologies (e.g. Illumina), and the majority of these are present uniquely in the genome because the number of fragments is very large. As library fragments are about 300-400 bp long, only a portion of the cDNA fragment can be covered by the sequencing read, so it is possible that two single-end reads map to the same start or end position within the fragment. Moreover single-end reads that include repetive sequence are excluded because they cannot mapped uniquely in the genome assembly lowering the number of mapped reads. Paired-end solution improves the sequencing experiment providing long-range positional information that is crucial in lncRNA discovery as these transcripts are not only expressed at low levels but they are also enriched for repetitive elements.

Stranded protocols: the information about the strand of origin of a transcript is lost with RNA-seq protocols that started from a double-strand cDNA for the preparation of the library. Recently, stranded protocols (also known as “directional”) have been developed [58]. This is an important advantage in RNA-seq technology, especially for the discovery of lncRNAs, because the conservation of the information about strand allows the discrimination between overlapping sense and antisense transcripts. When this information is not availble, the most used computational tools (e.g. Cufflinks) [156, 58], infers the transcriptional direction of a transcript using canonical splice site information, but this strategy isn’t useful when non-canonical splice sites are present or when a transcript is monoexonic. For this reason, stranded RNA-seq has clearly improved the understanding of pervasiveness of non coding sense and antisense transcription (Figure 7) [58].

Mapping algorithms: Following the primary analysis of quality control and possibly trimming of sequenced reads, the first step to computationally identify novel lncRNAs is the mapping of the sequenced reads to a reference genome if available [58]. Given the large number of reads produced in a RNA-seq experiment (ranging between 20-100s of million reads for single sample), a lot of algorithms and software have been developed such as Bowtie [77], BWA [82], Stampy [93] and GEM [95] to efficently map the reads to a reference genome. These tools store the genomic coordinates of short oligomers as an indexed genome using hash table indexing and compression with Burrows-Wheeler transform or other indexing scheme [95] and these features confer high performance. Several factors contribute to the accuracy of the alignments: first, most algorithms miss some read-genome matches due to non-exhaustive searching. Second, the sequenced reads include biological variability derived from SNPs variant, small indels as well as sequencing errors and finally, the sequenced reads could potentially map to two genomic regions with large gaps that span spliced introns [58]. This is a crucial point in RNA-seq technology because the detection of novel intron-spanning junctions is important for the discovery of lncRNAs. Several algorithms are being developed to efficiently and accurately identify these features, including TopHat [155], spliceMap [1], MapSplice, GSNAP [168], RUM [44] and STAR [28] that is the official ENCODE Project mapper. Most of these aligners provide an initial mapping to a reference genome and/or transcriptome using a short read alignment tool as Bowtie or BWA [155, 44, 168], afterwards the unmapped reads are aligned to inferred or known junctions [155] or to the genome using gapped aligner as BLAT [44]. Single-step methods have been developed, such STAR [44, 28], to improve the sensitivity. Moreover the memory usage and run time are important characteristics in this step of analysis, therefore the computing capacity available in the lab needs to be taken into account: for example, STAR is the fastest mapper with low false positive rate in junction discovery, but it uses a lot of memory [28], instead GSNAP poorly performs in terms of number of reads mapped and also memory usage. TopHat is the mapper that has been used for a long time, specially in the first studies on long non coding identification [12, 172, 150, 125, 112, 84, 5, 88]. TopHat is considered one of the best in junction mapping and, despite improvements of recently developed tools, it maintains the advantages in speed and memory usage. Nevertheless, the sequencing technology is under continuous development producing longer reads to improve detection therefore more sensitive and specific alignment tools such RUM and STAR will be increasingly used.

Transcripts reconstruction: It is possible to asses the identification and quantification of lncRNA transcripts by two approaches: quantification and coverage estimation of mapped reads over lncRNA existing annotation sets [12, 46, 47, 142, 25] or *ab initio* transcripts reconstruction and then novel annotation definition if no annotation is available for the species of interest. The second approach is also applied if the aim is the identification of novel lncRNAs in a specific biological context. The power of RNA-seq in the detection of novel lncRNAs (and transcripts in general) derives from the identification of novel splicing events and then definition of the exonic structure thanks to the distribution of sequenced reads across splice junction, whereas other sequencing technology like ChIP-seq provides information only about active transcriptional regions [58]. Two general strategy could be adopted for transcripts assembling: *mapping-first* approaches with Cufflinks [156] and Scripture [47] and *assembly-first* methods to perform *de novo* reconstruction of transcriptome, including Trinity [43], SOAPdenovo [83], transAbyss [137] and Oases [145]. The first strategy is often successful in reconstructing transcripts, recapitulating known isoforms as well as novel lncRNAs transcript models because it uses the splice reads to reconstruct individual transcripts after the alignment to reference genome (Figura 8). This strategy cannot be applied when a reference genome is not available or when the reference transcriptome includes many rearrangements, like in cancer genomes [58].

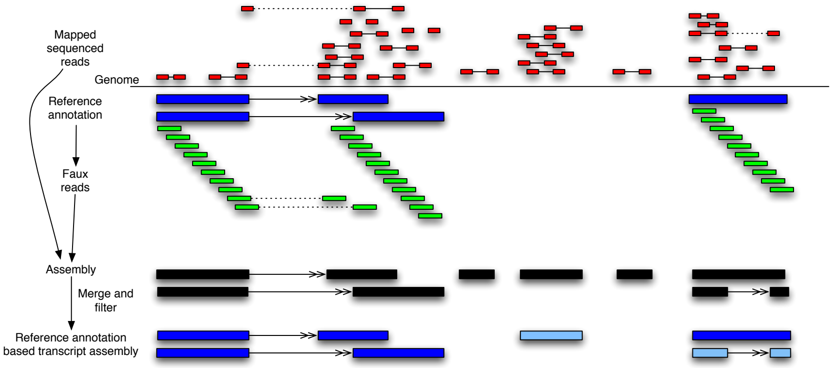


Figure 8. An overview of Cufflinks assembly method. First paired-end reads (mates shown connected by solid lines) are mapped to the genome using a spliced read mapper that can map reads across junctions (shown in dotted lines). The reference annotation (blue) is used to generate faux-read alignments that tile the transcripts (green). The faux-read alignments are used together with the spliced

read alignments to generate a reference genome based assembly (black). This assembly is merged with the reference annotation, and “noisy” read mappings are filtered resulting in all reference annotation transcripts in the output (blue) as well as novel transcripts (light blue) [136].

The *de novo* methods are based on strategy originally adopted for genome assembly: short sequences (k-mer) are extracted from sequenced reads and all the reads that share a partcular k-mer are overlapped and assembled into contigs, scaffolds and eventually transcript models [58] (Figura 9).

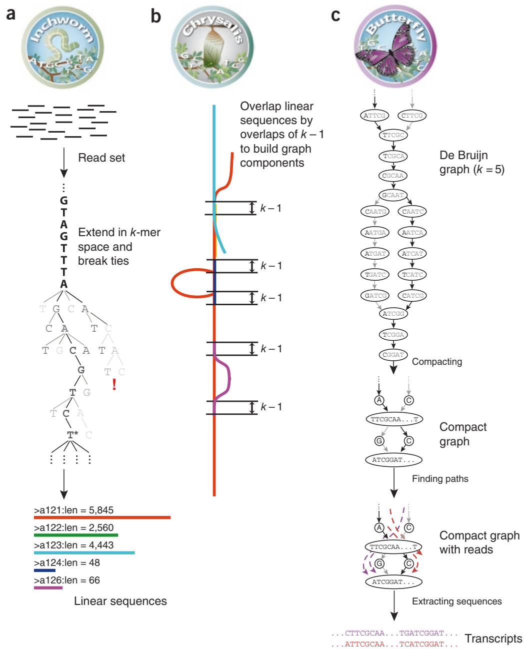


Figure 9. Overview of Trinity. (a) Inchworm assembles the read data set (short black lines, top) by greedily searching for paths in a k-mer graph (middle), resulting in a collection of linear contigs (color lines, bottom), with each k-mer present only once in the contigs. (b) Chrysalis pools contigs (colored lines) if they share at least one k − 1-mer and if reads span the junction between contigs, and then it builds

individual de Bruijn graphs from each pool. (c) Butterfly takes each de Bruijn graph from Chrysalis (top), and trims spurious edges and compacts linear paths (middle). It then reconciles the graph with reads (dashed colored arrows, bottom) and pairs (not shown), and outputs one linear sequence for each splice form and/or paralogous transcript represented in the graph (bottom, colored sequences) [50].

To avoid the problem of non-uniformity of sequenced reads across transcriptomes, it is possible to use different length of k-mers, generating different assemblies that can be merged to correctly reconstruct the transcriptome [145]. A problem that arises in novel lncRNA identification is that these transcripts are mainly lowly expressed and the *de novo* assemblers do not perform well in the detection of transcripts with low expression values. Consequently, if a good reference genome is available, a *mapping-first* approach using Cufflinks show an high performance compared to *de novo* methods in term of number of full-length transcripts reconstructed [145], albeit it generates more false positive transcripts.

Protein coding potential: The evaluation of the coding potential for a transcript is the most important step in the pipeline for the identification of lncRNAs. First of all, the assembled transcripts representing known protein coding genes need to be removed, so that all the transcripts that overlap known protein coding annotation available in different sources (Ensembl, UCSC or RefSeq) are not considered. However, the datasets are not completely overlapping in their annotations: it is possible that part of transcripts annotated as protein coding in one database haven't their counterparts in other sources. For this reason, if multiple annotation are available for the organism of interest, it is necessary to merge all the information to asses the novelty of the potential lncRNA [58]. The second point is the assessment of the coding potential of an identified transcript. Recent studies about ribosome association of lncRNA transcripts [60] suggested that some of these transcripts have the potential to be translated and generate small low abundance peptides, whose function is independent of the RNA-dependent functional moieties [58]. Two computational methods have been developed to evaluate the coding potential of lncRNAs transcripts: *open reading frame (ORF)-based* and *comparative sequence analysis*. In the first approach, the transcript is translated in all the six frames into aminoacids sequences and then compared, using BLASTX, against known protein coding sequences or domain (PFAM and SwissProt database). This method, though, does not allow a proper classification of transcripts as coding or non coding, in fact bona fide lncRNAs, such as transcribed pseudogenes that may derived from protein coding genes and thus maintain sequence similarities to their parent gene, could be misclassified as being protein coding. Moreover, another criteria considered in the evaluation of coding potential is the length of the ORF: the majority of protein coding genes have ORF longer than 200 nucleotides. The Coding Potential Calculator CPC [74] is commonly used by the scientific community for assessing the protein coding potential. It is a SVM (Support Vector Machine) framework that includes information on both ORFs homology and integrity: the SVM is trained on annotated set of protein coding and non coding genes and assigns a score to the analyzed transcript on the basis of the distance from the classification hyperplane [58]. The approaches based on comparative sequence analysis exploit information from multiple species alignments, considering the conservation of aminoacid sequence. PhyloCSF [90] is a method that discriminates between coding/non coding transcripts based on the codon substitution frequency (CSF): it relies on the expected frequency of nucleotide substitution at the three positions in a codon between a query sequence and homologous sequences [89]. The statistical method is based on the comparison of two model for each transcript: one with coding model parameters and one with non coding model parameters, then transcripts are classified as coding or non coding using maximum likehood estimation. Recently, other tools have been developed to improve accuracy and run time, e.g. iSeeRNA and CPAT [149, 165].

"gold standard" annotation: the current annotation of lncRNAs is not stable and is a source of debate because the discovery of these transcripts is clearly not yet saturated and still at a preliminary stage [100]. Several international consortia and platforms as well as HUGO Gene Nomenclature Committee (HGNC), that provided the first guidelines for lncRNA annotation, are attempting to develop e more unified system for lncRNA annotation. The term "lncRNA" itself is temporary because generally associated only to transcripts length, for this reason a unique identifier should be given to each novel transcripts. One proposed nomenclature comprises a catchy and easily recalled name with a unique numerical identifier of the sequence and genomic coordinates (referring to a specific genome assembly) for the exons of the transcript and its isoforms that can always be retrieved irrespective of changing genome coordinates and annotations [100]. Other ways to disseminate or describe lncRNA could be to provide contextual information in term of anatomical position relative to adjacent protein-coding genes (for example, overlapping antisense, overlapping sense, intronic, bidirectional and intergenic). To provide positional criteria that might help the scientific community to univocally identify lncRNAs in the literature, Mattick and Rinn recently proposed to name transcripts as follows:

Antisense transcripts: denoted as AS-GENENAME

Intronic transcripts: denoted as INT-GENENAME

Bidirectional transcripts: denoted as BI-GENENAME

Intergenic transcripts: denoted as LINC-X, where X represents a number. Although description of a lncRNA by naming it LINC-Y, where Y is the most 3'-adjacent gene, might be useful for its identification, this can generate problems when comparison across genomes is performed in which neighboring genes may no longer be syntenic.

Overlapping transcripts: denoted as OT-GENENAME

Descriptive transcripts: If function is ascribed to a given lncRNA through experimental studies, it is reasonable to name the gene accordingly. The HUGO committee has imposed some rules to prevent the use of commonly used or already existing annotations.

In the next years, the human annotation should become transcript centric rather than gene centric, because it is increasingly clear that the transcriptome is far more complex than the genome.

Functional interactions among microRNAs and long non coding RNAs

In recent years there is an increasing interest in understanding the cross-regulation between microRNAs and lncRNAs and of the molecular mechanisms underlying reciprocal control of their activity [169]. Several studies have begun to uncover the interaction among mammalian lncRNA and miRNAs and different mechanism are proposed: i) targeting of lncRNAs by microRNAs to reduce their stability ii) lncRNA as molecular decoys or sponges of microRNAs [141] iii) competition between microRNAs and lncRNAs for binding to shared target mRNAs and iv) lncRNAs as precursor for the generation of miRNAs to silence target mRNA [169].

microRNA-triggered lncRNA decay

microRNAs control the level of abundance of numerous lncRNAs (Figure 10, Table 2). Since the involvment of lncRNAs in different cell functions, as proliferation, differentiation, senescence, apoptosis and transformation, changes in the level of abundance of lncRNAs directly alter and influence the cellular response and pathologic processes [169]. Examples lncRNAs regulated by this mechanism are: *lincRNA-p21*, *HOTAIR*, *MALAT1*, *LOC285194*, *PTCSC3*, *H19* and *lincRNA-RoR*.

The stability of *lincRNA-p21*, a lncRNAs transcriptionally activated by p53, is related to changes in its turnover rate: miRNA let-7b together with RBP (RNA-binding protein), HuR and Ago2 contribute to lowering the stability of the lincRNA in human cervical carcinoma cells. This was demonstrated through overespression of let-7b that facilitated the degradation of the lincRNAs, as well as the depletion of HuR stabilized lincRNA-p21. The recruitment of let-7b and HuR also influence the stability of *HOTAIR*, suggesting that the mechanism of lncRNA decay mediated by HuR-enhanced microRNA interactions is widely shared [170]. In this model, HuR binds several microRNAs, including different members of let-7b family and also promotes the interation between Ago2 and let-7b [170, 108]. The high levels of *HOTAIR* are important for the formation of a platform to encourage the interaction of associated RBPs (for example Dzip3 and Mex3b, two E3 ligases), increasing the ubiquitination of Ataxin-1 and Snurportin-1. Moreover, miR-34a is involved in the decrease of *HOTAIR* stability in human prostate cancer cell lines [18]. Also *MALAT1* is targeted by microRNA-Ago2-RISC in human primary glioblastoma and human Hodkin cell line L428 [81]. In particular, silencing of Ago2 or miR-9 increased the steady-state of *MALAT1,* whereas the overexpression caused its decrease. This event was observed in the nucleus, highlighting a nuclear decay-promoting function for this miRNA. The tumor suppressor *LOC285194*, a p53-inducible lncRNA, is the target of Ago2 and miR-211 in colon cancer cell line HTC-116 [92], *PTCSC3* is another cancer-related lncRNA targeted by miR-574-5p in tyroid cancer [33], *H19* includes 4 let-7 target sites within its sequence [64] and its downmodulation by these miRNAs promoted myotube formation in C2C12 myoblats [64]. Finally, *lincRNA-RoR* is targeted by many miRNAs including the putative rugulators miR145-5p, miR-181a-5p and miR-99b-3p that specifically decrease the activity of the lncRNA in human embryonic stem cell [166].

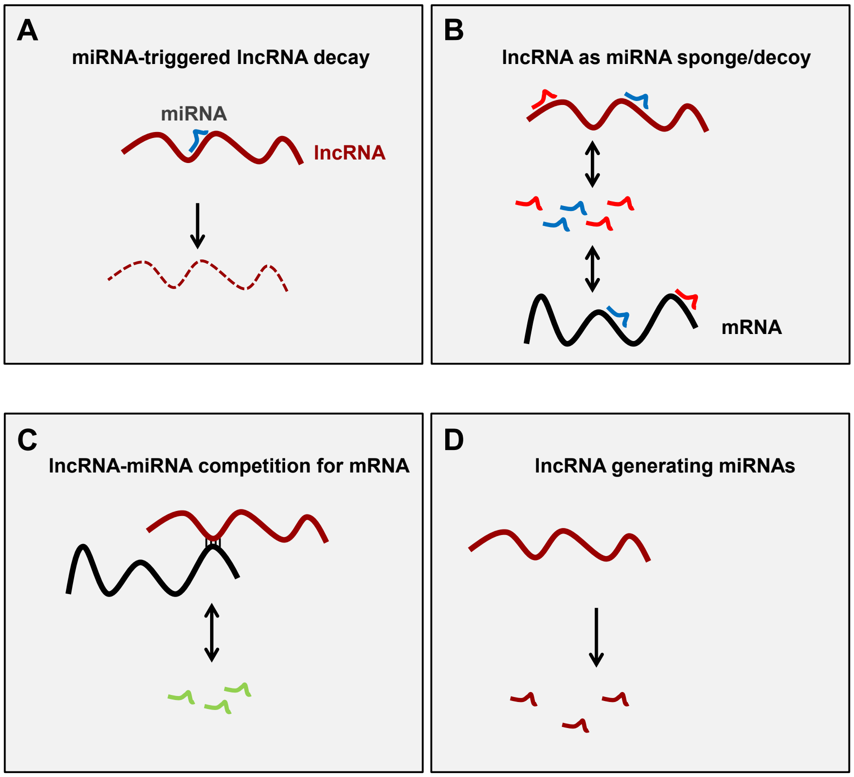

A
miRNA-triggered lncRNA decay
miRNA
lncRNA
B
lncRNA as miRNA sponge/decoy
mRNA
C
lncRNA-miRNA competition for mRNA
D
lncRNA generating miRNAs

Figure 10. Modes of direct post-transcriptional interaction among lncRNAs and miRNAs. Schematic of the major forms of interplay among lncRNAs and miRNAs. (A) Some lncRNAs are degraded by miRNAs, as described for lincRNA-p21 (degraded by let-7), HOTAIR (by let-7), MALAT1 (by miR-9), LOC285194 (by miR-211), PTCSC3 (by miR-574-5p), H19 (by let-7), and lincRNA-RoR (by miR145-5p and miR-181a-5p). (B) Other lncRNAs can serve as sponges/decoys for microRNAs, as described for linc-MD1 (sequestering miR-133 and miR-135), circular RNAs (sequestering miR-7, although many more examples are predicted to occur), lincRNA-RoR (miR-145-5p), and H19 (let-7). (C) A few examples of lncRNAs that compete with miRNAs for binding to

mRNAs have also been reported, including BACE1AS, which competes with miR-485-5p for binding to BACE1 mRNA, and ncNRFR, which competes with let-7 to derepress let-7 site-bearing mRNAs to promote carcinogenesis. (D) Several lncRNAs also produce microRNAs and other small RNAs, as shown for linc-MD1, which generates miR-206 and miR-133b, H19 (generates miR-675), and RMRP (generates miRNAs RMRP-S1 and RMRP-S2) [169].

lncRNAs as miRNA sponges/decoys

Recent evidence showed that also lncRNA can affect the levels and function of microRNAs (Figure 10, Table 2). It has been proposed that the levels of microRNAs in the cytoplasm (but also in the nucleus) can be titrated by lncRNAs that harbor similar microRNA target sequences so they can sequester microRNAs away from mRNAs [35]. These lncRNAs are also known as competing endogenous (ce)RNAs [141, 68, 153]. Some examples include *PTENP1*, that is the first ceRNA described (ref 56 yoon) involved in the regulation of the tumor suppressor PTEN. *PTENP1* RNA elicited this influence acting as a decoy for PTEN-mRNA-targeting microRNAs. Also *linc-MD1* modulates the muscle differentiation process by competing with miRNAs for the binding to specific mRNA target [15]. This lincRNAs if overexpressed, accellerates muscle differentiation in mouse enhancing the expression of MAML1 and MEF2C mRNAs by reducing the level of miR-133 and miR-135, respectively. Moreover, it has been recently proposed a feed-forward regulatory loops in which high level of *linc-MD1* in early stages of myogenesis sponges miR-133 away from HuR, allowing it to accumulate in cells and establish a myogenic expression gene program [80]. *lincRNA-RoR* and *H19* described above are also involved in sponge activity through self-regulatory loop: *lincRNA-RoR* transcription is activated by Nanog, Oct4 and Sox2 (in turn regulated by miR-145-5p) allowing the maintenance of hESC pluripotency whereas *H19* is an antagonist of let-7 during muscle differentiation [64].

lncRNA	microRNA	Interplay	Consequences
lincRNA-p21	let-7b	lncRNA decay	Translational repression of JUNBand CTNNBmRNAs
HOTAIR	let-7a, miR-34a	lncRNA decay	Ataxin-1 and Snurportin-1 ubiquitination
		lncRNA decay	Tumorigenesis
MALAT1	miR-9	lncRNA decay	Repression of MALAT1
LOC285194	miR-211	lncRNA decay	Tumorigenesis
PTCSC3	miR-574-5p	lncRNA decay	Tumorigenesis proliferation
H19	let7a,b,g,i	lncRNA decay, decoy	Decay of let-7 target mRNAs in muscle cells
	miR-675	miRNA production	Proliferation, myogenesis
lincRNA-RoR	miR145-5p	lncRNA decay	Stability of Nanog, Oct4, and Sox2 mRNAs
	miR-181a-5p	miRNA competition	
	miR-99b-3p	lncRNA decay	
		lncRNA decay	
Linc-MD1	miR-133	miRNA competition	Muscle differentiation, and abundance of MAML1 and MEF2CmRNAs
	miR-135	miRNA competition	
	miR-206	miRNA production	
	miR-133b	miRNA production	
CDR1as/ciRS-7	miR-7	miRNA sponge	Neuronal function
Sry	miR-138	miRNA sponge	
BACE1AS	miR-485-5p	miRNA competition	Brain
ncNRFR	let-7b,c,d,e,f,g,i	miRNA competition	Tumorigenesis
	miR-98	miRNA competition	
RMRP	RMRP-S1	miRNA production	CHH (Cartilage-Hair Hypoplasia) disease
	RMRP-S2	miRNA production	

Table 2. Examples of direct cross-regulation among lncRNAs and miRNAs. LncRNAs (column 1) affecting the levels and/or activity of microRNAs (column 2) and vice versa are summarized (column 3); if known, the consequences on gene expression are also indicated (column 4) [169].

lncRNAs competing with miRNAs for interaction with mRNAs

LncRNAa can also compete with miRNAs for binding to target mRNAs (Figure 10, Table 2). *BACE1AS* lncRNA promotes the stabilization of BACE1 mRNA, that is target of miR-485-5p in human embryonic kidney, protecting it from RISC-mediated degradation from the possible access and binding of the miRNA [32]. The tumor-promoting lncRNA *ncNRFR*, when overexpressed, suppresses let-7 actions in colonic epithelial cell line YAMC, increasing the level of let-7 target mRNAs. The lncRNA contains a 22nt sequence identical to let-7a and similar to other microRNAs of the same family [36].

lncRNAs generating miRNAs

Finally, mature miRNAs can be generated by lncRNAs processing (Figure 10, Table 2). For example, *linc-MD1* generates miR-206 and miR-133b from an intron and an exon, respectively [15] whereas lncRNA *H19* creates miR-675, a process that is repressed by HuR [71], in turn suppressing translation of the Insulin Growth Factor Receptor (Igf1r) mRNA in mouse. Finally, a mitochondrial lncRNA *RMRP,* produces two miRNAs, RMRP-S1 and RMRP-S2, that are putative regulators of two mRNAs (PTCH2 and SOX4) that encode proteins linked to human Cartilage Hair Hypoplasia [138]. Many evindeces reveals that lncRNAs and microRNAs work jointly to contribute to a robust and dynamic control of gene expression through complex post-transcriptional mechanisms.

Bioinformatic tools to investigate lncRNAs-miRNAs crosstalk

Different resources have been developed to discover miRNA-lncRNA interactions: miRcode (http://www.mircode.org/), DIANA-LncBase (http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=lncBase/index) and the most comprehensive starBase v2.0 (http://starbase.sysu.edu.cn/). The first provides "whole transcriptome" human microRNA target predictions based on the comprehensive GENCODE gene annotation, including >10,000 long non-coding RNA genes. This tool identifies putative target sites based on established principles: seed complementarity and evolutionary conservation. DIANA-LncBase unravels miRNA-lncRNA putative functional interactions, hosting transcriptome-wide experimentally verified and computationally predicted miRNA recognition elements (MREs) on human and mouse lncRNAs. The experimentally supported entries available correspond to >5000 interactions between 2,958 lncRNAs and 120 miRNAs, while the computationally predicted interactions exceed 10 million. starBase v2.0 is focused on the identification of general RNA–RNA and protein–RNA interaction networks from 108 CLIP-Seq (PAR-CLIP, HITS-CLIP, iCLIP, CLASH) data sets generated by 37 independent studies. By analyzing millions of RNA-binding protein binding sites, the tool collect ∼9000 miRNA-circRNA, 16000 miRNA-pseudogene and 285 000 protein–RNA regulatory relationships. It also provide the most comprehensive CLIP-Seq experimentally supported miRNA-mRNA and miRNA-lncRNA interaction networks. These new available resources will contribute to expand our understanding of ncRNA functions and their coordinated regulatory networks.

Long intergenic non coding RNA: novel drivers in human lymphocyte differentiation

With the advent of RNA-seq technologies and their application in the study of non coding world, the interest toward lncRNAs has been rapidly growing as well as the understanding of their multiple cellular functions and possible involvement in different pathologies. In the last years many studies were focused on lncRNAs quantification in different tissues and cell types, which generated different lncRNA catalogs with little overlap due to the high specificity of these RNA molecules [124]. In fact, unlike protein coding genes whose number has been remarkably stable over the years, lncRNAs seem to be more cell-specific than mRNAs with a lower but dinamically expression in different stages of differentiation. For this reason, the immune system is an excellent context in which the knowledge on lncRNAs could be gained. Indeed, little is known about the role of lncRNAs in human adaptive immune system, while the involvment of these molecules in innate branch is most studied [87, 59]. The adaptive immune system represents a great system for the study of the role of lincRNAs in differentiation [124]. In particular, upon antigene recognition CD4$^+$ naïve T cells differentiate into distinct T helper subsets characterized by the expression of specific master trascription factors and release of different cytokines. In recent years the concept of distinct T helper cell subsets as terminally differentiated lineages has been revisited; there are increasing evidences that CD4$^+$ T cell populations can alter the range of cytokines they produce in response to environmental clues, exhibiting substantial plasticity. In this context, lncRNAs have a foundamental role in governing flexibiliy and plasticity or in the maintenance of cell identity together with lineage-specific transcription factors, cytokines and other ncRNAs. In particular, ncRNAs seems to act as fine-tuners of fate choices: they are involved in changes of extrinsic signals that can alter the phenotype [120, 140, 158] or, in the context of the stability of lineage identity, they can also directly interact on histone and DNA modifiers redefining this enviroment or acting as maintainers of cell identity.

LncRNAs must be considered as minor players in a huge interconnected network [105], interacting with few partners. This allows them to be more flexible and sensitive to the variation, without affecting the integrity of the whole network. Moreover lncRNAs, differently from the protein coding counterparts, provide robustness to the network and are the fastest in evolving sequences in the genome [76, 114, 130, 144]. Indeed, few protein coding genes have been lost from worms to humans and mutations occour in pathologies [39, 118]. Recent single-case or genome wide studies on lncRNAs have been conducted in mouse immune system (Table 3), whereas only few studies are available in the human context.

Sample	LncRNAs	Function
Granulocytes, monocytes, NK, B, naïve $CD8^+$ and $CD4^+$, memory human T cells; in vitro polarized precursors T-helper, T_H0, T_H1, and T_H2 human cells	240 lncRNAs associated with autoimmune disease (AID) loci (RNA-seq)	Analysis of the expression profile of the AID-associated lncRNAs
$CD4^-CD8^-$, $CD4^+CD8^+$, $CD4^+CD8^-$, activated $CD4^+$ mouse T cells	31423 lncRNAs (lncRNA microarray)	Expression analysis and prediction off unction
17 T-cell leukemia cell lines	Thy-ncR1 (expression profiling of 10 thymus-specific ncRNA)	Enriched in human immature cells; acts as a cytoplasmic riboregulator that reduces the level of MFAP4 mRNA
Naïve, memory, activated, non-activated mouse $CD8^+$ T cells	Over 1000 mouse and human lncRNAs (microarray)	Expression and conservation analysis
$CD4^-CD8^-$, $CD4^+CD8^+$, $CD4^+$, $CD8^+$ mouse thymic T cells, and thymus-derived T_{reg} cells. In vitro differentiated T_H1, T_H2, T_H17, and induced T_{reg}	1524 lincRNA genes (RNA-seq); LincR-Ccr2-5′ AS	Expression analysis and ChIP-seq data analysis to identify lincRNA genes and possible regulators. LincR-Ccr2-5′ AS is T_H2-specific and it reduces the expression of Ccr1, Ccr2, Ccr3, and Ccr5. It contributes to the migration of T_H2 cells
Infected Namalwa B lymphocytes	IFNA1-AS	Cytoplasmic post-transcriptional stabilization of IFN-α1 RNA masking a miRNA-binding site
Jurkat cells, primary lymphomas, lymphoma cell lines, $CD19^+$ B cells	Saf/FAS-AS1	Regulates the alternative splicing of Fas which is impaired in non-Hodgkin's lymphomas associated with poor prognosis
Activated human $CD4^+$ T cells	BIC RNA (EST library analysis)	Proto-oncogene, induced upon activation, sensitive to immunosuppressive drugs
Jurkat cells	NRON (shRNA knock-down screening)	Regulates NFAT subcellular localization as part of an RNA–protein complex
CEM-C7 CKM1, jurkat JKM1, human primary lymphocytes	GAS5	Necessary and sufficient for growth arrest. Acts competing from GREs
Human $CD4^+$, $CD8^+$ cells, PBMC	NTT	Unknown, it shows a similar expression pattern to IFN γR
Thymocytes	TEA	Instruct the activity of Jα promoters and recombination
Human T_H1 cells	NeST/Tmevpg1/IFNG-AS1	Dependent on STAT4, T-bet, and NFκB. Contributes to Ifng expression by binding WDR5 and alter H3K4me3
Human primary $CD4^+$ and $CD8^+$ T cells, primary and polarized (from $CD4^+$ and $CD8^+$ T) $CD4^+$ CM, T_H1, T_H2, T_H17, and T_{reg} cells; neutrophils, basophils, $CD8^+$ CM, B cells	GATA3-AS1	Specifically expressed in T_H2 cells

Table 3. Studies on lncRNAs in the adaptive immune system [124].

The importance of lncRNAs investigation in human immune system is underlined by the many debates about the differences between experimental animal model and human in term of immune responses [147, 152, 157]. More important, ncRNAs seem to be poorly conserved between human and mouse [114, 115] and these sentences is supported by the fact that over 80% of the human lncRNAs that arose in the primate lineage, only 3% are conserved across tetrapods and also lack of known orthologs outside vertebrate [97].

The first functional study on adaptive immune system in mouse and human was focused in T_H1 and T_H2 lymphocytes. In this study the authors showed that lincRNA Tmevpg1, selectively expressed in T_H1 cells via STAT4 and T-bet, is involved in the induction of IFN-γ expression in response only to T_H1 differentiation program and not in other cellular context [21]. GATA3-AS1 is another investigated lincRNAs, specifically expressed in primary T_H2 cells and involved in a co-regulation with GATA3 [173]. Another study showed that GAS5 lncRNA, expressed in human T lymphocytes, is accumulated in starving condition and contributes to growth arrest suppressing GR-mediated transcription [73, 167]. Other genome-wide analyses aimed at profiling the lncRNA transcriptome, have been performed on $CD8^+$ and $CD4^+$ T cells in mice model, providing different set of specifically expressed lncRNA genes [123, 57]. In $CD8^+$context, hundreds lncRNAs were identified in mouse genome, many of which were lymphoid cell-specific and differentially expressed in naïve, memory and effector $CD8^+$ cells. Indeed, 1524 lincRNA gene clusters were identified in a panel of $CD4^+$ subsets, that exhibited dynamic, cell- and activation state-specific expression. Also in B cells, mediators of the antibody-dependent humoral arm of the adaptive immunity, express lncRNAs. The antisense lncRNA Fas-AS1 controls the production of soluble Fas receptor (sFas) that binds Fas ligand to regulate Fas-induced apoptosis in B cell lymphomas [146]. Fas-AS1 inhibits the alternative skipping of the exon 6 of Fas that is required for the generation of sFas mRNA. Because serum sFas levels are associated with poor prognosis in non-Hodgkin's lymphoma [117], the Fas-AS1 lncRNA is a potential therapeutic target in this setting. Finally in B cells, chromatin remodeling associated with V(D)J recombination has been potentially linked to a widespread antisense intergenic transcription that occurs in the variable (V) region of the immunoglobulin heavy chain (Igh) locus [8, 159]. These few examples are just clues of the importance of lncRNAs in human immune system and further deeper analyses are necessary to highlight their role in this context.

Scope of the thesis

In this thesis we investigated and provided the first comprehensive transcriptome analysis of human lymphocytes, focusing on the expression of long intergenic non coding RNAs (lincRNAs). Given the high specificity of these genes, we not only considered lincRNA collected in public databases, but we adopted a de novo approach to identify novel lincRNAs that are specifically expressed in our cells. Moreover, applying stringent filters, we identified lincRNAs "signature" that can play a key role in lymphocyte differentiation. We focused our attention on a T_H1-specific lincRNA, called linc-MAF-4, that we demonstrated to be involved in the maintenance of T_H1 cell identity via an epigenetic-repression of MAF gene.

References

1. Au KF, Jiang H, Lin L, Xing Y, Wong WH. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic acids research* 2010, 38(14): 4570-4578.

2. Ball P. DNA: Celebrate the unknowns. *Nature* 2013, 496(7446): 419-420.

3. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. *Cell* 2007, 129(4): 823-837.

4. Batista PJ, Chang HY. Long noncoding RNAs: cellular address codes in development and disease. *Cell* 2013, 152(6): 1298-1307.

5. Belgard TG, Marques AC, Oliver PL, Abaan HO, Sirey TM, Hoerder-Suabedissen A, et al. A transcriptomic atlas of mouse neocortical layers. *Neuron* 2011, 71(4): 605-616.

6. Berardi E, Pues M, Thorrez L, Sampaolesi M. miRNAs in ESC differentiation. *American journal of physiology Heart and circulatory physiology* 2012, 303(8): H931-939.

7. Bertani S, Sauer S, Bolotin E, Sauer F. The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin. *Molecular cell* 2011, 43(6): 1040-1046.

8. Bolland DJ, Wood AL, Johnston CM, Bunting SF, Morgan G, Chakalova L, et al. Antisense intergenic transcription in V(D)J recombination. *Nature immunology* 2004, 5(6): 630-637.

9. Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. *Nature* 2011, 478(7369): 343-348.

10. Brinkman BM. Splice variants as cancer biomarkers. *Clinical biochemistry* 2004, 37(7): 584-594.

11. Brosius J. The fragmented gene. *Annals of the New York Academy of Sciences* 2009, 1178: 186-193.

12. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development* 2011, 25(18): 1915-1927.

13. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, et al. The transcriptional landscape of the mammalian genome. *Science* 2005, 309(5740): 1559-1563.

14. Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, et al. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* 2012, 491(7424): 454-457.

15. Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M, et al. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* 2011, 147(2): 358-369.

16. Chen L, Tovar-Corona JM, Urrutia AO. Increased levels of noisy splicing in cancers, but not for oncogene-derived transcripts. *Human molecular genetics* 2011, 20(22): 4422-4429.

17. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 2005, 308(5725): 1149-1154.

18. Chiyomaru T, Yamamura S, Fukuhara S, Yoshino H, Kinoshita T, Majid S, et al. Genistein inhibits prostate cancer cell growth by targeting miR-34a and oncogenic HOTAIR. *PloS one* 2013, 8(8): e70372.

19. Chu C, Qu K, Zhong FL, Artandi SE, Chang HY. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Molecular cell* 2011, 44(4): 667-678.

20. Ciaudo C, Servant N, Cognat V, Sarazin A, Kieffer E, Viville S, et al. Highly dynamic and sex-specific expression of microRNAs during early ES cell differentiation. *PLoS genetics* 2009, 5(8): e1000620.

21. Collier SP, Collins PL, Williams CL, Boothby MR, Aune TM. Cutting edge: influence of Tmevpg1, a long intergenic noncoding RNA, on the expression of Ifng by Th1 cells. *Journal of immunology* 2012, 189(5): 2084-2088.

22. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012, 489(7414): 57-74.

23. Crick F. Central dogma of molecular biology. *Nature* 1970, 227(5258): 561-563.

24. Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, et al. Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome research* 2007, 17(6): 746-759.

25. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research* 2012, 22(9): 1775-1789.

26. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. *Nature* 2012, 489(7414): 101-108.

27. Djebali S, Kapranov P, Foissac S, Lagarde J, Reymond A, Ucla C, et al. Efficient targeted transcript discovery via array-based normalization of RACE libraries. *Nature methods* 2008, 5(7): 629-635.

28. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013, 29(1): 15-21.

29. Doolittle WF. Is junk DNA bunk? A critique of ENCODE.

Proceedings of the National Academy of Sciences of the United States of America 2013, 110(14): 5294-5300.

30. Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, et al. The UCSC Genome Browser database: extensions and updates 2011. *Nucleic acids research* 2012, 40(Database issue): D918-923.

31. Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, Morgan TE, et al. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nature medicine* 2008, 14(7): 723-730.

32. Faghihi MA, Zhang M, Huang J, Modarresi F, Van der Brug MP, Nalls MA, et al. Evidence for natural antisense transcript-mediated inhibition of microRNA function. *Genome biology* 2010, 11(5): R56.

33. Fan M, Li X, Jiang W, Huang Y, Li J, Wang Z. A long non-coding RNA, PTCSC3, as a tumor suppressor and a target of miRNAs in thyroid cancer cells. *Experimental and therapeutic medicine* 2013, 5(4): 1143-1146.

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

35. Franco-Zorrilla JM, Valli A, Todesco M, Mateos I, Puga MI, Rubio-Somoza I, et al. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature genetics* 2007, 39(8): 1033-1037.

36. Franklin JL, Rankin CR, Levy S, Snoddy JR, Zhang B, Washington MK, et al. Malignant transformation of colonic epithelial cells by a colon-derived long noncoding RNA. *Biochemical and biophysical research communications* 2013, 440(1): 99-104.

37. Garg M. MicroRNAs, stem cells and cancer stem cells. *World journal of stem cells* 2012, 4(7): 62-70.

38. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, et al. What is a gene, post-ENCODE? History and updated definition. *Genome research* 2007, 17(6): 669-681.

39. Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, Brown JB, et al. Comparative analysis of the transcriptome across distant species. *Nature* 2014, 512(7515): 445-448.

40. Gingeras TR. Origin of phenotypes: genes and transcripts. *Genome research* 2007, 17(6): 682-690.

41. Gong C, Maquat LE. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature* 2011, 470(7333): 284-288.

42. Gonzalez-Porta M, Calvo M, Sammeth M, Guigo R. Estimation of alternative splicing variability in human populations. *Genome research* 2012, 22(3): 528-538.

43. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 2011, 29(7): 644-652.

44. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, et al. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* 2011, 27(18): 2518-2528.

45. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome biology and evolution* 2013, 5(3): 578-590.

46. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 2009, 458(7235): 223-227.

47. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature biotechnology* 2010, 28(5): 503-510.

48. Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. *Nature* 2012, 482(7385): 339-346.

49. Ha TY. The Role of MicroRNAs in Regulatory T Cells and in the Immune Response. *Immune network* 2011, 11(1): 11-41.

50. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* 2013, 8(8): 1494-1512.

51. Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, et al. Natural RNA circles function as efficient microRNA sponges. *Nature* 2013, 495(7441): 384-388.

52. Hansen TB, Wiklund ED, Bramsen JB, Villadsen SB, Statham AL, Clark SJ, et al. miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. *The EMBO journal* 2011, 30(21): 4414-4422.

53. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* 2012, 22(9): 1760-1774.

54. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics* 2007, 39(3): 311-318.

55. Heinzen EL, Ge D, Cronin KD, Maia JM, Shianna KV, Gabriel WN, et al. Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS biology* 2008, 6(12): e1.

56. Heward JA, Lindsay MA. Long non-coding RNAs in the regulation of the immune response. *Trends in immunology* 2014, 35(9): 408-419.

57. Hu G, Tang Q, Sharma S, Yu F, Escobar TM, Muljo SA, et al. Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. *Nature immunology* 2013, 14(11): 1190-1198.

58. Ilott NE, Ponting CP. Predicting long non-coding RNAs using RNA sequencing. *Methods* 2013, 63(1): 50-59.

59. Imamura K, Akimitsu N. Long Non-Coding RNAs Involved in Immune Responses. *Frontiers in immunology* 2014, 5: 573.

60. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 2011, 147(4): 789-802.

61. Jalali S, Kapoor S, Sivadas A, Bhartiya D, Scaria V. Computational approaches towards understanding human long non-coding RNA biology. *Bioinformatics* 2015.

62. Jeon Y, Lee JT. YY1 tethers Xist RNA to the inactive X nucleation center. *Cell* 2011, 146(1): 119-133.

63. Kaikkonen MU, Spann NJ, Heinz S, Romanoski CE, Allison KA, Stender JD, et al. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Molecular cell* 2013, 51(3): 310-325.

64. Kallen AN, Zhou XB, Xu J, Qiao C, Ma J, Yan L, et al. The imprinted H19 lncRNA antagonizes let-7 microRNAs. *Molecular cell* 2013, 52(1): 101-112.

65. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, et al. Spatio-temporal transcriptome of the human brain. *Nature* 2011, 478(7370): 483-489.

66. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 2007, 316(5830): 1484-1488.

67. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, et al. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome research* 2005, 15(7): 987-997.

68. Karreth FA, Tay Y, Perna D, Ala U, Tan SM, Rust AG, et al. In vivo identification of tumor- suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. *Cell* 2011, 147(2): 382-395.

69. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, et al. Antisense transcription in the mammalian transcriptome. *Science* 2005, 309(5740): 1564-1566.

70. Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, et al. Functional annotation of a full-length mouse cDNA collection. *Nature* 2001, 409(6821): 685-690.

71. Keniry A, Oxley D, Monnier P, Kyba M, Dandolo L, Smits G, et al. The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and Igf1r. *Nature cell biology* 2012, 14(7): 659-665.

72. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 2009, 106(28): 11667-11672.

73. Kino T, Hurt DE, Ichijo T, Nader N, Chrousos GP. Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Science signaling* 2010, 3(107): ra8.

74. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic acids research* 2007, 35(Web Server issue): W345-349.

75. Kretz M, Siprashvili Z, Chu C, Webster DE, Zehnder A, Qu K, et al. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* 2013, 493(7431): 231-235.

76. Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, et al. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS genetics* 2012, 8(7): e1002841.

77. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* 2009, 10(3): R25.

78. Latos PA, Pauler FM, Koerner MV, Senergin HB, Hudson QJ, Stocsits RR, et al. Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science* 2012, 338(6113): 1469-1472.

79. Lee JT, Bartolomei MS. X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell* 2013, 152(6): 1308-1323.

80. Legnini I, Morlando M, Mangiavacchi A, Fatica A, Bozzoni I. A feedforward regulatory loop between HuR and the long noncoding RNA linc-MD1 controls early phases of myogenesis. *Molecular cell* 2014, 53(3): 506-514.

81. Leucci E, Patella F, Waage J, Holmstrom K, Lindow M, Porse B, et al. microRNA-9 targets the long non-coding RNA MALAT1 for degradation in the nucleus. *Scientific reports* 2013, 3: 2535.

82. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009, 25(14): 1754-1760.

83. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 2009, 25(15): 1966-1967.

84. Li T, Wang S, Wu R, Zhou X, Zhu D, Zhang Y. Identification of long non-protein coding RNAs in chicken skeletal muscle using next generation sequencing. *Genomics* 2012, 99(5): 292-298.

85. Li W, Notani D, Ma Q, Tanasa B, Nunez E, Chen AY, et al. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* 2013, 498(7455): 516-520.

86. Li X, Wu Z, Fu X, Han W. Long Noncoding RNAs: Insights from Biological Features and Functions to Diseases. *Medicinal research reviews* 2013, 33(3): 517-553.

87. Li Z, Rana TM. Decoding the noncoding: prospective of lncRNA-mediated innate immune regulation. *RNA biology* 2014, 11(8): 979-985.

88. Lin M, Pedrosa E, Shah A, Hrabovsky A, Maqbool S, Zheng D, et al. RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. *PloS one* 2011, 6(9): e23356.

89. Lin MF, Carlson JW, Crosby MA, Matthews BB, Yu C, Park S, et al. Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. *Genome research* 2007, 17(12): 1823-1836.

90. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 2011, 27(13): i275-282.

91. Lindsay MA. microRNAs and the immune response. *Trends in immunology* 2008, 29(7): 343-351.

92. Liu Q, Huang J, Zhou N, Zhang Z, Zhang A, Lu Z, et al. LncRNA loc285194 is a p53-regulated tumor suppressor. *Nucleic acids research* 2013, 41(9): 4976-4987.

93. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research* 2011, 21(6): 936-939.

94. Makrythanasis P, Kapranov P, Bartoloni L, Reymond A, Deutsch

S, Guigo R, et al. Variation in novel exons (RACEfrags) of the MECP2 gene in Rett syndrome patients and controls. *Human mutation* 2009, 30(9): E866-879.

95. Marco-Sola S, Sammeth M, Guigo R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature methods* 2012, 9(12): 1185-1188.

96. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research* 2008, 18(9): 1509-1517.

97. Marques AC, Ponting CP. Intergenic lncRNAs and the evolution of gene expression. *Current opinion in genetics & development* 2014, 27: 48-53.

98. Martianov I, Ramadass A, Serra Barros A, Chow N, Akoulitchev A. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* 2007, 445(7128): 666-670.

99. Mattick JS. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays : news and reviews in molecular, cellular and developmental biology* 2003, 25(10): 930-939.

100. Mattick JS, Rinn JL. Discovery and annotation of long non-coding RNAs. *Nature structural & molecular biology* 2015, 22(1): 5-7.

101. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 2013, 495(7441): 333-338.

102. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. *Nature reviews Genetics* 2009, 10(3): 155-159.

103. Mercer TR, Mattick JS. Understanding the regulatory and transcriptional complexity of the genome through structure. *Genome research* 2013, 23(7): 1081-1088.

104. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 2007, 448(7153): 553-560.

105. Mitsuya K, Meguro M, Lee MP, Katoh M, Schulz TC, Kugoh H, et al. LIT1, an imprinted antisense RNA in the human KvLQT1 locus identified by screening for differentially expressed transcripts using monochromosomal hybrids. *Human molecular genetics* 1999, 8(7): 1209-1217.

106. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 2010, 464(7289): 773-777.

107. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* 2008, 5(7): 621-628.

108. Mukherjee N, Corcoran DL, Nusbaum JD, Reid DW, Georgiev S, Hafner M, et al. Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Molecular cell* 2011, 43(3): 327-339.

109. Mutz KO, Heilkenbrinker A, Lonne M, Walter JG, Stahl F. Transcriptome analysis using next-generation sequencing. *Current opinion in biotechnology* 2013, 24(1): 22-30.

110. Nagalakshmi U, Waern K, Snyder M. RNA-Seq: a method for comprehensive transcriptome analysis. *Current protocols in molecular biology* / edited by Frederick M Ausubel [et al] 2010, Chapter 4: Unit 4 11 11-13.

111. Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R, et al. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* 2008, 322(5908): 1717-1720.

112. Nam JW, Bartel DP. Long noncoding RNAs in C. elegans.

Genome research 2012, 22(12): 2529-2540.

113. Necsulea A, Kaessmann H. Evolutionary dynamics of coding and non-coding transcriptomes. *Nature reviews Genetics* 2014, 15(11): 734-748.

114. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 2014, 505(7485): 635-640.

115. Nesterova TB, Slobodyanyuk SY, Elisaphenko EA, Shevchenko AI, Johnston C, Pavlova ME, et al. Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. *Genome research* 2001, 11(5): 833-849.

116. Ng SY, Johnson R, Stanton LW. Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *The EMBO journal* 2012, 31(3): 522-533.

117. Niitsu N, Sasaki K, Umeda M. A high serum soluble Fas/APO-1 level is associated with a poor outcome of aggressive non-Hodgkin's lymphoma. *Leukemia* 1999, 13(9): 1434-1440.

118. On T, Xiong X, Pu S, Turinsky A, Gong Y, Emili A, et al. The evolutionary landscape of the chromatin modification machinery reveals lineage specific gains, expansions, and losses. *Proteins* 2010, 78(9): 2075-2089.

119. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, et al. Long noncoding RNAs with enhancer-like function in human cells. *Cell* 2010, 143(1): 46-58.

120. Pagani M, Rossetti G, Panzeri I, de Candia P, Bonnal RJ, Rossi RL, et al. Role of microRNAs and long-non-coding RNAs in CD4(+) T-cell differentiation. *Immunological reviews* 2013, 253(1): 82-96.

121. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics* 2008, 40(12): 1413-1415.

122. Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L, Komorowski J, et al. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Molecular cell* 2008, 32(2): 232-246.

123. Pang KC, Dinger ME, Mercer TR, Malquori L, Grimmond SM, Chen W, et al. Genome-wide identification of long noncoding RNAs in CD8+ T cells. *Journal of immunology* 2009, 182(12): 7738-7748.

124. Panzeri I, Rossetti G, Abrignani S, Pagani M. Long Intergenic Non-Coding RNAs: Novel Drivers of Human Lymphocyte Differentiation. *Frontiers in immunology* 2015, 6: 175.

125. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome research* 2012, 22(3): 577-591.

126. Pearson H. Genetics: what is a gene? *Nature* 2006, 441(7092): 398-401.

127. Pennisi E. Genomics. DNA study forces rethink of what it means to be a gene. *Science* 2007, 316(5831): 1556-1557.

128. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010, 464(7289): 768-772.

129. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 2010, 465(7301): 1033-1038.

130. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, et al. Forces shaping the fastest evolving regions in the human genome. *PLoS genetics* 2006, 2(10): e168.

131. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* 2005, 33(Database issue): D501-504.

132. Qu H, Fang X. A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project. *Genomics, proteomics & bioinformatics* 2013, 11(3): 135-141.

133. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. *Annual review of biochemistry* 2012, 81: 145-166.

134. Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, et al. The transcriptional activity of human Chromosome 22. *Genes & development* 2003, 17(4): 529-540.

135. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 2007, 129(7): 1311-1323.

136. Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* 2011, 27(17): 2325-2329.

137. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. *Nature methods* 2010, 7(11): 909-912.

138. Rogler LE, Kosmyna B, Moskowitz D, Bebawee R, Rahimzadeh J, Kutchko K, et al. Small RNAs derived from lncRNA RNase MRP have gene-silencing activity relevant to human cartilage-hair hypoplasia. *Human molecular genetics* 2014, 23(2): 368-382.

139. Roh TY, Cuddapah S, Cui K, Zhao K. The genomic landscape of histone modifications in human T cells. *Proceedings of the National*

Academy of Sciences of the United States of America 2006, 103(43): 15782-15787.

140. Rossi RL, Rossetti G, Wenandy L, Curti S, Ripamonti A, Bonnal RJ, et al. Distinct microRNA signatures in human lymphocyte subsets and enforcement of the naive state in CD4+ T cells by the microRNA miR-125b. *Nature immunology* 2011, 12(8): 796-803.

141. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 2011, 146(3): 353-358.

142. Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature reviews Genetics* 2007, 8(6): 424-436.

143. Schmitz KM, Mayer C, Postepska A, Grummt I. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes & development* 2010, 24(20): 2264-2269.

144. Schorderet P, Duboule D. Structural and functional differences in the long non-coding RNA hotair in mouse and human. *PLoS genetics* 2011, 7(5): e1002071.

145. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 2012, 28(8): 1086-1092.

146. Sehgal L, Mathur R, Braun FK, Wise JF, Berkova Z, Neelapu S, et al. FAS-antisense 1 lncRNA and production of soluble versus membrane Fas in B-cell lymphoma. *Leukemia* 2014, 28(12): 2376-2387.

147. Seok J, Warren HS, Cuenca AG, Mindrinos MN, Baker HV, Xu W, et al. Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of*

Sciences of the United States of America 2013, 110(9): 3507-3512.

148. Shendure J. The beginning of the end for microarrays? *Nature methods* 2008, 5(7): 585-587.

149. Sun K, Chen X, Jiang P, Song X, Wang H, Sun H. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC genomics* 2013, 14 Suppl 2: S7.

150. Sun L, Zhang Z, Bailey TL, Perkins AC, Tallack MR, Xu Z, et al. Prediction of novel long non-coding RNAs based on RNA-Seq data of mouse Klf1 knockout study. *BMC bioinformatics* 2012, 13: 331.

151. Sun S, Del Rosario BC, Szanto A, Ogawa Y, Jeon Y, Lee JT. Jpx RNA activates Xist by evicting CTCF. *Cell* 2013, 153(7): 1537-1551.

152. Takao K, Miyakawa T. Genomic responses in mouse models greatly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences of the United States of America* 2015, 112(4): 1167-1172.

153. Tay Y, Kats L, Salmena L, Weiss D, Tan SM, Ala U, et al. Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell* 2011, 147(2): 344-357.

154. Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome research* 2012, 22(9): 1616-1625.

155. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009, 25(9): 1105-1111.

156. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* 2010, 28(5): 511-515.

157. Tuomela S, Lahesmaa R. Early T helper cell programming of gene expression in human. *Seminars in immunology* 2013, 25(4): 282-290.

158. Turner M, Galloway A, Vigorito E. Noncoding RNA and its associated proteins as regulatory elements of the immune system. *Nature immunology* 2014, 15(6): 484-491.

159. Verma-Gaur J, Torkamani A, Schaffer L, Head SR, Schork NJ, Feeney AJ. Noncoding transcription within the Igh distal V(H) region at PAIR elements affects the 3D structure of the Igh locus in pro-B cells. *Proceedings of the National Academy of Sciences of the United States of America* 2012, 109(42): 17004-17009.

160. Wan Y, Kertesz M, Spitale RC, Segal E, Chang HY. Understanding the transcriptome through RNA structure. *Nature reviews Genetics* 2011, 12(9): 641-655.

161. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008, 456(7221): 470-476.

162. Wang GS, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature reviews Genetics* 2007, 8(10): 749-761.

163. Wang J, Gong C, Maquat LE. Control of myogenesis by rodent SINE-containing lncRNAs. *Genes & development* 2013, 27(7): 793-804.

164. Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 2011, 472(7341): 120-124.

165. Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic acids research* 2013, 41(6): e74.

166. Wang Y, Xu Z, Jiang J, Xu C, Kang J, Xiao L, et al. Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. *Developmental cell* 2013, 25(1): 69-80.

167. Williams GT, Mourtada-Maarabouni M, Farzaneh F. A critical role for non-coding RNA GAS5 in growth arrest and rapamycin inhibition in human T-lymphocytes. *Biochemical Society transactions* 2011, 39(2): 482-486.

168. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 2010, 26(7): 873-881.

169. Yoon JH, Abdelmohsen K, Gorospe M. Functional interactions among microRNAs and long noncoding RNAs. *Seminars in cell & developmental biology* 2014, 34: 9-14.

170. Yoon JH, Abdelmohsen K, Kim J, Yang X, Martindale JL, Tominaga-Yamanaka K, et al. Scaffold function of long non-coding RNA HOTAIR in protein ubiquitination. *Nature communications* 2013, 4: 2939.

171. Yoon JH, Abdelmohsen K, Srikantan S, Yang X, Martindale JL, De S, et al. LincRNA-p21 suppresses target mRNA translation. *Molecular cell* 2012, 47(4): 648-655.

172. Young RS, Marques AC, Tibbit C, Haerty W, Bassett AR, Liu JL, et al. Identification and properties of 1,119 candidate lincRNA loci in the Drosophila melanogaster genome. *Genome biology and evolution* 2012, 4(4): 427-442.

173. Zhang H, Nestor CE, Zhao S, Lentini A, Bohle B, Benson M, et al. Profiling of human CD4+ T-cell subsets identifies the TH2-specific noncoding RNA GATA3-AS1. *The Journal of allergy and clinical immunology* 2013, 132(4): 1005-1008.

174. Zhang Y, Zhang XO, Chen T, Xiang JF, Yin QF, Xing YH, et al. Circular intronic long noncoding RNAs. *Molecular cell* 2013, 51(6): 792-806.

175. Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* 2008, 322(5902): 750-756.

176. Zimmerman AL, Wu S. MicroRNAs, cancer and cancer stem cells. *Cancer letters* 2011, 300(1): 10-19.

LincRNAs landscape in human lymphocytes highlights regulation of T cell differentiation by linc-MAF-4

Valeria Ranzani[1,3], Grazisa Rossetti[1,3], Ilaria Panzeri[1,3], Alberto Arrigoni[1,3], Raoul JP Bonnal[1,3], Serena Curti[1], Paola Gruarin[1], Elena Provasi[1], Elisa Sugliano[1], Maurizio Marconi[2], Raffaele De Francesco[1], Jens Geginat[1], Beatrice Bodega[1], Sergio Abrignani[1,*] & Massimiliano Pagani[1,*].

[1]Istituto Nazionale Genetica Molecolare "Romeo ed Enrica Invernizzi", 20122 Milano, Italy.

[2] IRCCS Ca' Granda Ospedale Maggiore Policlinico, 20122 Milan, Italy

[3] These authors contributed equally to this work

* Correspondence: pagani@ingm.org, abrignani@ingm.org

Nature immunology 2015, **16**(3)**:** 318-325

Abstract

Long non-coding-RNAs are emerging as important regulators of cellular functions but little is known on their role in human immune system. Here we investigated long intergenic non-coding-RNAs (lincRNAs) in thirteen T and B lymphocyte subsets by RNA-seq analysis and *de-novo* transcriptome reconstruction. Over five hundred new lincRNAs were identified and lincRNAs signatures were described. Expression of linc-MAF-4, a chromatin associated T_H1 specific lincRNA, was found to anti-correlate with MAF, a T_H2 associated transcription factor. Linc-MAF-4 down-regulation skews T cell differentiation toward T_H2. We identified a long-distance interaction between *linc-MAF-4* and *MAF* genomic regions, where linc-MAF-4 associates with LSD1 and EZH2, suggesting linc-MAF-4 regulated *MAF* transcription by chromatin modifiers recruitment. Our results demonstrate a key role of lincRNAs in T lymphocyte differentiation.

Introduction

Lymphocytes enable us to fight and survive infections, but are also major drivers of immune-mediated diseases, such as allergy and autoimmunity. These different type of immune responses are mostly coordinated by distinct CD4$^+$ T cell subsets through signals delivered both by cytokines and by cell-to-cell contacts[1]. Development and differentiation programs of CD4$^+$ T lymphocytes subsets with distinct effector functions have been extensively studied in terms of signalling pathways and transcriptional networks, and a certain degree of functional plasticity between different subsets has been recently established[2]. Indeed, CD4$^+$ T cell subset flexibility in the expression of genes coding for cytokines and transcription factors allows the immune system to dynamically adapt to the many challenges it faces[3]. As CD4$^+$ T lymphocyte subsets are no longer considered stable and terminally differentiated cell lineages, the question arises as to how lymphocyte phenotype and functions can be modulated and whether these new findings offer new therapeutic opportunities.

Besides the well-established role of transcription factors as instructive signals for cell differentiation toward a given lineage, other cues, such as epigenetic modifications, can regulate maintenance of cellular states[4]. In this context non-coding RNAs (ncRNAs) are emerging as a new regulatory layer impacting on both the development and the functioning of the immune system[5, 6]. Among the several classes of ncRNAs that play a specific role in lymphocyte biology, microRNAs are the best-characterized[7, 8, 9, 10, 11, 12]. As to long intergenic non-coding RNAs (lincRNAs), although thousands of them have been identified in the mammalian genome by bioinformatics analyses of transcriptomic data[13, 14], their functional characterization is still largely incomplete. The functional studies performed so far have shown that lincRNAs contribute to the control of cell differentiation and to the maintenance of cell identity through different modes of action[15]. Nuclear lincRNAs act mainly through their association with chromatin-modifying complexes[16, 17, 18]. Whereas, cytoplasmic lincRNAs can modulate translational control[19] and transcripts stability[20] directly by base pairing with specific targets or indirectly as competing endogenous RNAs[21, 22, 23]. Few examples of functional lincRNAs have been recently described in the mouse immune system. A broad analysis performed by interrogating naïve and memory CD8$^+$ cells purified from mouse spleen with a custom array of lincRNAs reported the identification of 96 lymphoid-specific lincRNAs and suggested a role for lincRNAs in lymphocyte differentiation and activation[24]. The lincRNA NeST has been found to be downregulated during lymphocyte activation in a reciprocal manner to IFN-g and to control susceptibility to Theiler's virus and Salmonella infection in mice through epigenetic regulation of the IFN-g locus[25, 26]. More recently, mouse lincRNA-Cox2 has been reported to be induced downstream Toll-like receptor signalling and to mediate the activation and repression of distinct sets of immune target genes involved in inflammatory responses[27]. Another study on mouse thymocytes and mature peripheral T cells allowed the identification of lincRNAs with specific cell expression pattern during T cell differentiation and of a $CD4^+$ T_H2 specific lincRNA - LincR-Ccr2-5'AS - involved in the regulation of $CD4^+$ T_H2 lymphocytes migration[28]. Although these studies highlight the relevance of lincRNAs in regulating immune responses, a thorough analysis of their expression profile and functional role in the human immune system is still lacking.

The present study is based on a RNA-seq analysis of thirteen highly purified primary human lymphocytes subsets. We performed a *de novo* transcriptome reconstruction, and discovered over five hundred new long intergenic non-coding RNAs (lincRNAs). We identified several lymphocyte subset-specific lincRNAs signatures, and found that linc-MAF-4, a chromatin associated $CD4^+$ T_H1 specific lincRNA, correlates inversely with the transcription factor MAF and that its down-regulation skews $CD4^+$ T cell differentiation toward T_H2 phenotype.

We provide the first comprehensive inventory of human lymphocytes lincRNAs and demonstrate that lincRNAs can be key to lymphocyte differentiation. This resource will likely help a better definition of lincRNAs role in lymphocytes differentiation, plasticity and effector functions.

Results

LincRNAs identify human lymphocyte subsets better than protein coding genes

To assess lincRNA expression in human primary lymphocytes, RNA was extracted from thirteen lymphocyte cell subsets (Table 1) purified from peripheral blood mononuclear cells (PBMCs) of five healthy donors[12]. The polyadenylated RNA fraction was then analysed by paired-end RNA sequencing obtaining about 1.7 billion mapped reads. In order to enrich for transcripts deriving from "bona fide" active genes we applied an expression threshold ("0.21" FPKM) defined through the integration of RNAseq and chromatin state ENCODE project data[29]. We found a total of 31,902 expressed genes (including both protein coding and non coding genes) in the 13 subsets (Table 1 and Supplementary Fig. 1a), of which 4,201 were lincRNAs annotated in public resources[13, 30] (Fig. 1a). In order to identify novel lincRNAs expressed in primary human lymphocytes, we used three *de novo* transcriptome reconstruction strategies that are based on the combination of two different sequence mappers, TopHat and Star[31, 32], with two different tools for *de nov*o transcripts assembly, Cufflinks and Trinity[33, 34]. LincRNAs were identified within the newly described transcripts exploiting the following process: *i*) selection of transcripts longer than 200 nucleotides and multiexonic, which did not overlap with protein coding genes (thus counting out unreliable single-exon fragments assembled from RNA-seq); *ii*) exclusion of transcripts that contain a conserved protein-coding region and transcripts with ORFs that contain protein domains catalogued in Pfam protein family database[35]; *iii*) exploitation of PhyloCSF, a comparative genomics method that assesses multispecies nucleotide sequence alignment based on a formal statistical comparison of phylogenetic codon models[36], which efficiently identifies non-coding RNAs as demonstrated by ribosome profiling experiments[37]. Finally we defined a stringent *de novo* lincRNA set including those genes for which at least one lincRNA isoform was reconstructed by two assemblers out of three. Through this conservatively multi-layered analysis we identified 563 novel lincRNAs genes, increasing by 11.8% the number of lincRNAs expressed in human lymphocytes. The different classes of RNAs are evenly distributed among different lymphocytes subsets (Supplementary Fig. 1b) and the ratio of already annotated and newly identified lincRNAs is similar across different chromosomes

(Supplementary Fig. 1c) and across various lymphocyte subsets (Supplementary Fig. 1d). As previously observed in different cell types[13, 33], also in human lymphocytes lincRNAs are generally expressed at lower levels than protein coding genes (Supplementary Fig. 1e). However, when transcripts were divided based on their expression in cell-specific and non specific (Supplementary Fig. 1f), we found that cell specific lincRNAs and cell specific protein coding genes, display similar expression levels (Supplementary Fig. 1e-g).

Lymphocytes subsets display very different migratory abilities and effector functions, yet they are very closely related from the differentiation point of view. As lincRNAs are generally more tissue specific than protein coding genes[13, 38], we assessed the lymphocyte cell-subset specificity of lincRNAs. We therefore classified genes according to their expression profiles by unsupervised K-means clustering and found that lincRNAs are defined by 15 clusters and protein coding genes by 24 clusters (Fig. 1b and Supplementary Fig. 1h). Remarkably, the percentage of genes assigned to the clusters specific for the different lymphocyte subsets is higher for lincRNAs (71%) than for protein coding genes (34%) (Fig. 1c). This superiority stands out even when lincRNAs are compared with membrane receptor coding genes (40%) (Fig. 1d), which are generally considered the most accurate markers of different lymphocyte subsets. Similar results were obtained also using the heuristic expression threshold of FPKM>1 (Supplementary Fig. 1i).

Altogether, based on RNA-seq analyses of highly purified primary T and B lymphocyte subsets, we provide a comprehensive landscape of lincRNAs expression in human lymphocytes. Exploiting a *de novo* transcriptome reconstruction we discovered 563 new lincRNAs, and found that lincRNAs are very effective in marking lymphocyte cell identity.

Identification of lincRNA expression signatures of human lymphocyte subsets

Next, we interrogated our dataset for the presence of lincRNAs signatures in the different lymphocyte subsets. We therefore looked for lincRNAs differentially expressed ($p<0.05$; non-parametric Kruskal-Wallis test) that had more than 2.5 fold expression difference in a given cell subset compared to all the other subsets and that were expressed in at least 3 out of 5 individuals and found 172 lincRNAs that met these criteria (Fig. 2a and Supplementary Fig. 2b-m). We integrated the human transcriptome database with our newly identified transcripts and thus created a new reference to assess more thoroughly expression of new transcripts, in other human tissues. Looking at lincRNAs signatures in a panel of sixteen human tissues (Human BodyMap 2.0 project) we found that lymphocytes signature lincRNAs are not only very poorly expressed in non-lymphoid tissues (Fig. 2a), but also that most signature lincRNAs are not detectable even in lymphoid tissues. These findings underscore the importance of assessing expression of lincRNAs (as well as of any highly cell-specific transcripts) in purified primary cells rather than in total tissues where a given cell-subset-specific transcript is diluted by the transcripts of all the other cell types of the tissue.

It is important to note that, the newly identified lincRNAs defined as signatures are more expressed (Fig. 2c) and more cell-specific (Supplementary Fig. 2b-m) than the already annotated lincRNAs defined as signatures. The representative data in Fig. 2b refer to the CD4$^+$ T$_H$1 cell subset; similar results were obtained for all the other subsets (Supplementary Fig. 2b-m).

Finally, to confirm and extend our signature data, we assessed the expression of CD4$^+$ T$_H$1 lincRNAs by RT-qPCR in a new set of independent samples of primary human CD4$^+$ naïve, T$_{reg}$ and T$_H$1 cells, as well as in naïve CD4$^+$ T cells that were activated *in vitro* and induced to differentiate toward T$_H$1 or T$_H$2 cells. Specific subset expression was confirmed for 90% of the CD4$^+$ T$_H$1 signature lincRNAs (Fig. 2d). Moreover, 90% of CD4$^+$ T$_H$1 signature lincRNAs that are expressed in resting CD4$^+$ T$_H$1 cells purified *ex vivo*, are highly expressed also in naïve CD4$^+$ T cells differentiated under T$_H$1 polarizing conditions *in vitro*, whereas they are poorly expressed in naïve CD4$^+$ T cells that are differentiated towards T$_H$2 *in vitro* (Fig. 2e). As a corollary to these findings, we observed by RNA-seq that CD4$^+$ naïve signature lincRNAs are mostly down-regulated during differentiation towards T$_H$0 cells *in vitro*, when T$_H$1, T$_H$2 and T$_H$17 signature lincRNAs are mostly up-regulated (Supplementary Fig. 2a).

Taken together our data demonstrate that lincRNAs provide excellent signatures of human lymphocyte subsets, and suggest that human CD4$^+$ T lymphocytes acquire most of their memory specific lincRNAs signatures during their activation-driven differentiation from naïve to memory cells.

Linc-MAF-4 downregulation skews CD4$^+$ T cell differentiation towards T$_H$2

As lincRNAs have been reported to influence the expression of neighbouring genes[25, 26, 28, 39], we asked whether protein coding genes proximal to lymphocytes signature lincRNAs were involved in key cell-functions. To this purpose we used the FatiGO tool from the Babelomics suite for functional enrichment analysis[40] and found that protein coding genes neighbouring to signature lincRNAs are enriched for Gene Ontology terms strongly correlated with lymphocyte T cell activation (Fig. 3a), pointing to a possible role of signature lincRNAs in important lymphocyte functions. In order to obtain proof of concept of this hypothesis, we chose to characterize in depth linc-MAF-4 (also referred to as linc-MAF-2 in LNCipedia database http://www.lncipedia.org[41]), a T$_H$1 signature lincRNA, localized 139.5 Kb upstream of the *MAF* gene. *MAF* encodes a transcription factor involved in T$_H$2 differentiation[42], which is also required for the efficient development of T$_H$17 cells[43] and controls IL4 transcription in CD4$^+$ T follicular helper cells[44]. Our sequencing data showed that high expression of linc-MAF-4 correlates with low levels of *MAF* transcript in CD4$^+$ T$_H$1 cells, conversely T$_H$2 cells have low expression levels of linc-MAF-4 and high levels of MAF transcript. The anti-correlation of expression between lincRNAs and their neighbouring genes is not a common feature of all lincRNAs ([13, 16]), and it is probably restricted to a limited number of cis-acting lincRNAs. This observation is confirmed also in our dataset (data not shown). Moreover, no correlation is observed between the expression linc-MAF-4 and its proximal upstream protein coding genes: CDYL2 and DYNLRB2 (Supplementary Fig. 3a).

The same inverse relation between linc-MAF-4 and MAF is observed when naïve CD4$^+$ T cells are differentiated *in vitro* towards T$_H$1 or T$_H$2 cells. In details, Fig. 3b shows that in T lymphocytes differentiating towards T$_H$1 cells, MAF transcript increases up to day 3 and then drops. Conversely, linc-MAF-4 is poorly expressed for the first three days but then increases progressively. In CD4$^+$ T lymphocytes differentiating towards T$_H$2 cells, we found the opposite situation, both MAF transcript and protein levels increase constantly up to day 8 while Iinc-MAF4 remains constantly low (Fig. 3b and Supplementary Fig. 3c), similarly to what observed in CD4$^+$ T lymphocytes differentiating towards T$_H$17 cells (Supplementary Fig. 3d).

We further characterized *MAF* transcriptional regulation by looking at H3K4 tri-methylation (H3K4me3) level and RNA polymerase II occupancy at *MAF* promoter region in T_H1 and T_H2 cells. Consistent with a higher active transcription of *MAF* in $CD4^+$ T_H2 cells, we found that H3K4me3 levels in T_H2 cells are greater than in T_H1 cells and that RNA polymerase II binding at *MAF* promoter is higher in T_H2 than in T_H1 cells (Fig. 3c). Intriguingly, linc-MAF-4 knock-down in activated $CD4^+$ naïve T cells leads to MAF increased expression (Fig. 3e and Supplementary Fig. 3e). All the above results indicate that modulation of *MAF* transcription in T cells depends on tuning of its promoter setting, and suggest a direct involvement of linc-MAF-4 in the regulation of *MAF* transcriptional levels.

We then assessed the overall impact of linc-MAF-4 knock-down on $CD4^+$ T cell differentiation by performing transcriptome profiling and Gene Set Enrichment Analysis (GSEA). We defined as reference Gene-Sets the genes upregulated in $CD4^+$ naïve T cells differentiated *in vitro* towards T_H1 or T_H2 types (Supplementary Table 1). We found that the $CD4^+$ T_H2 gene set is enriched for genes that are overexpressed in linc-MAF-4 knock-down cells, whereas the $CD4^+$ T_H1 gene set is depleted of these same genes (Fig. 3f). Concordant with these findings, the expression of *GATA3* and *IL4,* two genes characteristic of T_H2 cells, is increased after linc-MAF-4 knock-down (Fig. 3g and Supplementary Fig.3e).

Taken together these results demonstrate that linc-MAF-4 down regulation contributes to the skewing of $CD4^+$ T cells differentiation towards T_H2.

Epigenetic regulation of *MAF* transcription by linc-MAF-4

Since *linc-MAF-4* gene maps in relative proximity (139.5 Kb) to *MAF* gene we asked whether linc-MAF-4 can down-regulate *MAF* transcription, and, we investigated whether their genomic regions could physically interact. Chromosome conformation capture (3C) analysis was exploited to determine relative crosslinking frequencies among regions of interest. We tested the conformation of the *linc-MAF-4* - *MAF* genomic region in differentiated $CD4^+$ T_H1 cells. A common reverse primer mapping within the *MAF* promoter region, was used in combination with a set of primers spanning the locus, and interactions were analysed by PCR. Specific interactions between *MAF* promoter and 5' and 3' end regions of *linc-MAF-4* were detected (Fig.

4a,b and Supplementary Fig. 4a), indicating the existence of an *in cis* chromatin looping conformation that brings *linc-MAF-4* in close proximity to *MAF* promoter. Interestingly, the subcellular fractionation of *in vitro* differentiated CD4$^+$ T$_H$1 lymphocytes revealed a strong enrichment of linc-MAF-4 in the chromatin fraction (Fig. 4c). Because other chromatin-associated lincRNAs regulate neighbouring genes by recruiting specific chromatin remodellers, we tested in RNA immunoprecipitation (RIP) assays the interaction of linc-MAF-4 with different chromatin modifiers, including activators and repressors (data not shown), and found a specific enrichment of linc-MAF-4 in the immunoprecipitates of two repressors, EZH2 and LSD1 (Fig. 4d and Supplementary Fig. 4b). In agreement with these findings, we found that linc-MAF-4 knock-down in activated CD4$^+$ naïve T cells reduces both EZH2 and LSD1 levels and correlates with the reduction of EZH2 enzymatic activity at *MAF* promoter as demonstrated by the H3K27me3 reduction at this locus (Fig. 4e). Remarkably, H3K27me3 levels were reduced neither at *MYOD1* promoter region (a known target of EZH2) nor at a region within the chromatin loop between *linc-MAF-4* and *MAF* marked by H3K27me3 (Supplementary Fig. 4c).

Altogether, these results demonstrate that there is a long distance interaction between *linc-MAF-4* and *MAF* genomic regions, through which linc-MAF-4 could act as a scaffold to recruit both EZH2 and LSD1 and modulate the enzymatic activity of EZH2 on *MAF* promoter, thus regulating its transcription (Fig. 4f).

Discussion

Mammalian genomes encode more long non-coding RNAs than previously thought[16, 45] and the number of lincRNAs playing a role in cellular processes steadily grows. As there are relatively few examples of functional long non-coding RNAs in the immune system[24, 25, 26, 27, 28], with the present study we depict a comprehensive landscape of lincRNAs expression in thirteen subsets of human primary lymphocytes. Moreover, we identified a lincRNA (linc-MAF-4) that appear to play a key role in CD4$^+$ T helper cell differentiation.

LincRNAs have been reported to have high tissue specificity[13] and our study of lincRNAs expression in highly pure primary human lymphocyte provides an added value because it allows the identification of lincRNAs whose expression is restricted to a given lymphocyte cell subset. Interestingly, we found that lincRNAs define the cellular identity better than protein coding genes, even than surface receptor coding genes that are generally considered the most precise markers of lymphocytes subsets. Due to their specificity of expression, human lymphocytes lincRNAs that are not yet annotated in public resources would have not been identified without performing *de novo* transcriptome reconstruction. Indeed by exploiting three different *de novo* strategies we identified 563 novel lincRNAs and increased by 11.8% the number of lincRNAs expressed in human lymphocytes. As our conservative analysis was limited to thirteen cellular subsets, one may wonder how many novel lincRNAs could be identified by transcriptome analysis of all of the several hundreds human cell types.

We Compared our data with previous analyses of lincRNAs expression in mouse immune system[28] exploiting the LNCipedia database (http://www.lncipedia.org [41]) and we found that 51% of the human lincRNA signatures are conserved in mouse, that is similar to the overall conservation between human and mouse lincRNAs (60%). However further studies will be necessary to asses that also their function is conserved.

Based on our findings, signature lincRNAs might be exploited to discriminate and differentiate at the molecular level those cell subsets that cannot be distinguished easily based on cell surface markers because of their cellular heterogeneity, such as CD4$^+$ regulatory T cells (Treg cells). Furthermore, most lincRNA signatures defined for each of the thirteen lymphocytes subsets are not detected in human lymphoid tissues that include all the lymphocyte subsets we analyzed. Indeed, to get the best out of the enormous molecular resolution achievable with Next-Generation-Sequencing one should perform transcriptomic studies on single cells, or at least on functionally homogenous cell subsets. As lincRNAs expression in a tissue is averaged across all the cell types composing that tissue, a transcriptome analysis on unseparated tissue-derived cells will result in an underestimation both of the expression of a cell specific lincRNA and of its functional relevance.

The lincRNAs role in differentiation has been described in different cell types[17, 20, 23, 46, 47]. In the mouse immune system it has been found that lincRNAs expression changes during naïve to memory CD8$^+$ T cell differentiation[24] and during naïve CD4$^+$ T cells differentiation into distinct helper T cell lineages[28]. We show in human primary lymphocytes that activation induced differentiation of CD4$^+$ naïve T cells is associated with increased expression of lincRNAs belonging to the CD4$^+$ T$_H$1 signature suggesting that upregulation of T$_H$1 lincRNAs is part of the cell differentiation transcriptional program. Indeed, linc-MAF-4, one of the T$_H$1 signature lincRNA, is poorly expressed in T$_H$2 cells and its experimental downregulation skews differentiating T helper cells toward a T$_H$2 transcription profile. We have found that linc-MAF-4 regulates transcription exploiting a chromatin loop that brings its genomic region close to the promoter of *MAF* gene. We propose that the chromatin organization of this region allows linc-MAF-4 transcript to recruit both EZH2 and LSD1 and modulate the enzymatic activity of EZH2 negatively regulating *MAF* transcription with a mechanism of action similar to that shown for the lincRNAs HOTAIR[48] and MEG3 [49]. We therefore provide a mechanistic proof of concept that lincRNAs can be important regulators of CD4$^+$ T-cell differentiation. Given the number of specific lincRNAs expressed in the different lymphocytes subsets, it can be postulated that many other lincRNAs might contribute to cell differentiation and to the definition of cell identity in human lymphocytes.

These findings and the high cell specificity of lincRNAs suggest lincRNAs as novel and highly specific molecular targets for the development of new therapies for diseases (e.g. autoimmunity, allergy, and cancer) in which altered CD4$^+$ T-cell functions play a pathogenic role.

Online Methods

Purification of primary immunological cell subsets

Buffy-coated blood of healthy donors was obtained from the Ospedale Maggiore in Milan and peripheral blood mononuclear cells were isolated by Ficoll-hypaque density gradient centrifugation. The ethical committee of Istituto di Ricovero e Cura a Carattere Scientifico Policlinico Ospedale Maggiore approved the use of PBMCs from healthy donors for research purposes, and informed consent was obtained from subjects. Human blood primary lymphocyte subsets were purified >95% by cell sorting using different combinations of surface markers (Table 1). For *in vitro* differentiation experiments resting naïve $CD4^+$T cells were purified >95% by negative selection with magnetic beads with the isolation kit for human $CD4^+$ Naïve T cells of Miltenyi and stimulated with Dynabeads Human T-Activator CD3/CD28 (Life Technologies). IL-2 was added at 20 IU/ml (Novartis). T_H1 polarization was initiated with 10 ng/ml IL12 (R&D Systems) and T_H2 neutralizing antibody anti-IL4 (2 mg/ml). T_H2 polarization was induced by activation with Phytohaemagglutinin, PHA (4mg/mL) in the presence of IL-4 (R&D Systems) (10 ng/ml), and neutralizing antibodies to IFN-γ (2 mg/ml) and anti-IL12 (2 mg/ml). For GATA-3 and c-Maf intracellular staining, cells were harvested and then fixed for 30 min in Fixation/permeabilisation Buffer (Ebioscience) at 4°C. Cells were stained with antibodies anti-GATA-3 (BD bioscience) and anti-c-Maf (Ebioscience) in washing buffer for 30 min at 4°C. Cells were then washed two times, resuspended in FACS washing buffer and analysed by flow cytometry.

RNA isolation and RNA sequencing

Total RNA was isolated using mirVana Isolation Kit. Libraries for Illumina sequencing were constructed from 100 ng of total RNA with the Illumina TruSeq RNA Sample Preparation Kit v2 (Set A). The generated libraries were loaded on to the cBot (Illumina) for clustering on a HiSeq Flow Cell v3. The flow cell was then sequenced using a HiScanSQ (Illumina). A paired-end (2×101) run was performed using the SBS Kit v3 (Illumina). Real-time analysis and base calling was performed using the HiSeq Control Software Version 1.5 (Illumina).

RNA-seq and publicly available datasets

RNA-seq data representative of 13 lymphocyte populations were collected for transcriptome reconstruction. Five biological replicates were analyzed for all populations except for CD8$^+$ T$_{CM}$ and B CD5$^+$ (four samples). The whole dataset was aligned to GRCh37 (Genome Reference Consortium Human Build 37) with TopHat v.1.4.1[32] for a total of over 1.7 billions mapped paired-end reads (30 million reads per sample on average). These data were also mapped with the aligner STAR v.2.2.0[31]. RNA-seq datasets of 16 human tissues belonging to the Illumina Human BodyMap 2.0 project (ArrayExpress accession no. E-MTAB-513) were mapped following the same criteria.

Reference annotation

An initial custom reference annotation of unique, non-redundant transcripts was built by integrating the Ensembl database (version 67 from May 2012) with the lincRNAs identified by Cabili et al. 2011 using Cuffcompare v.2.1.1[33]. The annotated human lincRNAs were extracted from Ensembl using BioMart v.67 and subset by gene biotype 'lincRNA' (5,804 genes). Other classes of genes were integrated in the annotation: the list of protein coding genes (21,976 genes), the receptors genes collection defined in BioMart under GO term GO:000487 (2,043 genes with receptor activity function) and the class of genes involved in metabolic processes corresponding to GO term GO:0008152 (7,756 genes). Hence, the complete reference annotation consisted of 195,392 transcripts that referred to 62,641 genes, 11,170 of which are non-redundant lincRNA genes.

De novo genome-based transcripts reconstruction

A comprehensive catalogue of lincRNAs specifically expressed in human lymphocyte subsets was generated using a *de novo* genome-based transcripts reconstruction procedure with three different approaches. Two aligners were used: TopHat v.1.4.1 and STAR v. 2.2.0. The *de novo* transcriptome assembly was performed on the aligned sequences (samples of the same population were concatenated into one "population alignment") generated by STAR and TopHat using Cufflinks v. 2.1.1 with reference annotation to guide the assembly (-g option) coupled with multi-read (-u option) and fragment bias correction (-b option) to improve the accuracy of transcripts abundance estimates. With this method, about 30,000-50,000 new transcripts were identified in each lymphocyte population. The third approach employed the genome-guided Trinity software (http://pasa.sourceforge.net/#A_ComprehensiveTranscriptome), which generates novel transcripts performing a local assembly on previously mapped reads from specific location. The Trinity[50] default aligner was substituted with STAR. Each candidate transcript was then processed using the PASA pipeline, which reconstructs the complete transcript and gene structures, resolving incongruences derived from transcript misalignments and alternatively splices events, refining the reference annotation when there are enough evidences and proposing new transcripts and genes in case no previous annotation can explain the new data.

Novel lincRNA genes identification

Annotated transcripts and new isoforms of known genes were discarded, retaining only novel genes and their isoforms located in intergenic position. In order to filter out artifactual transcripts due to transcriptional noise or low polymerase fidelity, only multi-exonic transcripts longer than 200 bases were retained. Then, the HMMER3 algorithm[35] was run for each transcript in order to identify occurrences of any protein family domain documented in the Pfam database (release 26; used both PfamA and PfamB). All six possible frames were considered for the analysis, and the matching transcripts were excluded from the final catalogue.

The coding potential for all the remaining transcripts was then evaluated using PhyloCSF (phylogenetic codon substitution frequency)[36] (PhyloCSF was run on a multiple sequence alignment of 29 mammalian genomes (in MAF format) (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/) to obtain the best scoring ORF greater than 29 aminoacids across all three reading frames. To efficiently access the multialignment files (MAF) the bio-maf (https://github.com/csw/bioruby-maf) Ruby biogem[51] was employed. This library provides indexed and sequential access to MAF data, as well as performing fast manipulations on it and writing modified MAF files. Transcripts with at least one open reading frame with a PhyloCSF score greater than 100 were excluded from the final catalogue. The PhyloCSF score threshold of 100 was determined by Cabili et al. 2011 to optimize specificity and sensitivity when classifying coding and non coding transcripts annotated in RefSeq (RefSeq coding and RefSeq lincRNAs). PhyloCSF score =100 corresponds to a false negative rate of 6% for coding genes (i.e., 6% of coding genes are classified as non-coding) and a false positive rate of ~10% (i.e., 9.5% of noncoding transcripts are classified as coding).

De novo data integration

Duplicates among the transcripts identified with the same *de novo* method were resolved using Cuffcompare v2.1.1. In the same way, the resulting three datasets were further merged to generate a non-redundant atlas of lincRNAs in human lymphocytes and only genes identified by at least 2 out of 3 software were considered. A unique name was given to each newly identified lincRNA gene composed by the prefix "linc-" followed by the Ensembl gene name of the nearest protein coding gene (irrespective of the strand). The additional designation "up" or "down" defines the location of the lincRNA with respect to the sense of transcription of the nearest protein coding gene. In addition, either "sense" or "antisense" was added to describe the concordance of transcription between the lincRNA and its nearest coding gene. A numerical counter only of newly identified lincRNAs related to the same protein coding gene is added as suffix (such as 'linc-geneX-(up|down)-(sense|antisense)_#n'). This final non-redundant catalogue of newly identified lincRNAs includes 4,666 new transcripts referring to 3,005 new genes.

LincRNA signatures definition

A differential expression analysis among the thirteen cell subsets profiled was performed using Cuffdiff v.2.1.1. This analysis was run using --multi-read-correction (-u option) and upper quartile normalization (--library-norm-method quartile) to improve robustness of differential expression calls for less abundant genes and transcripts. Only genes expressed over 0.21 FPKM [29]were considered in the downstream analysis to filter out genes that are merely by-products of leaky gene expression, sequencing errors, and/or off-target read mapping. After adding a pseudo-count of 1 to the raw FPKM (fragments per kilobases of exons per million fragments mapped) for each gene, applying $\log_2$ transformation and Z-score normalization, K-means clustering with Euclidean metric was performed on lincRNAs expression values using

MultiExperiment Viewer v.4.6 tool. The same procedure was then applied to the expression values of protein coding, metabolic and receptors genes. The Silhouette function[52] was used to select an appropriate K (number of clusters). A K ranging from 13 to 60 was tested, and the value associated with the highest Silhouette score for each class of genes was selected. The number of clusters that maximizes the Silhouette score is 15 for lincRNA (Supplementary Figure 1h), 24 for protein coding genes and 23 and 36 for receptors and metabolic genes respectively. The centroid-expression profile of each cluster was then evaluated in order to associate each cluster to a single cellular population (Figure 1).

In order to select specifically expressed lincRNA genes, K-means results were subsequently intersected with the JS score, a cell-specificity measure based on Jensen–Shannon divergence and only the genes assigned to the same cellular population by both techniques were retained for further analysis. The estimation procedure for the JS score was adapted by building a reference model composed of 13 cell subsets. For the selected lincRNAs, the intrapopulation consistency among different samples was subsequently evaluated to minimize the biological variability: only genes expressed in at least 3/5 (or 3/4 replicates for $CD8^{+}{}_{CM}$ and $CD5^{+}$ B) of the profiled samples whose maximal expression value was >2.5 fold compared to all other lymphocyte subsets were considered. Finally, non-parametric Kruskal-Wallis test was applied to select only lincRNA genes with a significant difference across the medians of the different lymphocyte populations: a p-value lower than 0.05 was considered and the lincRNA genes that meet these selection criteria were selected as signature genes.

Gene Ontology Enrichment Analysis

A Gene Ontology (GO) enrichment analysis was performed for biological process terms associated with protein coding genes that are proximal to lincRNA signatures at genomic level. For each lincRNA signature, the proximal protein- coding gene was selected regardless of the sense of transcription. FatiGO tool of Babelomics suite (version 4.3.0) was used to identify the enriched GO terms of the 158 protein coding genes (input list). All protein coding genes that are expressed in lymphocyte subsets (19,246 genes) (except the genes proximal to a lincRNA signature gene [input list]) defined the background list. Only GO terms with adjusted pvalue lower than 0.01 were considered (10 GO terms). Moreover, we performed a gene ontology semantic similarity analysis on the 51 GO terms with adjusted pvalue lower than 0.1 resulting from previous analysis using G-SESAME tool. This analysis provides as a result a symmetric matrix where each value represents a similarity score between GO term pairs. Then, we carried out a hierarchical clustering based on semantic similarity matrix to group together all GO terms with common GO parent.

Naïve CD4$^+$ T cells siRNA transfection

Activated CD4$^+$ naïve T Cells, were transfected with 300 nM FITC-labelled- linc-MAF-4 siRNA or FITC-labelled-AllStars negative control (Qiagen) with Lipofectamine 2000 (Life Technologies) according to the manufacturer protocol. FITC positive cells were sorted and lysated 72 hours post transfection. See Supplementary Table 2 for siRNAs sequences.

Gene Expression Analysis

Gene expression analysis of transfected activated CD4$^+$ naive cells was performed with Illumina Direct Hybridization Assays according to the standard protocol (Illumina). Total RNA was isolated, quality controlled and quantified as described above; for each sample 500 ng of total RNA were reverse transcribed according to the Illumina TotalPrep RNA Amplification kit (AMIL1791 - LifeTechnologies) and cRNA was generated by *in vitro* transcription (14 hours). Hybridization was performed according to the standard Illumina protocol on Illumina HumanHT-12 v4 Expression BeadChip arrays (BD-103-0204 - Illumina). Scanning was performed on an Illumina HiScanSQ System and data were processed with Genome Studio; arrays were quantile normalized, with no background subtraction, and average signals were calculated on gene-level data for genes whose detection p-value was lower than 0.001 in at least one of the cohorts considered.

GSEA (Gene Set Enrichment Analysis)

GSEA is a statistical methodology used to evaluate whether a given gene set is significantly enriched in a list of gene markers ranked by their correlation with a phenotype of interest. In order to evaluate this degree of 'enrichment', the software calculates an enrichment score (ES) by moving down the ranked list, i.e., increasing the value of the sum if the marker is included in the gene set and decreasing this value if the marker is not in the gene set. The value of the increase depends on the gene-phenotype correlation. GSEA was performed comparing gene expression data obtained from activated CD4$^+$ naïve T cells transfected with linc-MAF-4 siRNAs vs. control siRNAs. The experimentally generated dataset from the *in vitro* differentiated cells (in T_H1 or T_H2 polarizing conditions respectively) derived from CD4+ naïve T cells of the same donors where linc-MAF-4 down-regulation was performed, were used to construct reference gene sets for T_H1 and a T_H2 cells. RNA for gene expression analysis of T_H1 and T_H2 differentiating cells was collected 72 hours after activation (i.e., the same time-point of RNA collection in the linc-MAF-4 downregulation experiments) but a fraction of cells was further differentiated up to day 8 to assess IFN-g and IL-13 production by T_H1 and T_H2 cells. The T_H1 and T_H2 datasets were ranked as $\log_2$ ratios of the expression values for each gene in the two conditions (T_H1/T_H2), and the most upregulated/downregulated genes (having log2 ratios ranging from |3| to |0.6|) were assigned to the T_H1 and T_H2 reference sets respectively.

Genes from the T_H1 gene list which were downregulated in a T_H1 vs. control-siRNA comparison and genes from the T_H2 gene list which were downregulated in a T_H2 vs. control-siRNA comparison were filtered out, obtaining a T_H1-specific gene set (74 genes) and a T_H2-specific gene set (141 genes) (Supplementary Table 1). GSEA was then performed on the linc-MAF-4 specific siRNA vs. control siRNA dataset. The metric used for the analysis is the $\log_2$ Ratio of Classes, with 1,000 gene set permutations for significance testing.

RT-qPCR Analysis

For reverse transcription, equal amounts of DNA-free RNA (500 ng) were reverse-transcribed with SuperScript III (LifeTechnologies) following the suggested conditions. Diluted cDNA was then used as input for RT-qPCR to assess MAF (Hs00193519_m1), IL4 (Hs00174122_m1), GATA3 (Hs01651755_m1), TBX21 (Hs00203436_m1), RORC (Hs01076119_m1), IL17 (Hs00174383_m1), Linc00339 (Hs04331223_m1), Malat1 (Hs01910177_s1), RNU2.1 (Hs03023892_g1) and GAPDH (Hs02758991_g1) gene expression levels with Inventoried TaqMan Gene Expression assays (LifeTechnologies) were used. For assessment of linc-MAF-4 and validation of CD4$^+$ T_H1 signature lincRNAs specific primers were designed and 2.5 mg of CD4$^+$ T$_H$1, T$_{reg}$ or naive cells RNA were used for reverse transcription with SuperScript III (LifeTechnologies). RT-qPCR was performed on diluted cDNA with PowerSyberGreen (LifeTechnologies) and specificity of the amplified products was monitored by performing melting curves at the end of each amplification reaction. The primers used in qPCR are listed in Supplementary Table 2.

Cell fractionation

In vitro differentiated T$_H$1 cells were resuspended in RLN1 buffer (50 mM Tris-HCl pH 8, 140 mM NaCl; 1.5 mM $MgCl_2$, 0.5% NP-40) supplemented with SUPERase□In (Ambion) for 10 minutes on ice. After a centrifugation at 300g for 2 minutes, the supernatant was collected as the cytoplasmic fraction. The pellet was resuspended in RLN2 buffer (50 mM Tris-HCl pH 8, 500 mM NaCl, 1.5 mM $MgCl_2$, 0.5% NP-40) supplemented with RNase inhibitors for 10 minutes on ice. Chromatin was pelletted at maximum speed for 3 minutes. The supernatant represents the nuclear fraction. All the fractions were resuspended in TRIzol (Ambion) to 1 ml and RNA was extracted following the standard protocol.

RNA immunoprecipitation (RIP)

In vitro differentiated T$_H$1 cells were UV-crosslinked at 400 mJ/cm^2 in ice-cold D-PBS and then pelleted at 1350 g for 5 minutes. The pellet was resuspended in ice-cold lysis buffer (25 mM Tris-HCl, 150 mM NaCl, 0.5% NP-40) supplemented with 0.5 mM β-mercaptoethanol, Protease Inhibitor Cocktail Tablets cOmplete, EDTA-free (Roche) and SUPERase□In (Ambion) and left rocking at 4°C until the lysis is complete. Debris was centrifuged at 13000 g for 10’. The lysate was precleared with Dynabeads® Protein G (Novex®) for 30 minutes at 4°C and then incubated for 2 hours with 7 mg of antibodies specific for EZH2 (Active Motif - 39875); LSD1 (Abcam – ab17721), or HA (Santa Cruz) as mock control. The lysate was coupled with Dynabeads® Protein G (Novex®) for 1 hour at 4°C. Immunoprecipitates were washed for five times with lysis buffer. RNA was then extracted following mirVana miRNA Isolation Kit (Ambion) protocol. Levels of Linc-MAF-4 or of the negative controls b-actin, RNU2.1 and a region upstream the TSS of linc-MAF-4 (linc-MAF-4 control) were assed by RT-qPCR.

Chromatin Immunoprecipitation analysis (ChIP)

In vitro differentiated T_H1 and T_H2 cells were crosslinked in their medium with 1/10 of fresh formaldehyde solution (50 mM Hepes-KOH pH 7.5, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 11% formaldehyde) for 12 minutes. Then they were treated with 1/10 of 1.25 M glycine for 5 minutes and centrifuged at 1350 g for 5 minutes at 4°C. Cell membranes were lysated in LB1 (50 mM Hepes-KOH pH 7.5, 10 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40 and 0.25% Triton X-100) supplemented with Protease Inhibitor Cocktail Tablets cOmplete, EDTA-free (Roche) and Phenylmethanesulfonyl fluoride (Sigma) at 4°C. Nuclei were pelletted at 1350 g for 5 minutes at 4°C and washed in LB2 (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA) supplemented protease inhibitors. Nuclei were again pelleted at 1350 g for 5 minutes at 4°C and resuspended with a syringe in 200 μl LB3 (10 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Na-deoxycholate, 0.5% N-lauroylscarcosine) supplemented with protease inhibitors. Cell debris were pelleted at 20000 g for 10 minutes at 4°C and a ChIP was set up in LB3 supplemented with 1% Triton X-100, protease inhibitors and antibodies against H3K4me3, H3K27me3 (Millipore), RNA polymerase II STD repeat YSPTSPS, LSD1 (Abcam), EZH2 (Active Motif) or no antibody (as negative control) o/n at 4°C. The day after Dynabeads® Protein G (Novex®) were added at left at 4°C rocking for 2 hours. Then the beads were washed twice with Low salt wash buffer (0.1% SDS, 2 mM EDTA, 1% Triton X-100, 20 mM Tris-HCl pH 8.0, 150 mM NaCl) and with High salt wash buffer (0.1% SDS, 2 mM EDTA, 1% Triton X-100, 20 mM Tris-HCl pH 8.0, 500 mM NaCl). Histones IPs were also washed with a LiCl solution (250 mM LiCl, 1% NP-40, 1 mM EDTA, 10 mM Tris-HCl pH 8.0). All samples were finally washed with 50 mM NaCl in 1X TE. Elution was performed o/n at 65°C in 50 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% SDS. Samples were treated with 0.02 μg/μl RNase A (Sigma) for 2 hours at 37 °C and with 0.04 μg/μl proteinase K (Sigma) for 2 hours at 55°C. DNA was purified with phenol/chloroform extraction.

Chromosome Conformation Capture (3C)

For 3C analysis cells were crosslinked and digested as described for ChIP[53]. Nuclei were resuspended in 500 μl of 1.2X NEB3 buffer (New England BioLabs) with 0.3% SDS and incubated at 37°C for 1h and then with 2% Triton X-100 for another 1h. Digestion was performed with 800U of BglII (New England BioLabs) o/n at 37°C shaking. Digestion was checked loading digested and undigested controls on a 0.6% agarose gel. Then the sample was incubated with 1.6% SDS for 25 minutes at 65°C and with 1.15X ligation buffer (New England BioLabs) and 1% Triton X-100 for 1 hour at 37°C. Ligation was performed with 1000U of T4 DNA ligase (New England BioLabs) for 8 hours at 16°C and at room temperature for 30 minutes. DNA was purified with phenol-chloroform extraction after RNase A (Sigma) and Proteinase K (Sigma) digestion. As controls, BACs corresponding to the region of interested were digested with 100U BglII in NEB3 buffer in 50 μl o/n at 37°C. Then fragments were ligated with 400U T4 DNA ligase o/n at room temperature in 40 μl. PCR products amplified with GoTaq Flexi (Promega) for BACs and samples were run on 2.5% agarose gels and quantified with ImageJ software. Primers are listed in Supplementary Table 3.

Accession numbers

ArrayExpress accession: E-MTAB-2319
Reviewer account:
Username: Reviewer_E-MTAB-2319
Password: ppkieb1o

Author contribution

V.R., A.A. and R.JP.B. setup all the bioinformatics pipelines performed the bioinformatics analyses and contributed to the preparation of the manuscript; G.R. and I.P. designed and performed the main experiments analysed the data and contributed to the preparation of the manuscript; B.B., S.C., P.G. E.P. and E.S. performed experiments and analysed the data; M.M. R.D.F. and J.G. discussed results, provided advice and commented on the manuscript; S.A. and M.P. designed the study, supervised research and wrote the manuscript. All authors discussed and interpreted the results.

Acknowledgments

We would like to thank C. Cheroni for support in statistical analysis; M. Moro and MC. Crosti for technical assistance with cell sorting; D. Gabellini, S.Biffo, P. Della Bona and A. Lanzavecchia for discussions and critical revision of the manuscript; B. J. Haas and A. Dobin for helping the integration of genome guided Trinity with STAR aligner. The INGM Bioinformatic facility for support. Google Summer of Code Project for supporting Clayton Wheeler in the development of https://github.com/csw/bioruby-maf. This study was supported by: the Flagship CNR-MIUR grant "EPIGEN", CARIPLO grant n° 2013-0955, AIRC grant n° IG2013-ID14596, ERC Advanced Grant n° 269022 to S.A, ERC Consolidator Grant n° 617978 to M.P, and by an unrestricted grant of the "Fondazione Romeo ed Enrica Invernizzi".

Figure and Table Legends

Table 1. Purification and RNA-sequencing of human primary lymphocyte subsets
Purity achieved (mean ± SD) by sorting 13 human lymphocyte subsets (isolated from peripheral blood lymphocytes) by various surface marker combinations (sorting phenotype) and number of expressed genes (FPKM> 0.21). Cells were sorted from 4-5 different individuals for each lymphocyte subset and RNA sequencing carried out for each sample separately.

Figure 1. Identification of lincRNAs expressed in human lymphocyte subsets
(a) RNA-seq data generated from 63 lymphocyte samples were processed according to two different strategies: quantification of lincRNAs already annotated in public resources and *de novo* Genome Based Transcripts Reconstruction for the quantification of new lincRNAs expressed in human lymphocytes. Three methods for the identification of new transcripts were adopted: Reference Annotation Based assembly by Cufflinks with two different aligners (TopHat and STAR) and an approach that integrates Trinity and PASA software. Only transcripts reconstructed by at least two assemblers were considered. Novel transcripts were filtered with a computational analysis pipeline to select for lincRNAs. The number of lincRNA genes and transcripts identified in lymphocytes subsets is indicated.
(b) Expression profiles of lincRNA and protein coding genes across 13 human lymphocyte subsets according to K-Means clusters definition. The black line represents the mean expression of the genes belonging to the same cluster. The peaks of expression profiles refer to the populations reported in legend according to numbering.
(c) Specificity of lincRNAs and protein coding genes. Rows and columns are ordered based on a K-Means clustering of lincRNAs and protein coding genes across 13 human lymphocyte populations. Colour intensity represents the Z-score $\log_2$-normalized raw FPKM counts estimated by Cufflinks. 79% of lincRNAs genes and 39% of protein coding genes are assigned to specific clusters. See also Supplementary Fig. 1h.
(d) As in (c), performed on receptors and metabolic processes genes.

Figure 2. Definition of lincRNA signatures in human lymphocyte subsets

(a) Heatmap of normalized expression values of lymphocytes signature lincRNAs selected on the basis of fold change (>2.5 with respect to all the other subsets), intrapopulation consistency (expressed in at least 3 out of 5 samples) and non parametric Kruskal-Wallis test (pval < 0.05). Signature lincRNAs relative expression values were calculated as $\log_2$ ratios between lymphocyte subsets and a panel of human lymphoid and non lymphoid tissues of the Human BodyMap 2.0 project (See also Supplementary Fig. 2b-m).

(b) $CD4^+$ T_H1 signature lincRNAs extracted from panel (A). The barcode on the left indicates already annotated lincRNAs (white) and newly described lincRNAs (brick red). For newly described lincRNAs name, 'S' and 'AS' indicates 'sense' and 'antisense' respectively.

(c) Average expression levels of already annotated (white) and newly described (brick red) lincRNAs in human lymphocyte subsets and lymphoid or non-lymphoid human tissues.

(d) Validation of T_H1 signature lincRNAs expression by RT-qPCR on primary $CD4^+$ naïve, T_H1 and Treg cells sorted from PBMC of healthy donors (average of three independent experiments ± SEM).

(e) RT-qPCR analysis of T_H1 signature lincRNAs expression in a time course of $CD4^+$ naïve T cells differentiated in T_H1 and T_H2 polarizing conditions presented as relative quantity (RQ) relative to time zero (average of three independent experiments).

Figure 3. Linc-MAF-4 contributes to T_H1 cell differentiation.

(a) Gene Ontology (GO) semantic similarity matrix of protein coding genes proximal to lincRNA signatures. The semantic similarity scores for all GO term pairs were clustered using hierarchical clustering method. On the right of the matrix a bar plot of the adjusted p-values for each GO term is reported. Red bars represent GO terms that are significantly enriched in Gene Ontology analysis. Common ancestor is reported for each cluster.

(b) Expression of linc-MAF-4 and MAF assessed at different time points by RT-qPCR in activated CD4$^+$ naïve T cells differentiated in T_H1 or T_H2 polarizing conditions (average of four technical replicates ± SEM). See also Supplementary Fig. 3c.

(c) ChIP-qPCR analysis of H3K4me3 and RNA polymerase II occupancy at *MAF* locus in CD4$^+$ naïve T cells differentiated in T_H1 or T_H2 polarizing conditions at day 8 post activation. Enrichment is a percentage of input (average of at least 5 independent experiments ± SEM). One-tailed t-test * $p < 0.05$.

(d) As in (c) at *IFNG* locus as control (average of at least 10 independent experiments ± SEM). One-tailed t-test * $p < 0.05$; ** $p < 0.01$.

(e) Linc-MAF-4 and MAF expression levels determined by RT-qPCR in activated CD4$^+$ naïve T cells (in the absence of polarizing cytokines) and transfected at the same time with linc-MAF-4 siRNA (black) or ctrl siRNA (white). Transcripts expression was detected 72 hours post transfection (average of six independent experiments ± SEM). One-tailed t-test ** $p < 0.01$; * $p < 0.05$.

(f) Results of GSEA (Gene Set Enrichment Analysis) performed on gene expression data obtained from siRNA mediated knock-down of linc-MAF-4 in activated CD4 naïve T cells. Activation and transfection conditions were as in (e). The red and blue line represent the observed enrichment score profile of genes in the linc-MAF-4 / ctrl siRNA treated cells compared to the CD4 T_H1 and T_H2 reference gene sets respectively (average of four independent experiments). Nominal p-val <0.05

(g) GATA3 and IL4 expression levels determined by RT-qPCR in activated CD4$^+$ naïve T cells transfected with linc-MAF-4 siRNA (black) or ctrl siRNA (white) (average of six independent experiments ± SEM). One-tailed t-test ** $p < 0.01$; * $p < 0.05$.

Figure 4. Epigenetic characterization of linc-MAF4/MAF genomic locus

(a) Schematic representation of the region analyzed by 3C. The M1 primer, located near the 5'-end of *MAF*, was used as bait. Primers spanning the region between *linc-MAF-4* and *MAF* were tested for interaction. 3C results show the relative frequency of interaction between *MAF* 5'-end and *linc-MAF-4* 5'- (L7 primer) and 3'- (L12 primer) ends in CD4$^+$ naïve T cells differentiated in T_H1 polarizing conditions (day 8) (average of three independent experiments ± SEM). (b) Sequencing results with pertaining electropherograms and BLAST alignments for M1-L7 and M1-L12 amplicons.

(c) Relative abundance of linc-MAF-4 transcript in cytoplasm, nucleus and chromatin in $CD4^+$ naïve T cells differentiated in T_H1 polarizing conditions (day 8). Linc-00339, Malat1 and RNU2.1 were used respectively as cytoplasmic, nuclear and chromatin-associated controls (average of three independent experiments ± SEM).

(d) RIP assay for LSD1 and EZH2 in $CD4^+$ naïve T cells differentiated in T_H1 polarizing conditions (day 8). The enrichment of linc-MAF-4 is relative to mock. β-actin, RNU2.1 and a region upstream the TSS of linc-MAF-4 were chosen as controls (average of six independent experiments ± SEM). The statistical significance was determined with ANOVA and Dunnet post-hoc test: *$p<0.05$; **$p<0.01$.

(e) ChIP-qPCR analysis of EZH2, H3K27me3 and LSD1 occupancy at *MAF* locus in activated $CD4^+$ naïve T cells transfected with linc-MAF-4 siRNA (black) or ctrl siRNA (white) (average of at least three independent experiments ± SEM). One-tailed t-test * $p < 0.05$.

(f) Model for linc-MAF-4-mediated *MAF* repression in T_H1 lymphocytes. When linc-MAF-4 is expressed, it recruits chromatin remodelers (i.e. LSD1 and EZH2) at *MAF* 5'-end, taking advantage of a DNA loop that brings in close proximity *linc-MAF-4* 5'- and 3'- end and *MAF* 5'-end. This event causes the downregulation of *MAF* transcription and enforces T_H1 cell fate, contrasting T_H2 differentiation.

Subset	**Purity (%)**	**Sorting phenotype**	**Genes**
CD4$^+$ naïve	99.8 ± 0.1	CD4$^+$ CCR7$^+$ CD45RA$^+$ CD45RO$^-$	20061
CD4$^+$ T$_H$1	99.9 ± 0.05	CD4$^+$ CXCR3$^+$	20855
CD4$^+$ T$_H$2	99.7 ± 0.3	CD4$^+$ CRTH2$^+$ CXCR3$^-$	19623
CD4$^+$ T$_H$17	99.1 ± 1	CD4$^+$ CCR6$^+$ CD161$^+$ CXCR3$^-$	20959
CD4$^+$ T$_{reg}$	99.0 ± 0.8	CD4$^+$ CD127$^-$ CD25$^+$	21435
CD4$^+$ T$_{CM}$	98.4 ± 2.8	CD4$^+$ CCR7$^+$ CD45RA$^-$ CD45RO$^+$	20600
CD4$^+$ T$_{EM}$	95.4 ± 5.5	CD4$^+$ CCR7$^-$ CD45RA$^-$ CD45RO$^+$	19800
CD8$^+$ T$_{CM}$	98.3 ± 0.8	CD8$^+$ CCR7$^+$ CD45RA$^-$ CD45RO$^+$	20901
CD8$^+$ T$_{EM}$	96.8 ± 0.9	CD8$^+$ CCR7$^-$ CD45RA$^-$ CD45RO$^+$	21813
CD8$^+$ naïve	99.3 ± 0.2	CD8$^+$ CCR7$^+$ CD45RA$^+$ CD45RO$^-$	20611
B naïve	99.9 ± 0.1	CD19$^+$ CD5$^-$ CD27$^-$	21692
B memory	99.1 ± 0.8	CD19$^+$ CD5$^-$ CD27$^+$	21239
B CD5$^+$	99.1 ± 0.8	CD19$^+$ CD5$^+$	22499

Figure 1

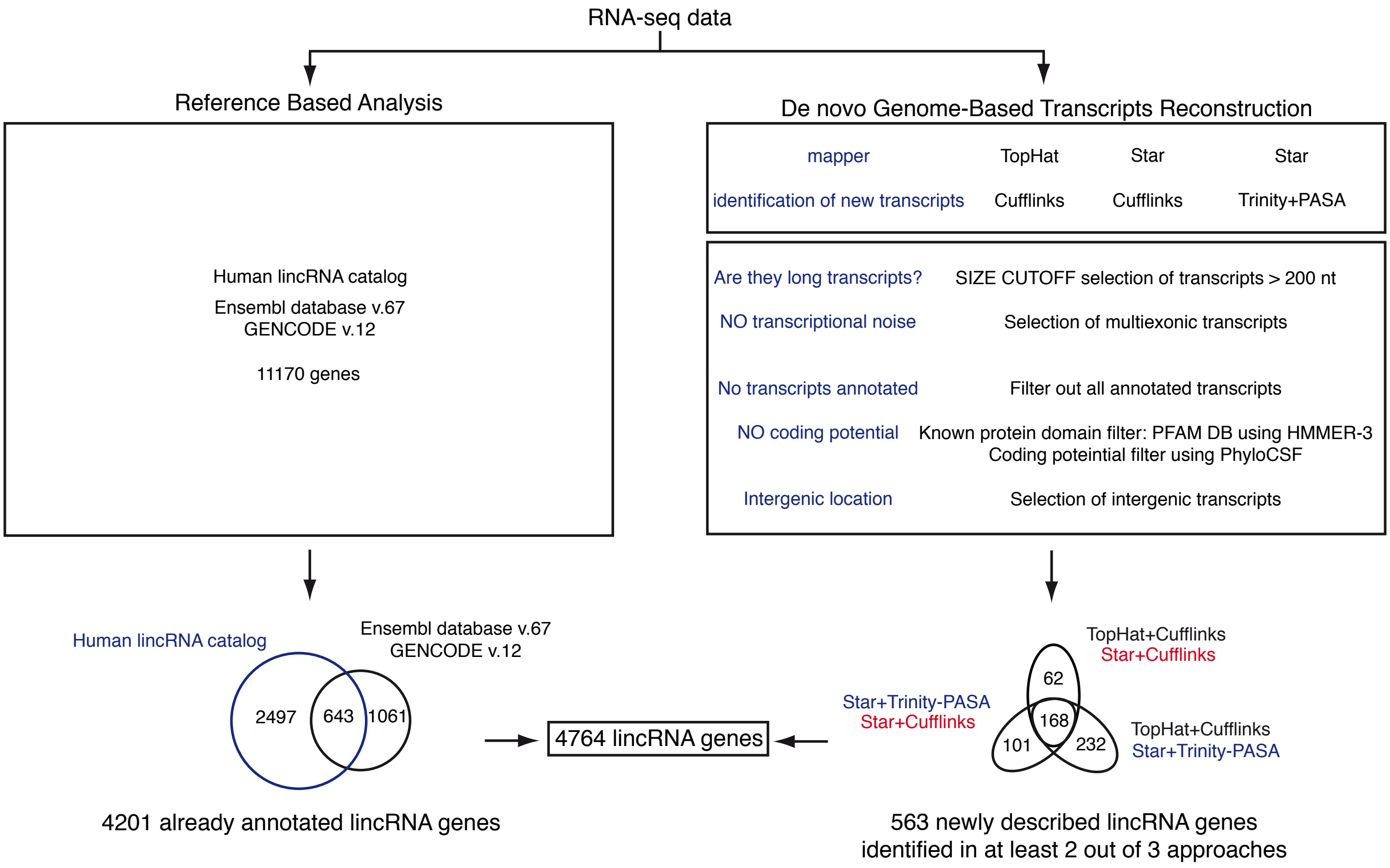

RNA-seq data
Reference Based Analysis
Human lincRNA catalog
Ensembl database v.67
GENCODE v.12
11170 genes
De novo Genome-Based Transcripts Reconstruction
mapper
identification of new transcripts
TopHat
Cufflinks
Star
Cufflinks
Star
Trinity+PASA
Are they long transcripts?
SIZE CUTOFF selection of transcripts > 200 nt
NO transcriptional noise
Selection of multiexonic transcripts
No transcripts annotated
Filter out all annotated transcripts
NO coding potential
Known protein domain filter: PFAM DB using HMMER-3
Coding poteintial filter using PhyloCSF
Intergenic location
Selection of intergenic transcripts
Human lincRNA catalog
Ensembl database v.67
GENCODE v.12
2497
643
1061
4201 already annotated lincRNA genes
4764 lincRNA genes
TopHat+Cufflinks
Star+Cufflinks
Star+Trinity-PASA
Star+Cufflinks
62
168
101
232
TopHat+Cufflinks
Star+Trinity-PASA
563 newly described lincRNA genes
identified in at least 2 out of 3 approaches

Figure 2

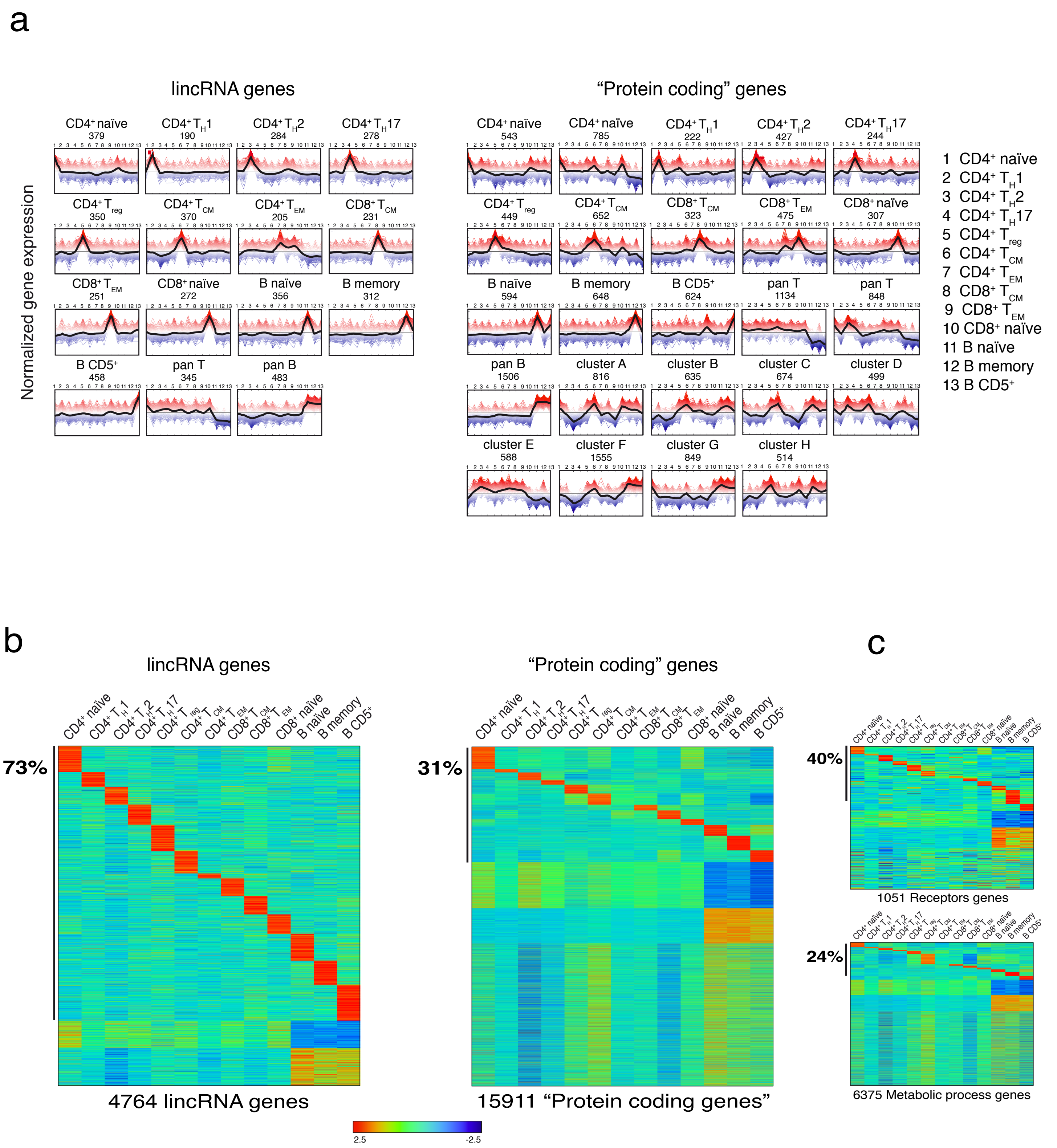
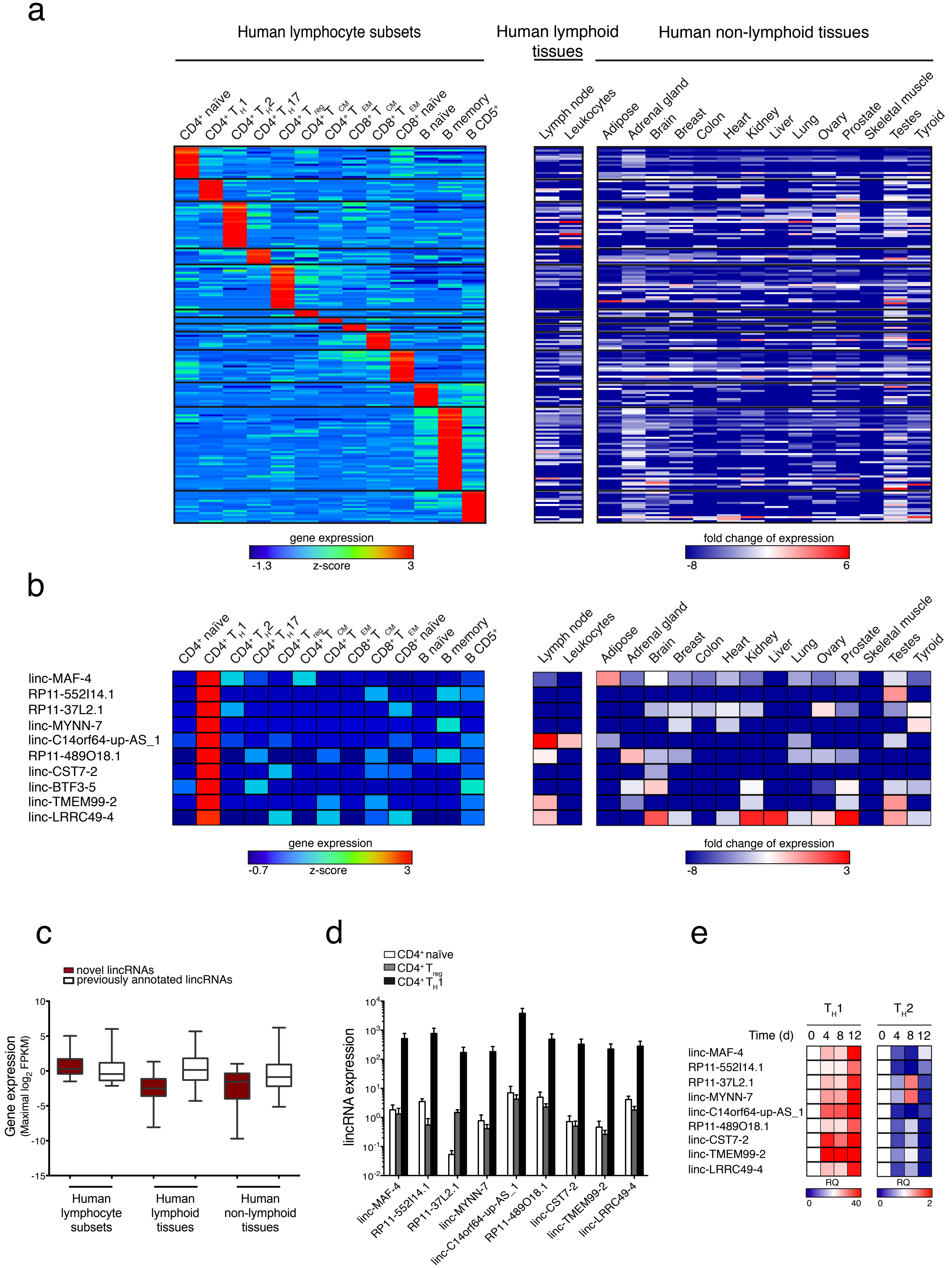

a
lincRNA genes
"Protein coding" genes
Normalized gene expression
CD4⁺ naïve 379
CD4⁺ TH1 190
CD4⁺ TH2 284
CD4⁺ TH17 278
CD4⁺ Treg 350
CD4⁺ TCM 370
CD4⁺ TEM 205
CD8⁺ TCM 231
CD8⁺ TEM 251
CD8⁺ naïve 272
B naïve 356
B memory 312
B CD5⁺ 458
pan T 345
pan B 483
CD4⁺ naïve 543
CD4⁺ naïve 785
CD4⁺ TH1 222
CD4⁺ TH2 427
CD4⁺ TH17 244
CD4⁺ Treg 449
CD4⁺ TCM 652
CD8⁺ TCM 323
CD8⁺ TEM 475
CD8⁺ naïve 307
B naïve 594
B memory 648
B CD5⁺ 624
pan T 1134
pan T 848
pan B 1506
cluster A 816
cluster B 635
cluster C 674
cluster D 499
cluster E 588
cluster F 1555
cluster G 849
cluster H 514
1 CD4⁺ naïve
2 CD4⁺ TH1
3 CD4⁺ TH2
4 CD4⁺ TH17
5 CD4⁺ Treg
6 CD4⁺ TCM
7 CD4⁺ TEM
8 CD8⁺ TCM
9 CD8⁺ TEM
10 CD8⁺ naïve
11 B naïve
12 B memory
13 B CD5⁺
b
lincRNA genes
"Protein coding" genes
CD4⁺ naïve
CD4⁺ TH1
CD4⁺ TH2
CD4⁺ TH17
CD4⁺ Treg
CD4⁺ TCM
CD4⁺ TEM
CD8⁺ TCM
CD8⁺ TEM
CD8⁺ naïve
B naïve
B memory
B CD5⁺
73%
31%
4764 lincRNA genes
15911 "Protein coding genes"
2.5
-2.5
c
40%
1051 Receptors genes
24%
6375 Metabolic process genes

Figure 3



a
Human lymphocyte subsets
Human lymphoid tissues
Human non-lymphoid tissues
CD4⁺ naïve
CD4⁺ T_H1
CD4⁺ T_H2
CD4⁺ T_H17
CD4⁺ T_reg
CD4⁺ T_CM
CD4⁺ T_EM
CD8⁺ T_CM
CD8⁺ T_EM
CD8⁺ naïve
B naïve
B memory
B CD5⁺
Lymph node
Leukocytes
Adipose
Adrenal gland
Brain
Breast
Colon
Heart
Kidney
Liver
Lung
Ovary
Prostate
Skeletal muscle
Testes
Tyroid
gene expression
-1.3
z-score
3
fold change of expression
-8
6
b
linc-MAF-4
RP11-552I14.1
RP11-37L2.1
linc-MYNN-7
linc-C14orf64-up-AS_1
RP11-489O18.1
linc-CST7-2
linc-BTF3-5
linc-TMEM99-2
linc-LRRC49-4
gene expression
-0.7
z-score
3
fold change of expression
-8
3
c
novel lincRNAs
previously annotated lincRNAs
Gene expression
(Maximal log2 FPKM)
10
5
0
-5
-10
-15
Human lymphocyte subsets
Human lymphoid tissues
Human non-lymphoid tissues
d
CD4⁺ naïve
CD4⁺ T_reg
CD4⁺ T_H1
lincRNA expression
10^4
10^3
10^2
10^1
10^0
10^-1
10^-2
linc-MAF-4
RP11-552I14.1
RP11-37L2.1
linc-MYNN-7
linc-C14orf64-up-AS_1
RP11-489O18.1
linc-CST7-2
linc-TMEM99-2
linc-LRRC49-4
e
T_H1
T_H2
Time (d)
0 4 8 12
0 4 8 12
linc-MAF-4
RP11-552I14.1
RP11-37L2.1
linc-MYNN-7
linc-C14orf64-up-AS_1
RP11-489O18.1
linc-CST7-2
linc-TMEM99-2
linc-LRRC49-4
RQ
0
40
RQ
0
2

Figure 4

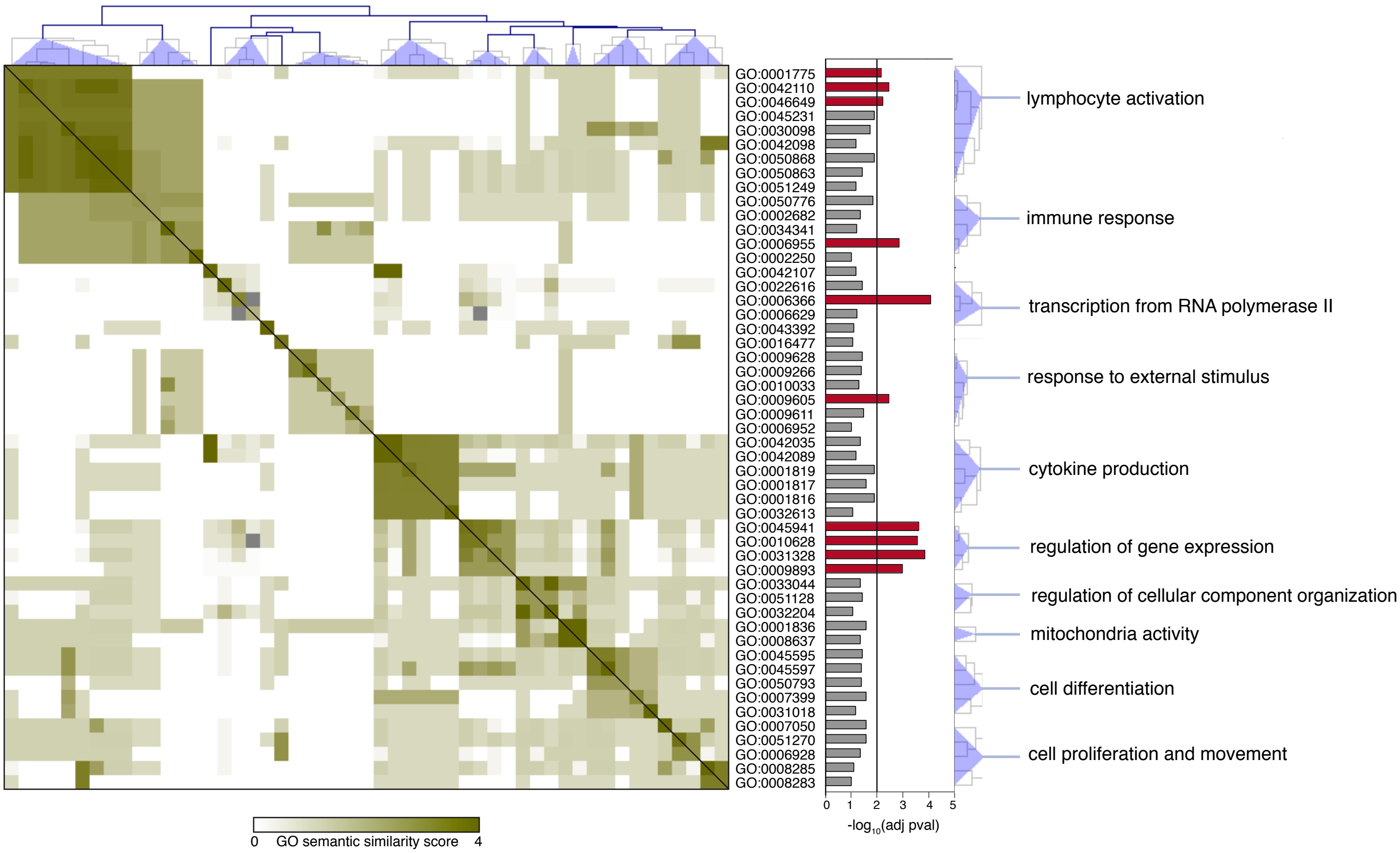

GO:0001775
GO:0042110
GO:0046649
GO:0045231
GO:0030098
GO:0042098
GO:0050868
GO:0050863
GO:0051249
GO:0050776
GO:0002682
GO:0034341
GO:0006955
GO:0002250
GO:0042107
GO:0022616
GO:0006366
GO:0006629
GO:0043392
GO:0016477
GO:0009628
GO:0009266
GO:0010033
GO:0009605
GO:0009611
GO:0006952
GO:0042035
GO:0042089
GO:0001819
GO:0001817
GO:0001816
GO:0032613
GO:0045941
GO:0010628
GO:0031328
GO:0009893
GO:0033044
GO:0051128
GO:0032204
GO:0001836
GO:0008637
GO:0045595
GO:0045597
GO:0050793
GO:0007399
GO:0031018
GO:0007050
GO:0051270
GO:0006928
GO:0008285
GO:0008283
0
1
2
3
4
5
-log$_{10}$(adj pval)
0
GO semantic similarity score
4
lymphocyte activation
immune response
transcription from RNA polymerase II
response to external stimulus
cytokine production
regulation of gene expression
regulation of cellular component organization
mitochondria activity
cell differentiation
cell proliferation and movement

Figure 5

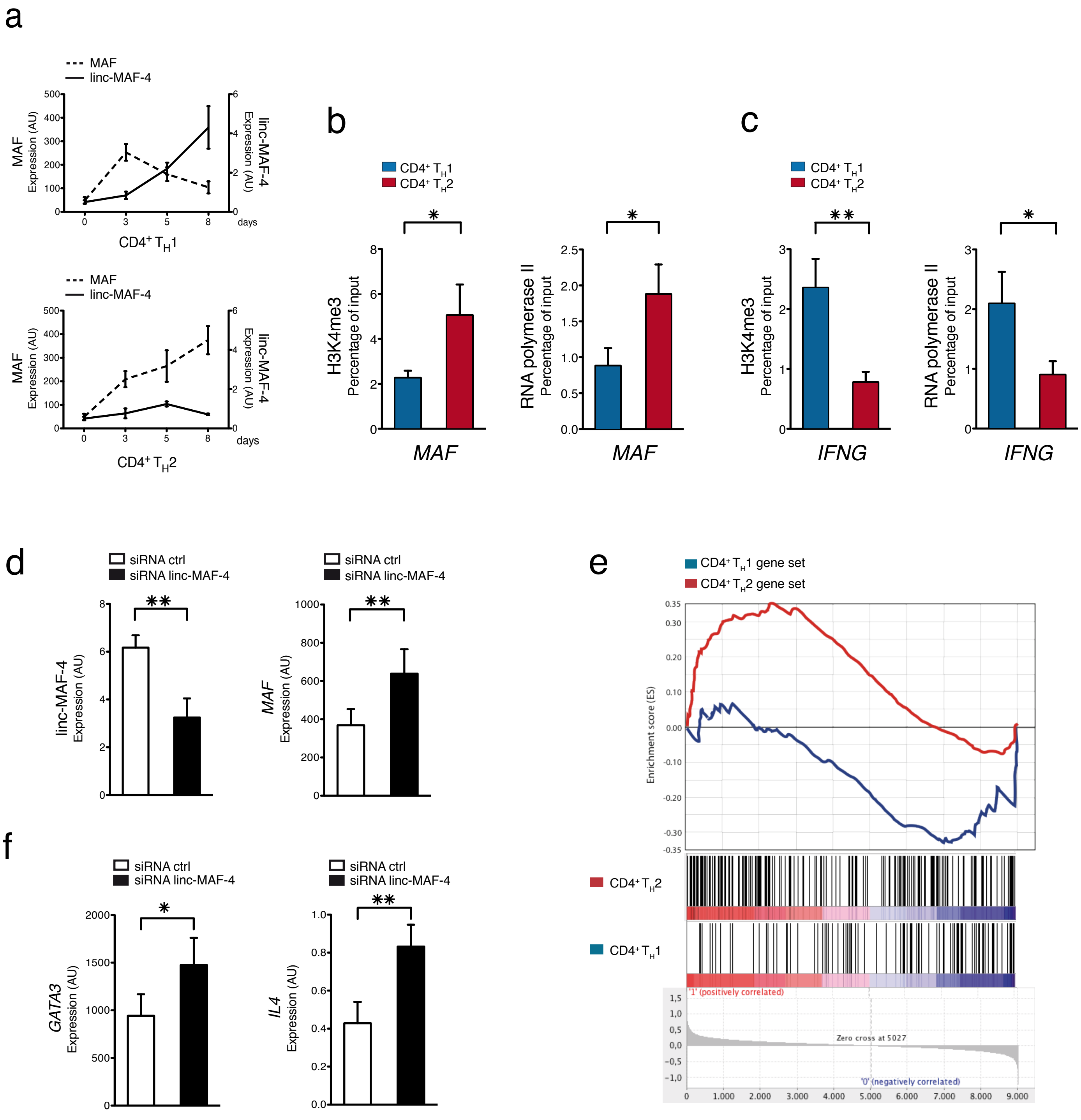
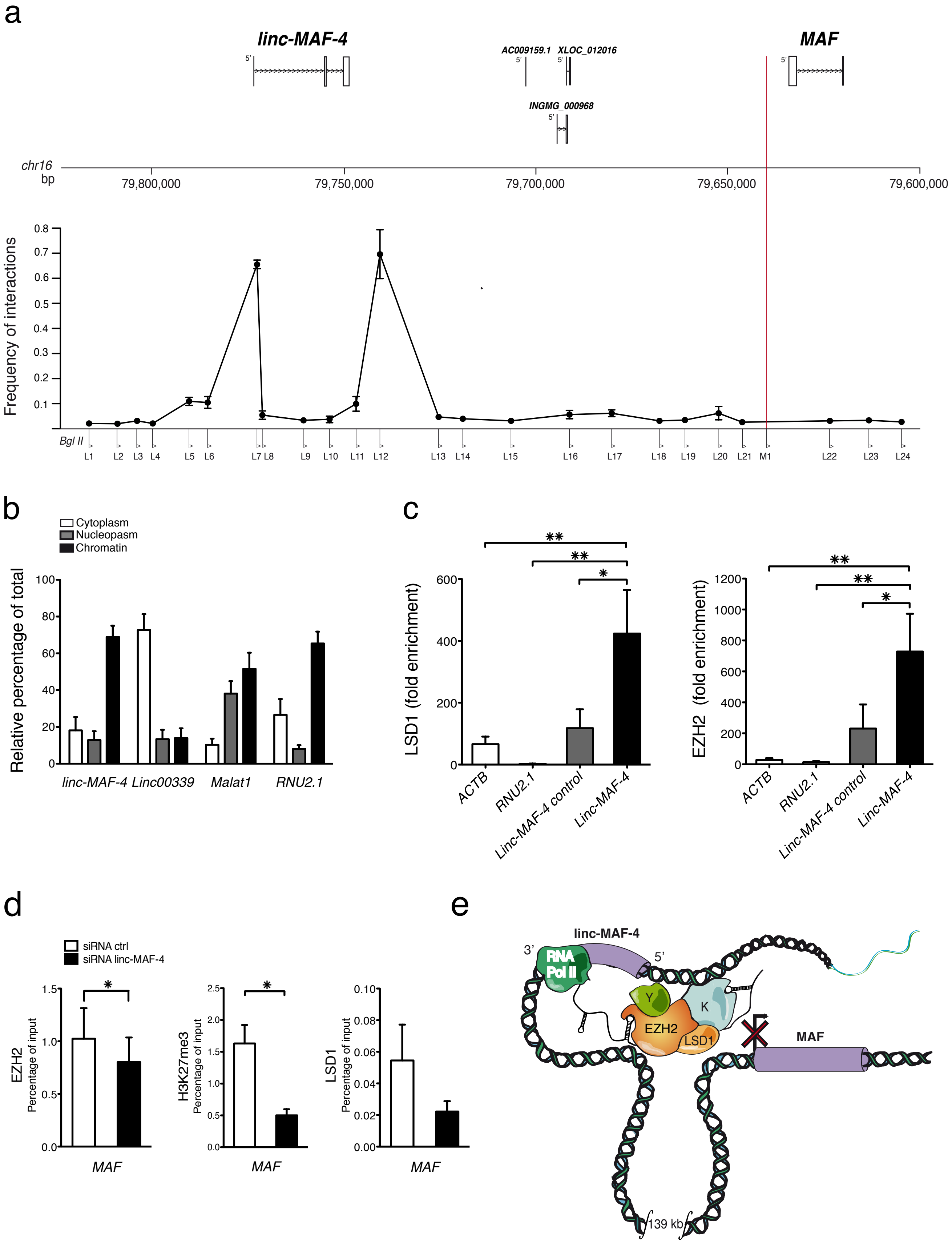
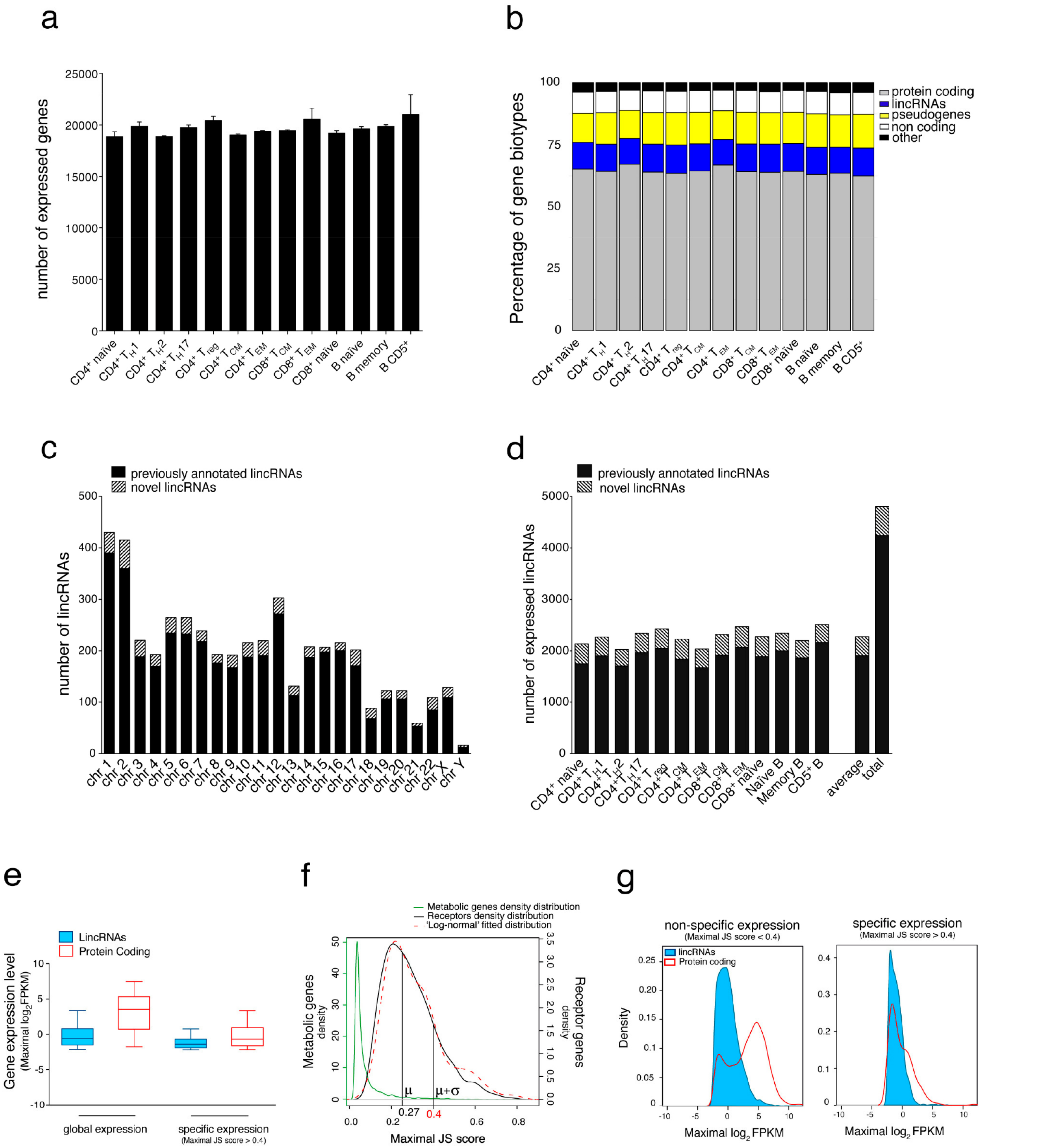

a
MAF
linc-MAF-4
MAF Expression (AU)
linc-MAF-4 Expression (AU)
500
400
300
200
100
0
6
4
2
0
0
3
5
8
days
CD4+ TH1
CD4+ TH2
b
CD4+ TH1
CD4+ TH2
*
H3K4me3 Percentage of input
RNA polymerase II Percentage of input
8
6
4
2
0
2.5
2.0
1.5
1.0
0.5
0.0
MAF
MAF
c
CD4+ TH1
CD4+ TH2
**
*
3
2
1
0
IFNG
IFNG
d
siRNA ctrl
siRNA linc-MAF-4
**
linc-MAF-4 Expression (AU)
MAF Expression (AU)
8
6
4
2
0
1000
800
600
400
200
0
e
CD4+ TH1 gene set
CD4+ TH2 gene set
Enrichment score (ES)
0.35
0.30
0.20
0.10
0.00
-0.10
-0.20
-0.30
-0.35
CD4+ TH2
CD4+ TH1
'1' (positively correlated)
Zero cross at 5027
'0' (negatively correlated)
1,5
1,0
0,5
0,0
-0,5
-1,0
0
1.000
2.000
3.000
4.000
5.000
6.000
7.000
8.000
9.000
f
siRNA ctrl
siRNA linc-MAF-4
*
**
GATA3 Expression (AU)
IL4 Expression (AU)
2000
1500
1000
500
0
1.0
0.8
0.6
0.4
0.2
0.0

Figure 6


a
linc-MAF-4
AC009159.1
XLOC_012016
INGMG_000968
MAF
5'
chr16
bp
79,800,000
79,750,000
79,700,000
79,650,000
79,600,000
Frequency of interactions
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Bgl II
L1 L2 L3 L4 L5 L6 L7 L8 L9 L10 L11 L12 L13 L14 L15 L16 L17 L18 L19 L20 L21 M1 L22 L23 L24
b
Cytoplasm
Nucleopasm
Chromatin
Relative percentage of total
100
80
60
40
20
0
linc-MAF-4
Linc00339
Malat1
RNU2.1
c
LSD1 (fold enrichment)
600
400
200
0
EZH2 (fold enrichment)
1200
1000
800
600
400
200
0
**
*
ACTB
RNU2.1
Linc-MAF-4 control
Linc-MAF-4
d
siRNA ctrl
siRNA linc-MAF-4
EZH2
Percentage of input
1.5
1.0
0.5
0.0
H3K27me3
Percentage of input
2.5
2.0
1.5
1.0
0.5
0.0
LSD1
Percentage of input
0.10
0.08
0.06
0.04
0.02
0.00
MAF
*
e
linc-MAF-4
3'
5'
RNA Pol II
Y
K
EZH2
LSD1
MAF
139 kb



Supplementary Figure 1

Distribution and expression of lincRNAs in primary human lymphocytes subsets.

(a) Bar plots of expressed genes across a panel of 13 lymphocyte subsets. Average expression (± sdev) of at least four samples for each subset is reported

(b) Stacked bar plots of expressed genes percentages according to their biotype (protein coding, lincRNAs, pseudogenes, non-coding genes and other) across the analyzed human lymphocyte subsets

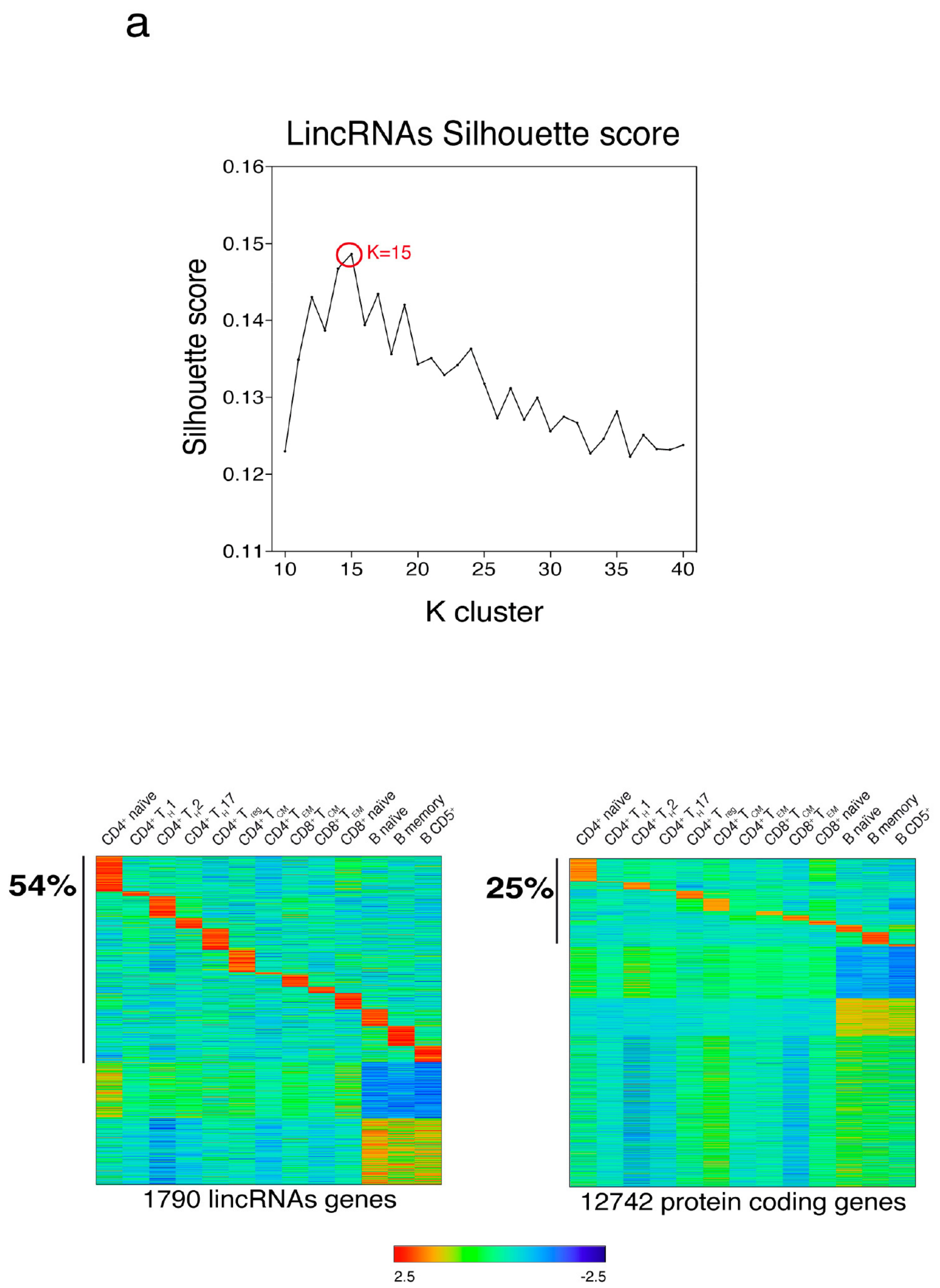
(c) Distribution of novel (striped) and previously annotated (black) lincRNAs in all human chromosomes

(d) Distribution of expressed novel (striped) and previously annotated (black) lincRNAs across the analyzed human lymphocyte subsets.

(e) Boxplots of gene expression values of lincRNA (blue) and protein coding genes (red) on either the whole dataset (global expression) or on a dataset filtered according to the specificity score (specific expression, Maximal JS score > 0.4)

(f) The density distribution of JS score for cell-specific receptor genes (black line) was fitted to a log-normal distribution (dotted red line). In order to derive a threshold for the cell-specificity score, we calculated the JS score value corresponding to one standard deviation away from the mean value of the fitted distribution (0.27). As a reference, the JS density distribution for the metabolic genes is reported (green line)

(g) Density distributions of maximal expression values of lincRNAs (blue area plot) and protein coding genes (red line), divided according to cellular specificity (maximal JS score < 0.4 or JS score > 0.4)

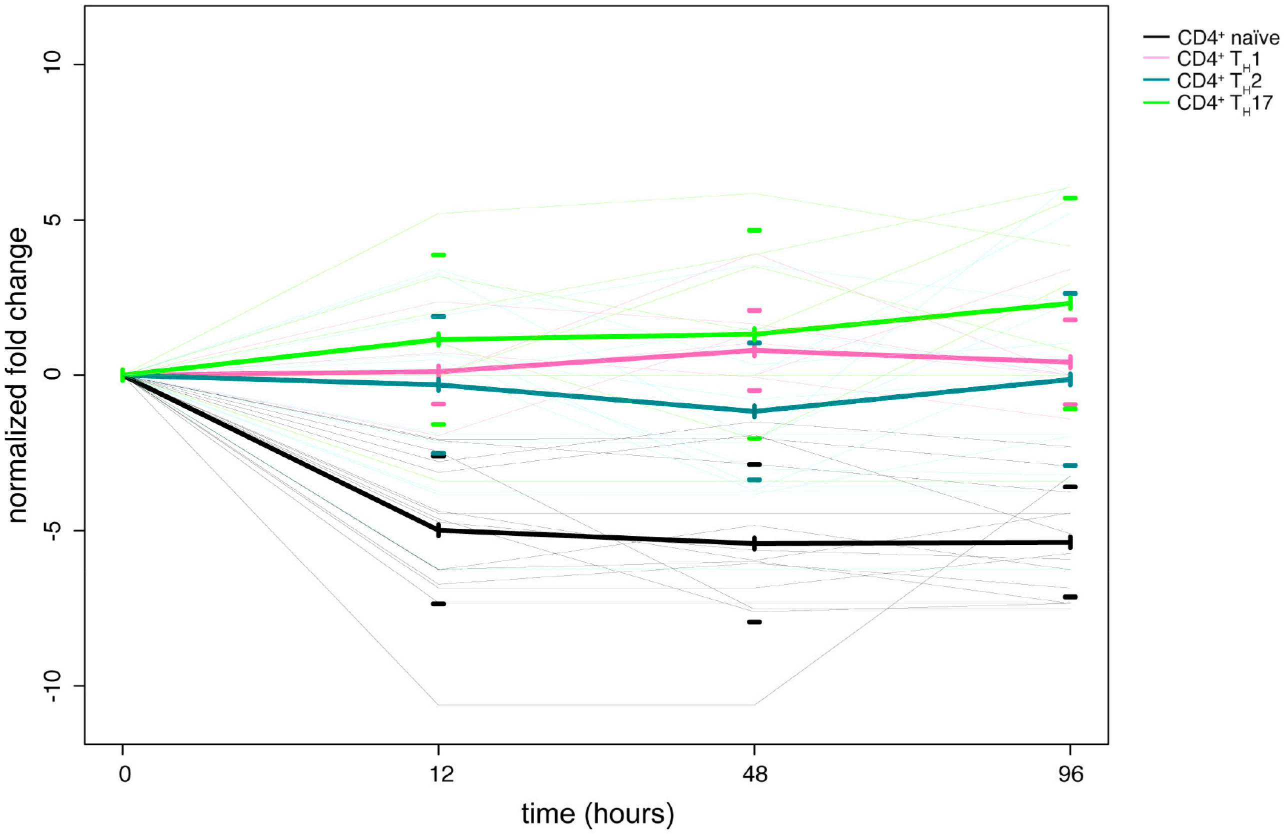


Supplementary Figure 2

Specificity of lincRNAs and protein-coding genes in primary human lymphocytes subsets.

(a) Silhouette scores (y-axis) are reported as a function of K (x-axis), the number of clusters used to partition the gene expression dataset of lincRNA genes. The average Silhouette value was calculated by taking the average of each clusters's average Si. In the graph Si data are reported for lincRNAs genes, for which the highest Si value (implying better clustering of the data) is 15

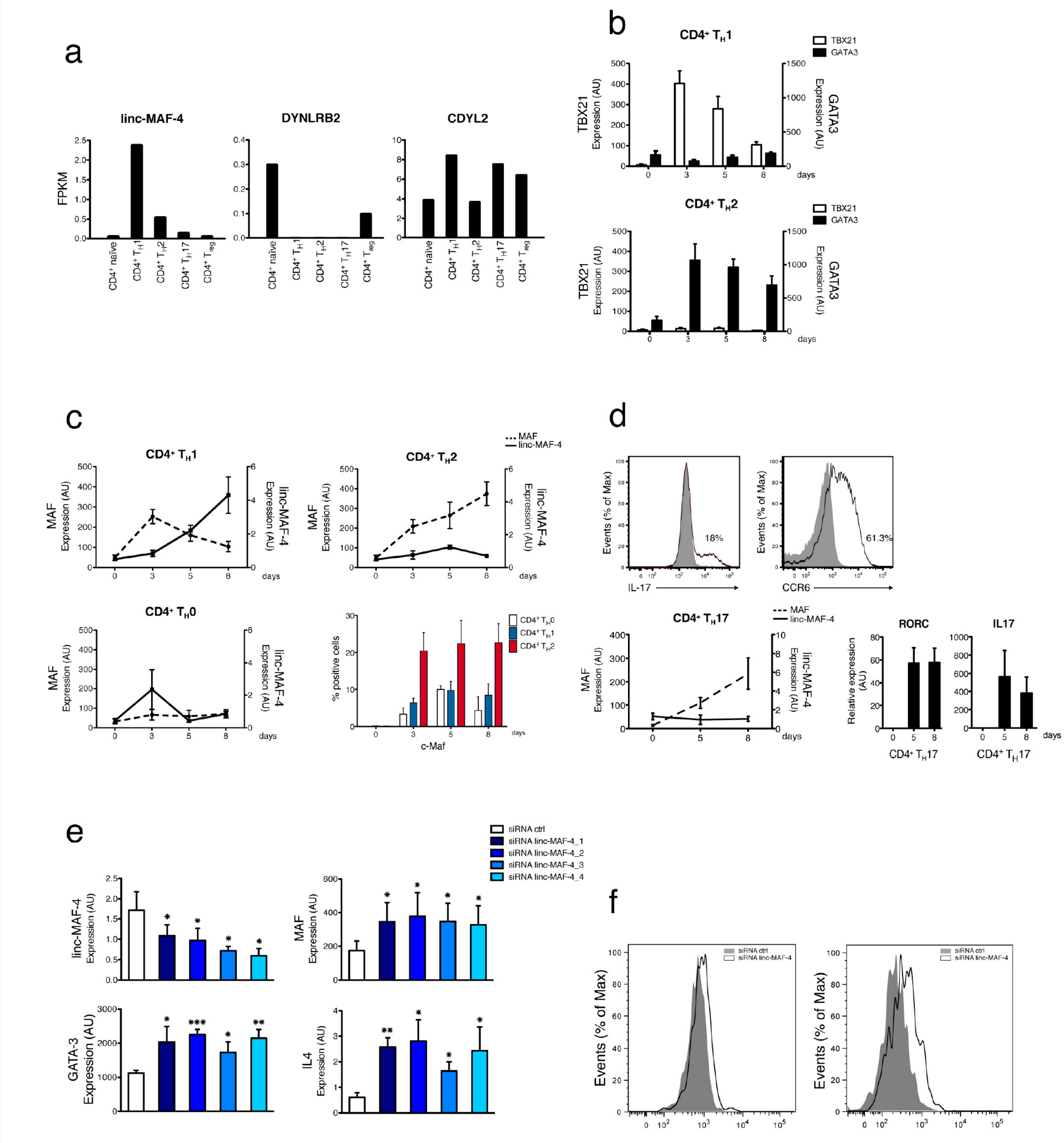
(b) Specificity of lincRNAs and protein coding genes (FPKM >1) by K-Means clustering across 13 human lymphocyte populations. Colour intensity represents the Z-score log2-normalized raw FPKM counts estimated by Cufflinks



Supplementary Figure 3

LincRNA signatures in a differentiation time course.

CD4$^+$ naïve, T$_H$1, T$_H$2 and T$_H$17 signature lincRNAs trends in CD4$^+$ naïve T cells differentiated in T$_H$0 conditions. RNA was collected at different time points during CD4+ naïve T cells differentiation and RNA-seq experiments were performed. Thin lines represent the trends of each signature lincRNA. Bold lines represent the average trend of all signature lincRNAs for each subset. Data are represented as a log2 normalized ratio between each time point and the relative time 0.



Supplementary Figure 4

Regulation of *MAF* transcription by linc-MAF-4.

(a) Expression levels (FPKM) of linc-MAF-4 and its neighboring protein coding genes DYNLRB2 and CDYL2 in CD4+ T cell subsets

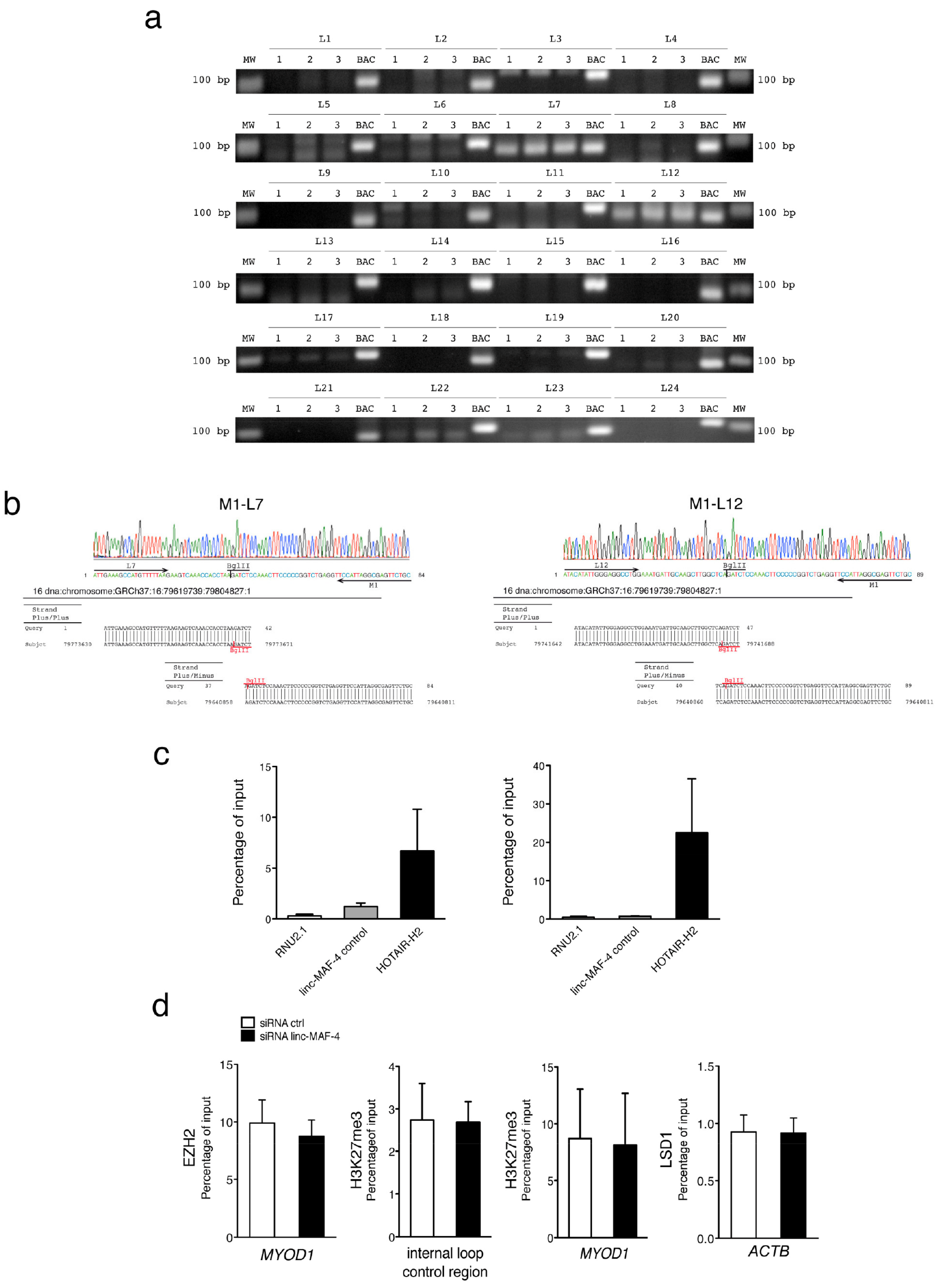
(b) Expression of TBX21 an GATA3 in activated CD4+ naïve T cells differentiated in T_H1 or T_H2 polarizing conditions assessed at different time points by RT-qPCR (average of four independent experiments ± SEM)

(c) Expression of linc-MAF-4 and MAF assessed at different time points by RT-qPCR in activated CD4+ naïve T cells differentiated in T_H1, T_H2 and T_H0 polarizing conditions. Bar plot of the percentage of c-Maf positive cells determined by intracellular staining at different time points is also shown (average of four independent experiments ± SEM)

(d) CD4+ naïve T cells differentiated in T_H17 polarizing conditions according to Kleinewietfeld et al. (Nature 2013; 496, 518). Upper panels: intracellular staining of IL-17 and CCR6 protein expression at day 8 of differentiation (data are representative of four independent experiments) Lower panels: linc-MAF-4, MAF, RORC and IL17 transcript levels assessed at different time points by RT-qPCR (average of four independent experiments ± SEM)

(e) Test of linc-MAF-4 siRNAs in CD4+ naïve T cells. Four siRNA sequences were transfected independently in activated CD4+ naïve T cells and linc-MAF-4, MAF, GATA3 and IL4 transcript levels were assessed by RT-qPCR at day 3 post-transfection and activation (average of five independent experiments ± SEM)

(f) Intracellular staining of c-Maf and GATA-3 in naive CD4+ T cells stimulated with anti-CD3 and anti-CD28 and transfected with a control siRNA or linc-MAF-4 siRNA assessed at day 4 post-transfection and activation. Data are representative of five independent experiments


a
L1
L2
L3
L4
MW
1
2
3
BAC
100 bp
L5
L6
L7
L8
L9
L10
L11
L12
L13
L14
L15
L16
L17
L18
L19
L20
L21
L22
L23
L24
b
M1-L7
L7
BglII
1 ATTGAAAGCCATGTTTTTAAGAAGTCAAACCACCTAAGATCTCCAAACTTCCCCCGGTCTGAGGTTCCATTAGGCGAGTTCTGC 84
M1
16 dna:chromosome:GRCh37:16:79619739:79804827:1
Strand
Plus/Plus
Query 1 ATTGAAAGCCATGTTTTTAAGAAGTCAAACCACCTAAGATCT 42
Subjct 79773630 ATTGAAAGCCATGTTTTTAAGAAGTCAAACCACCTAAGATCT 79773671
BglII
Strand
Plus/Minus
BglII
Query 37 AGATCTCCAAACTTCCCCCGGTCTGAGGTTCCATTAGGCGAGTTCTGC 84
Subjct 79640858 AGATCTCCAAACTTCCCCCGGTCTGAGGTTCCATTAGGCGAGTTCTGC 79640811
M1-L12
L12
BglII
1 ATACATATTGGGAGGCCTGGAAATGATTGCAAGCTTGGCTCAGATCTCCAAACTTCCCCCGGTCTGAGGTTCCATTAGGCGAGTTCTGC 89
M1
16 dna:chromosome:GRCh37:16:79619739:79804827:1
Strand
Plus/Plus
Query 1 ATACATATTGGGAGGCCTGGAAATGATTGCAAGCTTGGCTCAGATCT 47
Subjct 79741642 ATACATATTGGGAGGCCTGGAAATGATTGCAAGCTTGGCTCAGATCT 79741688
BglII
Strand
Plus/Minus
BglII
Query 40 TCAGATCTCCAAACTTCCCCCGGTCTGAGGTTCCATTAGGCGAGTTCTGC 89
Subjct 79640860 TCAGATCTCCAAACTTCCCCCGGTCTGAGGTTCCATTAGGCGAGTTCTGC 79640811
c
Percentage of input
15
10
5
0
RNU2.1
linc-MAF-4 control
HOTAIR-H2
Percentage of input
40
30
20
10
0
RNU2.1
linc-MAF-4 control
HOTAIR-H2
d
siRNA ctrl
siRNA linc-MAF-4
EZH2
Percentage of input
15
10
5
0
MYOD1
H3K27me3
Percentageof input
4
3
2
1
0
internal loop
control region
H3K27me3
Percentageof input
15
10
5
0
MYOD1
LSD1
Percentage of input
1.5
1.0
0.5
0.0
ACTB

Supplementary Figure 5

Chromosome-conformation capture on *in vitro*–differentiated CD4$^+$ T$_H$1 cells.

(a) 2.5% agarose gel of the experimental triplicate used for 3C followed by BAC controls amplified with different primers that span the region between linc-MAF-4 and MAF

(b) Sequencing results with pertaining electropherograms and BLAST alignments for M1-L7 and M1-L12 amplicons

(c) Validation of anti-LSD1 and EZH2 antibodies used in RIP assay. LSD1 and EZH2 immunoprecipitates specifically retrieve HOTAIR RNA in HeLa cells as shown by Tsai et al. Science 329, 689 (2010). RNU2.1 and a region upstream the TSS of linc-MAF-4 were used as negative controls

(d) ChIP-qPCR analysis of EZH2 and H3K27me3 at MYOD1 locus, of H3K27me3 at a control region within the chromatin loop and of LSD1 at beta-actin locus in activated CD4+ naïve T cells transfected with linc-MAF-4 siRNA (black) or ctrl siRNA (white) (average of at least three independent experiments ± SEM)

Additional considerations for *de novo* genome-based transcripts reconstruction

Three different approaches were adopted to define a new catalog of lincRNA specifically expressed in human lymphocyte subsets. These approaches are based on the application of two different mappers TopHat v.1.4.1 (Trapnell et al. 2009) and STAR v. 2.2.0 (Dobin et al. 2012) and two tools for new transcripts reconstruction: Cufflinks v. 2.1.1 (Trapnell et al. 2010) and Trinity (Grabherr et al. 2011) .

TopHat was used in combination with Cuffilinks, while STAR mapper both with Cufflinks and Trinity.

TopHat is a spliced read mapper that detects splice sites *ab initio* by identifying reads that span exon junctions. The pipeline is divided into two steps: mapping of all reads to the reference genome using Bowtie (Langmead et al. 2009), an ultra-fast short-read mapping program. Then TopHat assembles the mapped reads extracting the sequences and inferring them to be a putative exons while the reads that do not map are set aside (unmapped reads). These reads are afterwards indexed and aligned to potential splice junction that are sequences flanking potential donor/acceptor splice sites within neighbouring regions.

STAR is the RNA-seq aligner used by the ENCODE Project and is designed to align the non-contiguous sequences directly to the reference genome making this software faster than other RNA-seq aligners. Initially STAR searches for each read the maximum mappable length and the matches to the genome create a lot of seeds. If the read comprises a splice junction, the search is repeated for the unmapped portions of the read. The sequential application of the search of maximum read match to the genome only to the unmapped portion of the reads makes STAR extremely fast. Later the software builds alignments of the read sequence clustering the seeds within a genomic window defined. All these seeds are stitched together according to a local alignment scoring scheme and the stitched combination with highest score is chosen as the best alignment of a read.

The number of mapped reads are similar between both aligners for all samples analyzed.

These two tools were used because they map reads over exon/intron junctions, which is a critical feature when aligning RNA-seq reads to a reference genome. Moreover, by improving alignment precision and sensitivity, exon junctions and splicing events are better defined in the reconstruction of new transcripts.

The alignments generated by STAR and TopHat were then considered as input for software that perform identification of new transcripts. Samples belonging to the same population were concatenated into one “population alignment” to improve coverage depth. Cufflinks v. 2.1.1 and Trinity were both evaluated for this purpose. Cufflinks, which uses a mapping-first approach, first aligns all the reads to a reference genome and then merges sequences with overlapping alignment, spanning splice junctions with paired-end reads. To identify a set of novel transcripts expressed in human lymphocyte subsets, a reference annotation is considered to guide the assembly (-g option, RABT assembly) coupled with multi-read (-u option) and fragment bias correction (-b option) to improve the accuracy of transcripts abundance estimates.

The third approach exploits STAR in combination with the genome-guided Trinity software. To address the computational complexity of assembling the human transcriptome by de novo approach, Trinity uses a specifc pipeline named “Genome-guided Trinity” combined with the Program to Assemble Spliced Alignments (PASA). The pipeline has two major steps.

The first uses the “Genome-guided Trinity” where reads are initially aligned to the genome and partitioned according to locus, followed by the “classic” Trinity de novo transcriptome assembly at each locus. In particular, the Trinity default aligner (GSNAP) was substituted with STAR which performs better in terms of accuracy and computing time. The “Genome-guided Trinity” was used with the paramenters suggested in the main documentation and the input alignments were generated using STAR with the default parameters.

The second phase of the pipeline runs PASA having in input all the putative transcripts generated by the first step above. Initially PASA maps transcripts and aligns them to the reference genome; in this case we customized PASA to use START for long reads. STAR required to be customized changing the variables “MAX_READ_LENGTH = 100.000” inside the file “IncludeDefine.h” and recompiled from source code using “make STARlong” which makes available the “COMPILE_FOR_LONG_READS” option. The resulting alignments were validated as nearly perfect with an identity of 95% and percentage of transcript length of about 90% (default PASA’s parameters). The valid transcript alignments are clustered based on genome mapping location and assembled into gene structures; those alignment assemblies which are located in the same locus with a significant overlap and are predicted to be on the same strand are clustered together. Finally, comparing the provided annotation with the clusters, PASA reconstructs the complete transcript and gene structures, resolving incongruencies, refining the reference annotation when there are enough evidences and proposing new transcripts and genes in case any previous annotation can explain the new data.

K- means clustering of gene expression patterns: the Silhouette function

For the clusters presented in this paper K=16 was used for lincRNA genes after optimizing the selection of K to minimize the distances of data within clusters while maximizing the distance between clusters using a Silhouette function (Rousseeuw 1987).

Briefly, K-means clustering was used with different values of K (k=13,14..20..40). For each run, the Silhouette function was calculated on each gene’s expression pattern e^i:

$$Si(e^i) = \frac{b(e^i) - a(e^i)}{\max(a(e^i), b(e^i))}$$

where:

$a(e^i) = E(Dist(e^i, e^j) | e^i \ \in c^x \ and \ e^j \ \in c^x)$, where c^x is the cluster to which e^i was assigned. $a(e^i)$ corresponds to the average dissimilarity between i and all other points of the cluster to which i belongs

and:

$$b(e^i) = min_{co} x E(Dist(e^i, e^j) | e^i not \ \in co^x \ and \ e^j \ \in co^x)$$

$b(e^i)$ can be seen as the dissimilarity between i and its “neighbor” cluster, i.e., the nearest one to which it does *not* belong

The Silhouette graph (shown in Supplementary Figure 1h) reports the optimal number of clusters (bins) that the K-means algorithm needs in order to categorize the dataset in a reliable and reproducible way (when the algorithm reaches convergence). The $S(i)$ function calculates for each datum i (in our case the expression profile of a single gene) the average dissimilarity with all other data within the same cluster, and confronts these results with the lowest average dissimilarity of i (the 'neighbouring cluster') to any other cluster which i is not a member. The final Silhouette score is averaged over all data points in the dataset, and reported in the aforementioned graph (Supplementary Figure 1h).

Specificity score of gene expression patterns: Jensen-Shannon divergence

The clustering results were integrated with an entropy-based methodology that assigns a cell-specificity score to each gene based on Jensen–Shannon divergence (Trapnell et al., 2010).

The JS divergence of two discrete probability distributions $p1$, $p2$, is defined to be:

$$JS(p^1, p^2) = H\left(\frac{p^1 + p^2}{2}\right) - \frac{H(p^1) + H(p^2)}{2}$$

where H is the entropy of a discrete probability distribution:

$p = (p_1, p_2 .. p_n), 0 \leq p_i \leq 1$ and $\sum_{i=1}^{n} p_i = 1$

$$H(p) = -\sum_{i=1}^{n} p_i \log(p_i)$$

Relying on the theorem that the square root of the JS divergence is a metric (Fuglede and Topsoe 2004), the distance between two expression patterns, e^1 and e^2 , $e^i = (e_1^i, .. e_n^i)$, was defined as

$$JS_{dist}(e^1, e^2) = \sqrt{JS(e^1, e^2)}$$

This metric quantifies the similarity between a transcript's expression pattern and another predefined pattern that represent an extreme case in which a transcript is expressed in only one condition. In our case we built a reference model composed of 13 cell subsets. Then, the JS method captures the shape of the distribution and the general trend of expression assigning a gene X to the population for whom it appears to be more specific. The integration of these two approaches has the power to group gene expression profiles according to their cell-specificity.

In order to define a JS score threshold that roughly identifies specifically expressed genes, a log-normal fitting was performed on the JS score density distribution of receptor genes (Supplementary Fig. 1f), that are generally considered the most precise markers of lymphocytes subsets. The metabolic genes density distribution (the non-specific counterpart) is reported as reference.

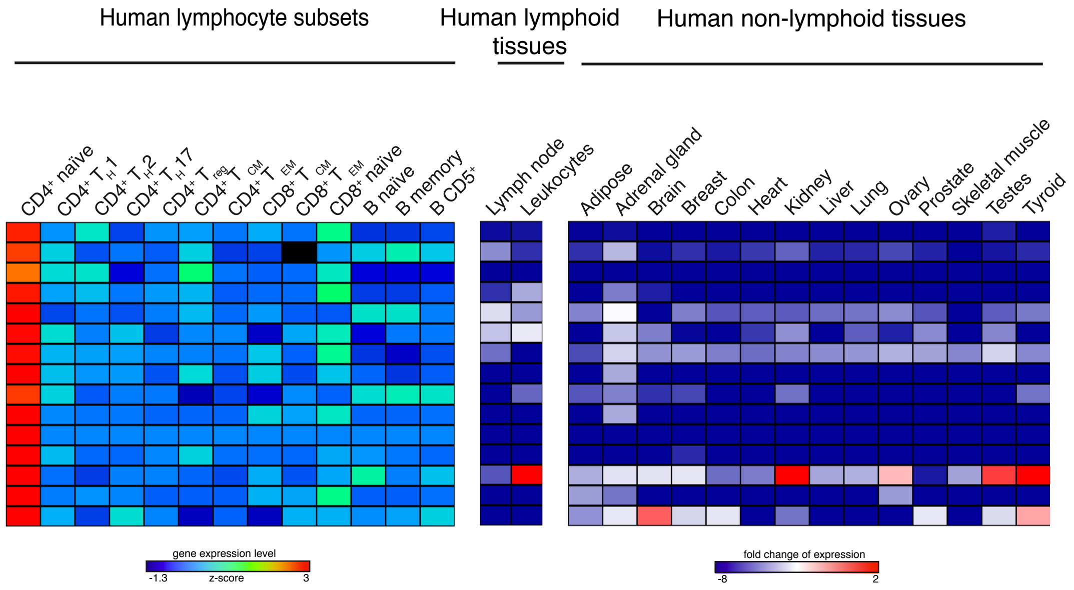
The threshold value for the JS score was calculated by considering one standard deviation away from the mean of the fitted distribution (0.4).

The value corresponding to one standard deviation away (0.4) from the mean of the fitted distribution (0.27) was used as a threshold to define a specific expression.

Table 1

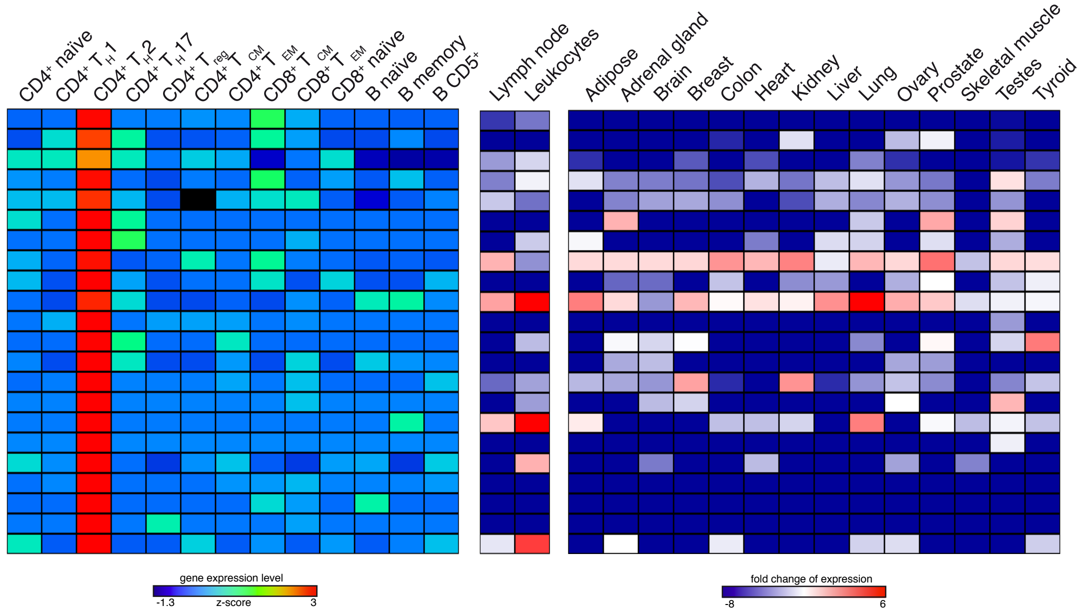
CD4$^+$ naïve signature

gene id	locus	strand	n° isoforms
INGMG_000614	12:53021754-53024658	-	2
ENSG00000262992	17:18932636-18935795	+	1
XLOC_009373	11:11173943-11177996	-	1
INGMG_001448	2:95740576-95742212	+	7
ENSG00000262292	17:19063665-19065046	+	1
ENSG00000254802	8:61878360-61880334	-	1
INGMG_002593	8:27447901-27450875	+	2
INGMG_003003	Y:23173821-23190659	-	2
INGMG_002507	7:2546245-2548666	-	2
INGMG_000615	12:53034890-53038221	-	3
XLOC_004392	5:55354876-55363199	+	1
INGMG_001950	3:59704033-59712944	-	1
XLOC_006012	7:23245631-23247664	+	1
INGMG_001405	2:7512184-7513642	+	1
XLOC_004989	5:126567724-126618000	-	1



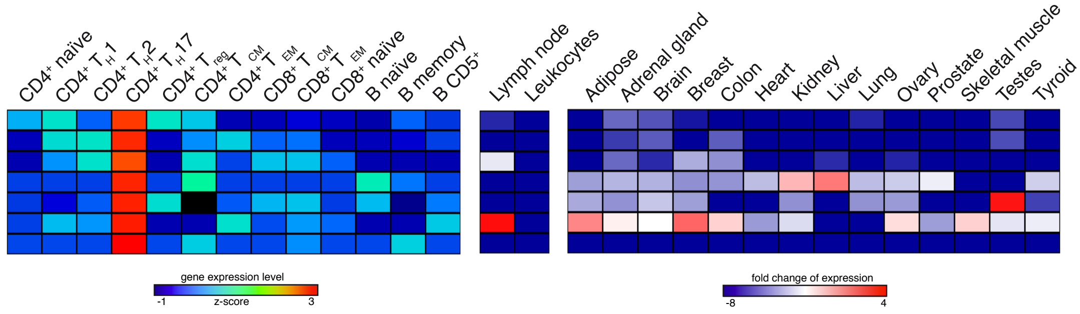
CD4$^+$ T$_H$2 signature

gene id	locus	strand	n° isoforms
INGMG_000354	10:9036884-9061427	-	1
XLOC_008357	10:3985204-4006403	+	2
XLOC_003738	4:153021905-153025384	+	1
XLOC_011681	16:27297150-27301839	+	2
ENSG00000260517	16:29150981-29228027	+	2
XLOC_009457	11:62178649-62179162	-	1
XLOC_009659	12:10393134-10412929	+	1
ENSG00000250786	5:9546311-9550721	+	2
XLOC_011680	16:27280222-27296191	+	2
ENSG00000224397	20:48884022-48896332	+	4
INGMG_000045	1:83243143-83368591	+	1
XLOC_011052	14:65170510-65170923	-	1
ENSG00000254757	11:3490548-3552558	+	1
XLOC_009153	11:63287300-63292203	+	1
XLOC_007934	X:16599799-16601770	+	1
XLOC_007722	9:71158456-71161505	-	2
XLOC_008385	10:8939951-8956559	+	1
XLOC_010236	12:125510477-125513897	-	1
INGMG_000264	1:229114082-229116130	-	1
XLOC_001683	2:136835461-136836083	+	1
XLOC_008383	10:8340859-8343630	+	1
XLOC_009037	11:4415041-4432109	+	1



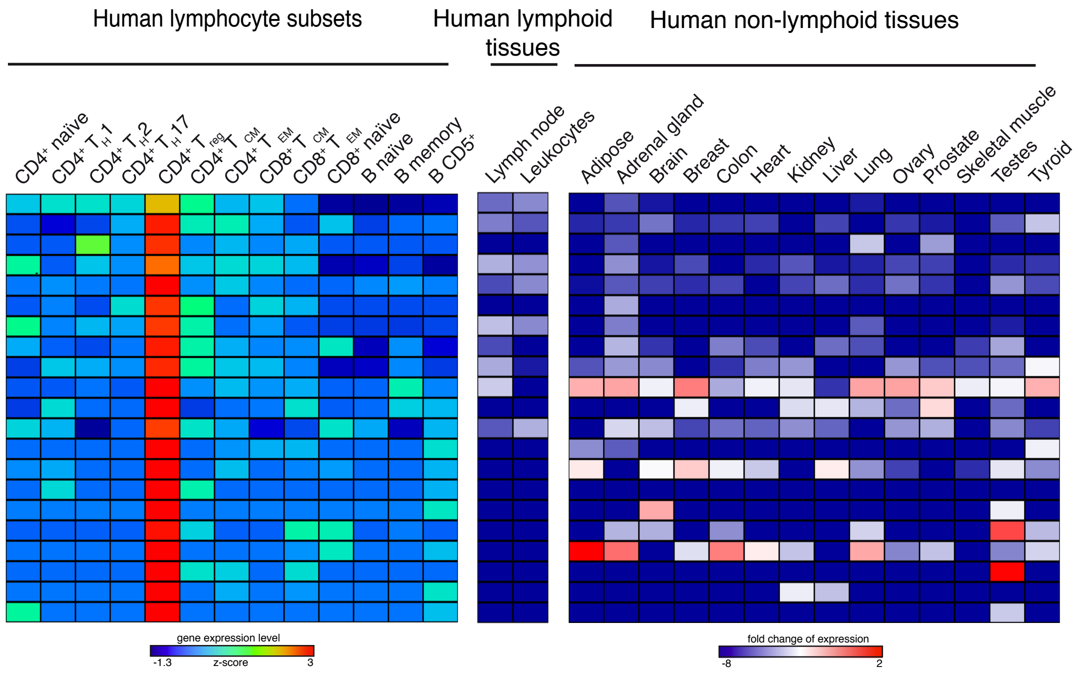
CD4$^+$ T$_H$17 signature

gene id	locus	strand	n° isoforms
INGMG_001733	21:47026508-47032208	+	3
INGMG_001408	2:7797732-7811547	+	10
INGMG_001410	2:7860237-7865579	+	2
XLOC_009027	11:2397410-2398419	+	2
XLOC_002630	3:44465601-44470995	+	5
XLOC_011112	14:95988348-95992377	-	1
ENSG00000260673	6:4599520-4602654	-	1



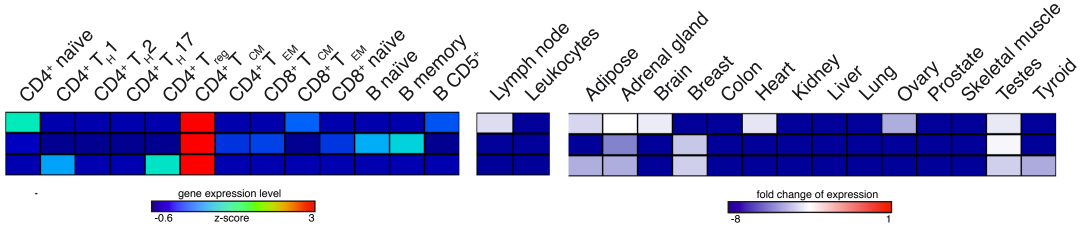
CD4$^+$ T$_{reg}$ signature

gene id	locus	strand	n° isoforms
INGMG_001638	20:21537986-21541942	+	3
INGMG_001237	19:5978323-5980738	+	1
ENSG00000236481	16:26596075-26606134	-	1
ENSG00000253522	5:159895274-159914433	+	1
XLOC_008164	X:49121663-49123331	-	1
INGMG_001500	2:204738800-204762117	+	2
ENSG00000235304	X:39164209-39186616	-	2
INGMG_000762	14:76035364-76039390	+	1
INGMG_001569	2:87538500-87551898	-	6
ENSG00000237697	3:8613467-8615561	+	1
XLOC_003002	3:195869506-195887761	+	4
XLOC_012323	17:76311809-76343879	+	1
XLOC_002477	2:214101740-214103567	-	1
ENSG00000259347	15:67278698-67351591	-	3
XLOC_005276	6:36907862-36912451	+	1
XLOC_010192	12:108646295-108647414	-	1
XLOC_001626	2:112365417-112370095	+	1
XLOC_012881	18:71336694-71358564	-	4
XLOC_003962	4:59646790-59853878	-	7
ENSG00000261729	1:185624133-185626300	+	1
ENSG00000248870	5:81882594-81883230	-	1



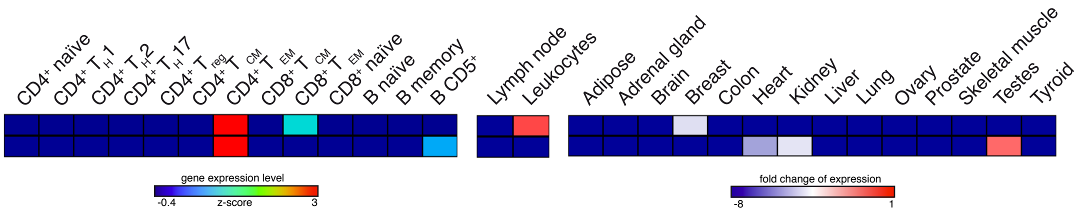
CD4$^+$ T$_{CM}$ signature

gene id	locus	strand	n° isoforms
ENSG00000254538	8:74582675-74645132	+	2
XLOC_013842	20:61775639-61783415	-	1
ENSG00000237899	1:41134760-41153260	-	1



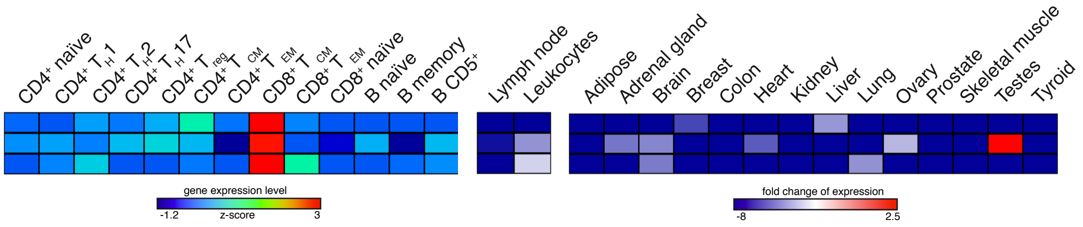
CD4$^+$ T$_{EM}$ signature

gene id	locus	strand	n° isoforms
XLOC_000627	1:235092977-235095736	+	1
XLOC_005870	6:148454944-148458540	-	1



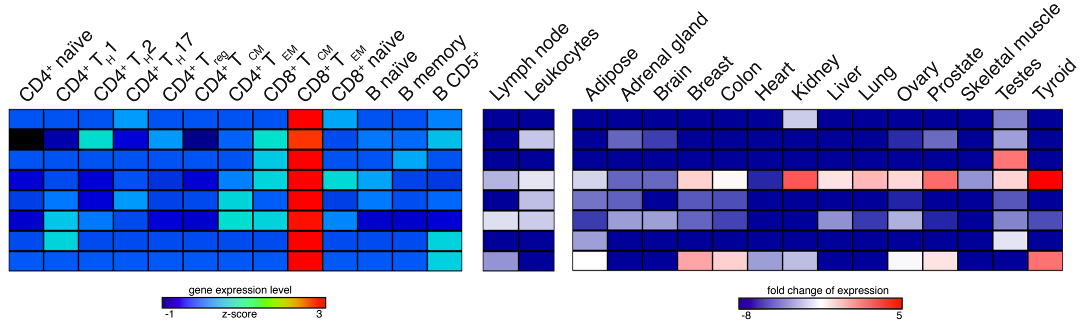
CD8$^+$ T$_{CM}$ signature

gene id	locus	strand	n° isoforms
INGMG_000280	10:8257084-8259668	+	1
XLOC_005737	6:45523579-45545334	-	11
ENSG00000255484	11:112404944-112426525	-	1



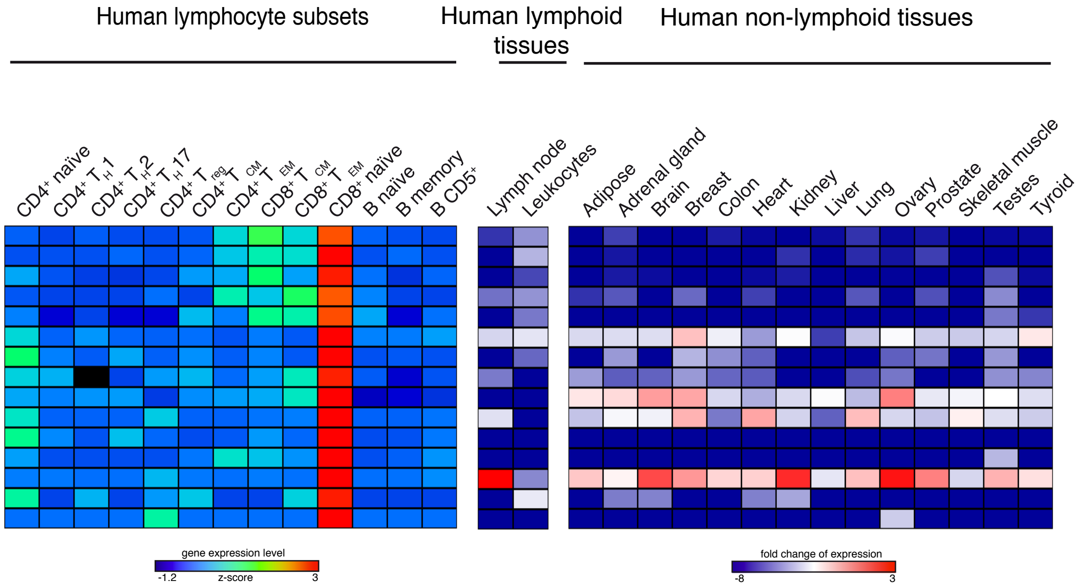
CD8$^+$ T$_{EM}$ signature

gene id	locus	strand	n° isoforms
XLOC_004238	5:524819-526709	+	1
XLOC_001288	1:244393072-244401962	-	1
XLOC_009505	11:75469515-75470461	-	1
XLOC_013703	20:24911283-24913619	-	1
INGMG_002670	8:128677375-128686846	-	3
INGMG_001017	17:34513843-34516804	+	3
ENSG00000254135	5:157912197-157961446	+	2
XLOC_009361	11:2900624-2902339	-	1



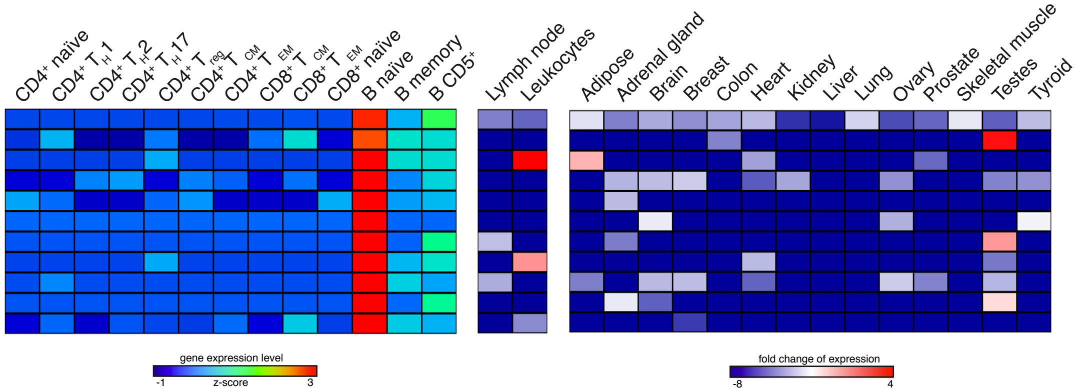
CD8$^+$ naïve signature

gene id	locus	strand	n° isoforms
XLOC_009661	12:10705978-10710816	+	1
XLOC_009662	12:10725616-10727581	+	1
INGMG_000685	13:114920047-114941975	+	5
INGMG_001014	17:34401948-34404160	+	1
XLOC_010517	13:114944062-114944563	+	1
ENSG00000100181	22:17082776-17179521	+	5
XLOC_010859	14:69446486-69448265	+	1
XLOC_013744	20:39763825-39765073	-	1
XLOC_006248	7:130033936-130035446	+	1
ENSG00000256540	12:276021-291565	-	2
INGMG_000819	14:98501239-98503269	-	1
INGMG_000599	12:10652210-10653289	-	1
XLOC_006507	7:79085480-79096779	-	2
INGMG_002390	6:110359651-110361374	-	1
ENSG00000259503	15:70613914-70619081	+	1



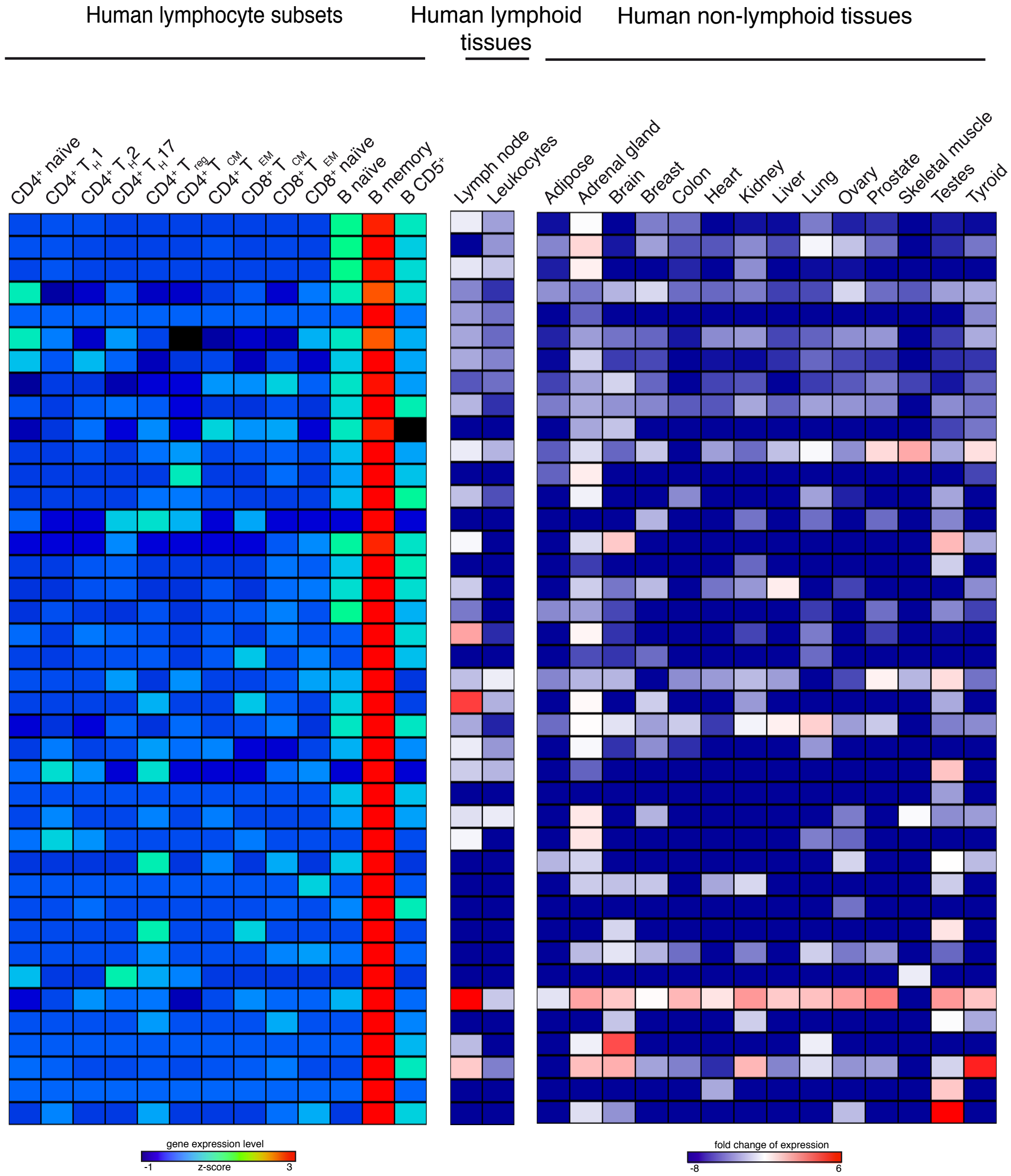
B Naïve signature

gene id	locus	strand	n° isoforms
XLOC_012849	18:53440547-53448952	-	3
XLOC_005155	6:7427115-7453025	+	6
XLOC_011132	14:101586896-101587425	-	1
INGMG_002736	9:99483725-99486063	+	1
ENSG00000256875	12:133038827-133039312	+	1
XLOC_002735	3:98621202-98623886	+	1
XLOC_011265	15:57611128-57617222	+	1
ENSG00000223929	2:60586350-60618510	-	2
XLOC_004483	5:96840399-97006750	+	1
XLOC_000150	1:38940867-38942156	+	1
XLOC_001589	2:100824715-100867946	+	2



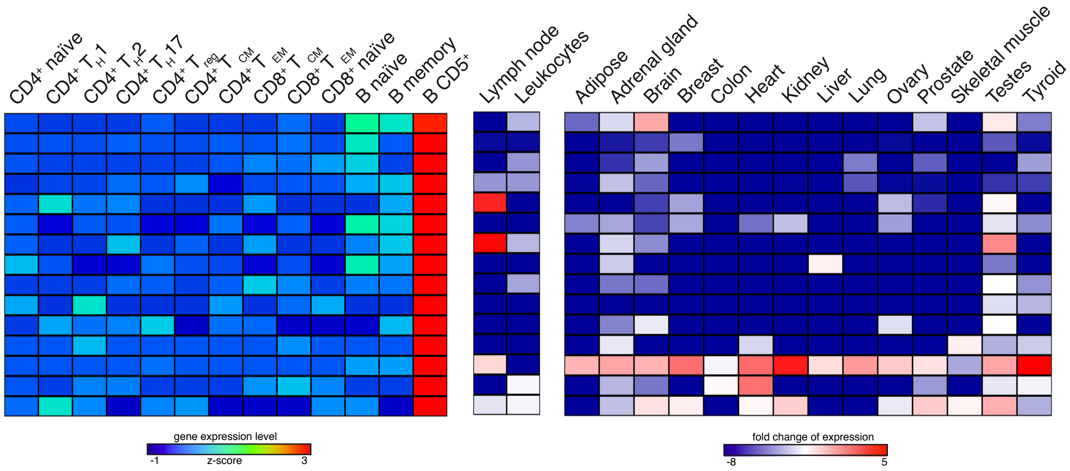
B memory signature

gene id	locus	strand	n° isoforms
ENSG00000253701	14:106170300-106170939	-	1
ENSG00000253364	14:106110832-106115394	-	2
XLOC_000268	1:81001439-81112834	+	4
ENSG00000237438	22:17517459-17539682	+	2
XLOC_002342	2:143628157-143628636	-	1
XLOC_001181	1:207978670-207980881	-	1
XLOC_006293	7:150130741-150145228	+	5
XLOC_007718	9:70501271-70505069	-	1
INGMG_002776	9:14604047-14610947	-	2
XLOC_007388	9:70843566-70844228	+	1
XLOC_005810	6:113943170-113971276	-	10
ENSG00000227468	14:106064027-106066420	-	2
INGMG_000121	1:221250832-221279410	+	2
XLOC_011623	16:2693654-2696114	+	1
XLOC_005264	6:34203397-34204471	+	1
ENSG00000258048	12:80083923-80172231	+	1
XLOC_004625	5:163151151-163158626	+	1
XLOC_009603	11:130086479-130087479	-	1
XLOC_014268	22:46533091-46539488	+	1
INGMG_001754	22:18539268-18555853	+	4
XLOC_008392	10:11715226-11722506	+	1
XLOC_000837	1:53832339-53833917	-	1
ENSG00000203386	2:33931952-34522820	+	2
INGMG_001510	2:226164095-226261637	+	2
ENSG00000260896	16:80862631-80926492	-	4
XLOC_008116	X:13405670-13438072	-	2
XLOC_005811	6:114189182-114194729	-	2
XLOC_011054	14:65708498-65714846	-	1
XLOC_005856	6:139777986-139795737	-	3
XLOC_000835	1:53798052-53812604	-	2
XLOC_001369	2:16704443-16710706	+	2
ENSG00000224565	20:46020672-46041071	-	1
XLOC_002514	2:231450742-231451708	-	1
XLOC_010173	12:102317558-102318599	-	1
ENSG00000242290	3:114172439-114238979	+	1
INGMG_000285	10:10367755-10370619	+	1
XLOC_005372	6:84732773-84734272	+	1
INGMG_002537	7:55424963-55432075	-	1
ENSG00000255595	12:126843847-126845611	+	1
XLOC_007275	9:6704178-6707763	+	1
ENSG00000253686	5:173134616-173173214	-	3
INGMG_001924	3:193508865-193509944	+	1
XLOC_005323	6:52529198-52533951	+	1
XLOC_001882	2:224904214-224907185	+	1
ENSG00000225554	1:241587591-241596792	+	2
ENSG00000261786	3:44158790-44163857	+	1
XLOC_004621	5:159003427-159012901	+	1
XLOC_001866	2:219838968-219844350	+	6



B CD5+ signature

gene id	locus	strand	n° isoforms
XLOC_008554	10:91613272-91674712	+	5
INGMG_002637	8:239189-246382	-	1
INGMG_002250	6:868139-875388	+	2
XLOC_002130	2:65128973-65132271	-	1
XLOC_002613	3:34242414-34310303	+	2
INGMG_000383	10:96871114-96872423	-	1
XLOC_002612	3:34200825-34604551	+	9
CABG_006664	7:155061985-155069592	-	1
INGMG_002582	17:55330358-55332237	-	1
XLOC_006231	8:6639119-6646394	+	1
XLOC_006739	7:124638323-124641124	+	1
ENSG00000256568	9:139155770-139159083	+	1
XLOC_005764	22:18879967-18882205	-	1
ENSG00000234323	6:72117910-72130506	-	4
XLOC_005470	9:109040672-109367076	+	2



Heatmaps of signature lincRNAs expression for each lymphocytes subset. For each lincRNA gene id, locus, strand prediction and number of isoforms are also reported. Right panel represents signature lincRNAs relative expression values in a panel of 16 human tissues (Human BodyMap 2.0 project).

Table 3

VALIDATION PRIMERS

PRIMER ID	TARGET	SEQUENCE (5'-> 3')
linc-MAF-4 F	XLOC_012017	GGCTACGTCTCCATTGTTT
linc-MAF-4 R	XLOC_012017	TGGTGTTTGGGATCATTTGT
T6F	ENSG00000257860	TTTCATGGTGAGGGAGAATGG
T6R	ENSG00000257860	CTGGGTCTTGCCTCTTAATGT
T8F2	INGMG_000772	AGCCTGGGCTTTGGAGTC
T8R3	INGMG_000772	GGCTTTGCCAGGATCTCACA
T21F2	ENSG00000234535	GAAATGCCAATGAAGCAGAAAG
T21R2	ENSG00000234535	GTGCAAAGAATAGGAGGTTTGA
T24F1	XLOC_002906	GTTATCTGTTGCCAGTTGTT
T24R1	XLOC_002906	ACCTCTGCTTATTGCTGATT
T27F1	ENSG00000253988	ACATGGATGCAGCTGGAG
T27R1	ENSG00000253988	TGAGAACATGCCTTTCTTGG
T28F4	XLOC_013498	TACAGCCTCCACCTATTGATT
T28R4	XLOC_013498	ATGGCTTACAGGTAGGAGTTT
T30F3	XLOC_012199	CTGGGTGAACACTGTCTAA
T30R3	XLOC_012199	GCTCAGAGTAAACGGCTAA
T31F1	XLOC_011294	TCGTGTGGGTGAGGAGAA
T31R1	XLOC_011294	AGTGTAGGAGGGCAGTGT

siRNA

siRNA ID	TARGET	SEQUENCE (5'-> 3')
T2_si1	XLOC_012017	GGACCAACCTCTTGTCTTA
T2_si2	XLOC_012017	GTACTGCAAAGGTCTAATA
T2_si3	XLOC_012017	CCGCATACTTTCAGACTTT
T2_si4	XLOC_012017	GCTTGAACTCACAAAGAAA

ChIP PRIMERS

PRIMER ID	TARGET	SEQUENCE (5'-> 3')	REFERENCE
GAS1f	MAF-promoter	TTAAGTGCAGTGCTATAAAGTTGTT	Rani et al., 2011
GAS1r	MAF-promoter	GGGGAAGACCATTCTGAAGTG	Rani et al., 2011
IFNgf	IFNγ-promoter	AAATACCAGCAGCCAGAGGA	
IFNgr	IFNγ-promoter	AGCTGATCAGGTCCAAAGGA	
ILCRf	Internal loop control region	TGAGCAGAGAAAGTGCATAG	
ILCRr	Internal loop control region	TCACAGGCATTCTTTGTACC	
MyoD1f	MyoD1 5' regulatory region	ACGTGCAGATTTAGATGGAG	
MyoD1r	MyoD1 5' regulatory region	ATCGGAGATTGCTGCTAAAG	
ACTBcf	ACTB-promoter	AAAGAGCGAGAGCGAGAT	
ACTBcr	ACTB-promoter	AACGCCAAAACTCTCCCT	

3C PRIMERS

PRIMER ID	SEQUENCE (5'-> 3')
M1	GCAGAACTCGCCTAATGG
L1	TGATTAATGCTGGGTAAAGG
L2	TTCAGCCTTTGTTTTTCTCC
L3	GGTCTTCAATTACAATAGCC
L4	CCAATTGGAAGTCTGAAGGC
L5	ACTGCCCTTCAAGTCCTTGC
L6	ACAGGGAGAGCTGACCTTTG
L7	ATTGAAAGCCATGTTTTTAAG
L8	ACTGCATGGCATTTGTCTGG
L9	CCTTTTTCGCTAGTAGAGCC
L10	TCTCTGGCTGACAGTCTACC
L11	GTACAGCAGCCTCCACAAAG
L12	ATACATATTGGGAGGCCTGGAA
L13	GCTGCAAATCTTGGGATTGG
L14	GCTGAGGTCACAGAGCTAGG
L15	TGCAGGCTCCAAAATAAACC
L16	AGTACAGTAGGCCTCCTTTC
L17	TTTGGGTGTTCTGGGATCTG
L18	TGCCTATGAGTGCTACTGAG
L19	AGGCCCTGCAATATGCACAC
L20	TCCAGCCAGGGCATCCAATC
L21	ACACCCACCAACTTTATTGG
L22	ATAGCGCTGTCTGTGTCTAC
L23	CCCTATCAGCCTGATTTGAG
L24	AGGCCAAACGTAGTGGGTTC

RIP PRIMERS

PRIMER ID	TARGET	SEQUENCE (5'-> 3')	REFERENCE
Actin_sy-F2	β-actin	CATCCTCACCCTGAAGTACC	
Actin_sy-R2	β-actin	CACGCAGCTCATTGTAGAAG	
LincM_pr-F1	linc-MAF-4 (control)	AGGTCATGAGGCAGAGGAGA	
LincM_pr-R1	linc-MAF-4 (control)	TCCCTTTGGGAGGTAAAACC	
HOTAIR/H2-F	HOTAIR/H2	GGTAGAAAAAGCAACCACGAAGC	Tsai et al., 2010
HOTAIR/H2-R	HOTAIR/H2	ACATAAACCTCTGTCTGTGAGTGCC	Tsai et al., 2010

References

1. Zhu J, Yamane H, Paul WE. Differentiation of effector CD4 T cell populations (*). *Annual review of immunology* 2010, **28:** 445-489.

2. Zhou L, Chong MM, Littman DR. Plasticity of CD4+ T cell lineage differentiation. *Immunity* 2009, **30**(5)**:** 646-655.

3. O'Shea JJ, Paul WE. Mechanisms underlying lineage commitment and plasticity of helper CD4+ T cells. *Science* 2010, **327**(5969)**:** 1098-1102.

4. Kanno Y, Vahedi G, Hirahara K, Singleton K, O'Shea JJ. Transcriptional and epigenetic control of T helper cell specification: molecular mechanisms underlying commitment and plasticity. *Annual review of immunology* 2012, **30:** 707-731.

5. O'Connell RM, Rao DS, Chaudhuri AA, Baltimore D. Physiological and pathological roles for microRNAs in the immune system. *Nature reviews Immunology* 2010, **10**(2)**:** 111-122.

6. Pagani M, Rossetti G, Panzeri I, de Candia P, Bonnal RJ, Rossi RL*, et al.* Role of microRNAs and long-non-coding RNAs in CD4(+) T-cell differentiation. *Immunol Rev* 2013, **253**(1)**:** 82-96.

7. Cobb BS, Nesterova TB, Thompson E, Hertweck A, O'Connor E, Godwin J*, et al.* T cell lineage choice and differentiation in the absence of the RNase III enzyme Dicer. *The Journal of experimental medicine* 2005, **201**(9)**:** 1367-1373.

8. Koralov SB, Muljo SA, Galler GR, Krek A, Chakraborty T, Kanellopoulou C*, et al.* Dicer ablation affects antibody diversity and cell survival in the B lymphocyte lineage. *Cell* 2008, **132**(5)**:** 860-874.

9. Li QJ, Chau J, Ebert PJ, Sylvester G, Min H, Liu G*, et al.* miR-181a is an intrinsic modulator of T cell sensitivity and selection. *Cell* 2007, **129**(1)**:** 147-161.

10. O'Connell RM, Kahn D, Gibson WS, Round JL, Scholz RL, Chaudhuri AA*, et al.* MicroRNA-155 promotes autoimmune inflammation by enhancing inflammatory T cell development. *Immunity* 2010, **33**(4)**:** 607-619.

11. Rodriguez A, Vigorito E, Clare S, Warren MV, Couttet P, Soond DR*, et al.* Requirement of bic/microRNA-155 for normal immune function. *Science* 2007, **316**(5824)**:** 608-611.

12. Rossi RL, Rossetti G, Wenandy L, Curti S, Ripamonti A, Bonnal RJ*, et al.* Distinct microRNA signatures in human lymphocyte subsets and enforcement of the naive state in CD4+ T cells by the microRNA miR-125b. *Nature immunology* 2011, **12**(8)**:** 796-803.

13. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A*, et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 2011, **25**(18)**:** 1915-1927.

14. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H*, et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research* 2012, **22**(9)**:** 1775-1789.

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

16. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D*, et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 2009, **458**(7235)**:** 223-227.

17. Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G*, et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 2011, **477**(7364)**:** 295-300.

18. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D*, et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(28)**:** 11667-11672.

19. Yoon JH, Abdelmohsen K, Srikantan S, Yang X, Martindale JL, De S*, et al.* LincRNA-p21 suppresses target mRNA translation. *Molecular cell* 2012, **47**(4)**:** 648-655.

20. Kretz M, Siprashvili Z, Chu C, Webster DE, Zehnder A, Qu K*, et al.* Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* 2013, **493**(7431)**:** 231-235.

21. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 2010, **465**(7301)**:** 1033-1038.

22. Sumazin P, Yang X, Chiu HS, Chung WJ, Iyer A, Llobet-Navas D*, et al.* An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* 2011, **147**(2)**:** 370-381.

23. Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M*, et al.* A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* 2011, **147**(2)**:** 358-369.

24. Pang KC, Dinger ME, Mercer TR, Malquori L, Grimmond SM, Chen W*, et al.* Genome-wide identification of long noncoding RNAs in CD8+ T cells. *J Immunol* 2009, **182**(12)**:** 7738-7748.

25. Collier SP, Collins PL, Williams CL, Boothby MR, Aune TM. Cutting edge: influence of Tmevpg1, a long intergenic noncoding RNA, on the expression of Ifng by Th1 cells. *J Immunol* 2012, **189**(5)**:** 2084-2088.

26. Gomez JA, Wapinski OL, Yang YW, Bureau JF, Gopinath S, Monack DM*, et al.* The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-gamma locus. *Cell* 2013, **152**(4)**:** 743-754.

27. Carpenter S, Aiello D, Atianand MK, Ricci EP, Gandhi P, Hall LL*, et al.* A long noncoding RNA mediates both activation and repression of immune response genes. *Science* 2013, **341**(6147)**:** 789-792.

28. Hu G, Tang Q, Sharma S, Yu F, Escobar TM, Muljo SA*, et al.* Expression and regulation of intergenic long

noncoding RNAs during T cell development and differentiation. *Nature immunology* 2013, **14**(11)**:** 1190-1198.

29. Hart T, Komori HK, LaMere S, Podshivalova K, Salomon DR. Finding the active genes in deep RNA-seq gene expression studies. *BMC genomics* 2013, **14:** 778.

30. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S*, et al.* Ensembl 2013. *Nucleic acids research* 2013, **41**(Database issue)**:** D48-55.

31. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S*, et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013, **29**(1)**:** 15-21.

32. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009, **25**(9)**:** 1105-1111.

33. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ*, et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* 2010, **28**(5)**:** 511-515.

34. Rhind N, Chen Z, Yassour M, Thompson DA, Haas BJ, Habib N*, et al.* Comparative functional genomics of the fission yeasts. *Science* 2011, **332**(6032)**:** 930-936.

35. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE*, et al.* The Pfam protein families database. *Nucleic acids research* 2010, **38**(Database issue)**:** D211-222.

36. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 2011, **27**(13)**:** i275-282.

37. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 2013, **154**(1)**:** 240-251.

38. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS. Specific expression of long noncoding RNAs in the mouse brain. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(2)**:** 716-721.

39. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G*, et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* 2010, **143**(1)**:** 46-58.

40. Al-Shahrour F, Minguez P, Vaquerizas JM, Conde L, Dopazo J. BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic acids research* 2005, **33**(Web Server issue)**:** W460-464.

41. Volders PJ, Helsens K, Wang X, Menten B, Martens L, Gevaert K*, et al.* LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic acids research* 2013, **41**(Database issue)**:** D246-251.

42. Ho IC, Lo D, Glimcher LH. c-maf promotes T helper cell type 2 (Th2) and attenuates Th1 differentiation by both interleukin 4-dependent and -independent

mechanisms. *The Journal of experimental medicine* 1998, **188**(10)**:** 1859-1866.

43. Liu X, Nurieva RI, Dong C. Transcriptional regulation of follicular T-helper (Tfh) cells. *Immunol Rev* 2013, **252**(1)**:** 139-145.

44. Sato K, Miyoshi F, Yokota K, Araki Y, Asanuma Y, Akiyama Y*, et al.* Marked induction of c-Maf protein during Th17 cell differentiation and its implication in memory Th cell development. *The Journal of biological chemistry* 2011, **286**(17)**:** 14963-14971.

45. Mattick JS. The genetic signatures of noncoding RNAs. *PLoS genetics* 2009, **5**(4)**:** e1000459.

46. Klattenhoff CA, Scheuermann JC, Surface LE, Bradley RK, Fields PA, Steinhauser ML*, et al.* Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell* 2013, **152**(3)**:** 570-583.

47. Cabianca DS, Casa V, Bodega B, Xynos A, Ginelli E, Tanaka Y*, et al.* A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. *Cell* 2012, **149**(4)**:** 819-831.

48. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F*, et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 2010, **329**(5992)**:** 689-693.

49. Kaneko S, Bonasio R, Saldana-Meyer R, Yoshida T, Son J, Nishino K*, et al.* Interactions between JARID2 and noncoding RNAs regulate PRC2 recruitment to chromatin. *Molecular cell* 2014, **53**(2)**:** 290-300.

50. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J*, et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* 2013, **8**(8)**:** 1494-1512.

51. Bonnal RJ, Aerts J, Githinji G, Goto N, MacLean D, Miller CA*, et al.* Biogem: an effective tool-based approach for scaling up open source software development in bioinformatics. *Bioinformatics* 2012, **28**(7)**:** 1035-1037.

52. Rousseeuw PJ, Leroy AM, John Wiley & Sons. Robust regression and outlier detection. *Wiley series in probability and mathematical statistics Applied probability and statistics*. New York: Wiley,; 1987.

53. Bodega B, Ramirez GD, Grasser F, Cheli S, Brunelli S, Mora M*, et al.* Remodeling of the chromatin structure of the facioscapulohumeral muscular dystrophy (FSHD) locus and upregulation of FSHD-related gene 1 (FRG1) expression during human myogenic differentiation. *BMC biology* 2009, **7:** 41.

miRiadne: a web tool for consistent integration of miRNA nomenclature

Raoul J. P. Bonnal[1,*,†], Riccardo L. Rossi[1,*,†], Donatella Carpi[1], Valeria Ranzani[1], Sergio Abrignani[1] and Massimiliano Pagani[1,2,*]

[1]Istituto Nazionale Genetica Molecolare 'Romeo ed Enrica Invernizzi', Via F. Sforza 35, 20122 Milan, Italy

[2]Department of Medical Biotechnology and Translational Medicine, Universita` degli Studi di Milano, Via Festa del Perdono 7, 20122 Milano, Italy

[*]To whom correspondence should be addressed.
Email: bonnal@ingm.org
Correspondence may also be addressed to Riccardo L. Rossi. Email: rossi@ingm.org
Correspondence may also be addressed to Massimiliano Pagani Email: pagani@ingm.org

[†]These authors contributed equally to the paper as first authors.

Nucleic Acids Res. 2015 Apr 20. pii: gkv381.

Abstract

The miRBase is the official miRNA repository which keeps the annotation updated on newly discovered miRNAs: it is also used as a reference for the design of miRNA profiling platforms. Nomenclature ambiguities generated by loosely updated platforms and design errors lead to incompatibilities among plat- forms, even from the same vendor. Published miRNA lists are thus generated with different profiling plat- forms that refer to diverse and not updated annotations. This greatly compromises searches, comparisons and analyses that rely on miRNA names only without taking into account the mature sequences, which is particularly critic when such analyses are carried over automatically. In this paper we introduce miRiadne, a web tool to harmonize miRNA nomenclature, which takes into account the original miRBase versions from 10 up to 21, and annotations of 25 common profiling platforms from nine brands that we manually curated. miRiadne uses the miRNA mature sequence to link miRBase versions and/or platforms to prevent nomenclature ambiguities. miRiadne was designed to simplify and support biologists and bioinformaticians in re-annotating their own miRNA lists and/or data sets. As Ariadne helped Theseus in escaping the mythological maze, miRiadne will help the miRNA researcher in escaping the nomenclature maze. miRiadne is freely accessible from the URL http://www.miriadne.org.

Introduction

MicroRNAs (miRNAs) are short endogenous non-coding RNAs varying in length from 19 to 25 nucleotides. They inhibit expression of their cognate target genes by post transcriptional regulation (1), playing a relevant role in a vast array of molecular and cellular processes (2). miRNAs have been shown to be involved in several diseases, such as type 2 diabetes, cardiovascular disease and cancer (3,4), to be key regulators of gene expression and differentiation (5,6) and they have been exploited as circulating biomarkers both in health and disease (7). The first miRNA was identified in *Caenorhabditis elegans* in 1993 (8,9), while the name ‘miRNA’ was introduced almost a decade later in 2001 (10). Subsequently miRNA research gained momentum and an organic naming system was proposed (11), then implemented into a curated and regularly updated repository largely adopted by the scientific community: the miRBase (12). miRBase collects precursor and mature miRNA sequences and is now the authoritative reference for miRNA nomenclature constantly addressing the increasing complexity using and updating suffixes of mature miRNA names. Microarray and RTqPCR-based expression profiling methods of mature miRNAs have become increasingly common in recent years establishing themselves as methods of election for miRNA expression studies. Next-generation sequencing technologies are now pushing forward the discovery of new sequences significantly increasing the number of known miRNAs in each specific organism (13). The correct categorization and collection of mature miRNAs into consistently annotated catalogs is thus an instrumental process for managing and integrating miRNA expression data sets. Unfortunately the fast changing rate of miRNA nomenclature generates confusion in miRNA research that is acknowledged by the scientific community: this problem has been addressed by the development of webservers that allows to browse across versions of the miRBase such as the ‘miRBase Tracker’ (14) and ‘miRSystem’ (15). While the first allows to browse miRNA annotations according to the miRBase, the second is mainly devoted to the characterization of miRNA targets and contains a feature used to con- vert miRNAs names; both of them use the miRBase only as underlying source of heterogeneous annotations. miRBase also provides a list of previous names of its entries, nevertheless none of these resources deal with sequences generated by the profiling platforms. The highly detrimental miRNA nomenclature maze it is not only due to the evolving annotations, but also—and somehow mainly—due to the use of heterogeneously annotated profiling data sets generated by diverse and not updated detection platforms: this often impairs the straightforward integration of different data sets to deliver meta analyses that could be beneficial to miRNA research. On these premises we decided to implement an application specifically devoted to the re-annotation of miRNA signatures and profiling data sets, natively embedding the annotation and probe sequences of 40 common profiling platforms (27 human and 13 rodents) from nine brands integrated with the annotations of miRBase from version 10 to the most recent. Our webserver application, miRiadne, goes beyond browsing miRNA nomenclature across miRBase versions, by offering translation and updating of miRNA namelists and data sets annotations enforcing consistency between miRNA name and their mature sequences. The novelty of our application is the curation of experimental profiling platforms' annotations checking for the consistency between miRNA names used by these platforms and their probes' sequences. Such consistency is a prerequisite that would make meta analysis of miRNA expression data sets easier. While the community is moving toward higher quality standards for miRNA repositories (13) and efforts are made to re-annotate miRNA entries with data from next-generation sequencing experiments (16), heterogeneous and inconsistent miRNA profile nomenclature will remain in the literature affecting future meta analyses; moreover, many detection platforms that adopt different standards are still in use thus generating more inconsistencies. Therefore miRiadne aims at improving interoperability among miRNA data sets and at exploiting all the information contained in already produced miRNA profiling data without any loss of knowledge.

Implementation

The tool was implemented as a web application written in Ruby on Rails. miRBase data from version 10 to 21 were downloaded (in the form of the officially released txt files), from the miRBase FTP server and reorganized in a SQLite database. Probes' Annotation Information Files from 40 different detection platforms of nine vendors (Supplementary Tables S1 and S2 and Supplementary Figures S1 and S2) were retrieved and used in the application as reference for the miRNA name to mature sequence correspondence of each single probe (Supplementary Figure S3). Probes annotation files were stored in the SQLite database. miRiadne is fully compliant with HTML5 and using Twitter Bootstrap framework provides a responsive web site for variable desktop sizes and mobile devices.

Features and inputs

miRiadne application presents two main features: the first one includes four slightly different methods, collectively referred to as ‘Rosetta functions’, and allows the user to trans- late miRNA names from diverse sources into more established and controlled ones: they are the functions ‘Translation, and Overlap’ (in the ‘Rosetta Stone’ menu) and the functions ‘Update and Intersect’ (in the ‘Rosetta Data’ menu). While the first two work on miRNA names only, the third and fourth accept also whole miRNA expression data sets. The second main feature is the ‘Time Warp’ function that allows the user to browse the evolution of miRNA annotations across miRBase versions and serves as a standalone function or as a substantial consistency crosscheck for any miRNA list whose annotation has been translated or converted with the Rosetta functions (Figure 1).

Data can be entered into the miRiadne application in a few different ways: all miRiadne’s functions are accessed by a main ‘input data window’ where miRNA names can be typed one by one or they can be pasted from external documents; either typed or pasted, miRNA names must be separated by a space or a new line. Lists can also be imported directly dragging files containing the query miRNA list into the data window (files being in tab delimited or comma separated values format).

Rosetta Stone—Translation

This feature allows to translate miRNA name lists from a specific platform’s annotation to a different one, from a specific miRBase version to a different one or between a combination of platform and miRBase versions: any of the four possible paired combinations among platforms and miRBases is allowed and can be selected choosing one of the four buttons below the data input window (Supplementary Figure S4). The user can enter an miRNA name or name list as it appears in a given vendor platform (among the plat- forms implemented as shown in Supplementary Tables S1 and S2) or as reported in a list referring to a specific miRBase version, and obtain a translation either (i) into names of a different, 'destination' platform, or (ii) into names ac- cording to a different, usually more recent, miRBase version. After entering the miRNA names in the specific field and choosing one of the available conversions, a contextual menu appears offering the user the possibility to specify sources and destination (i.e. from which specific profiling platform to which miRBase version); then a simple click to the 'convert' button results in the elaboration and production of the Result Table. Submission of such a query will result in a converted list of names, with graphical evidence for miRNAs that cannot be translated due to different reasons (a different background for dead, changed or not present miRNAs or for miRNAs with multiple names). The Result Table can also be downloaded as csv files.

Rosetta Stone—Overlap

The 'Overlap' function does not need any input query file because it is used to find and highlight the miRNAs in common between two different detection platforms (i.e. present on both platforms with the same mature sequence regardless names variations), between two miRBase versions or between platforms and miRBase versions. The first case (Platform versus Platform) can be used to evaluate the potential detection ability of different platforms (Supplementary Figure S5), the second case (Platform versus miRBase) is useful to evaluate the miRBase coverage of different detection platforms; finally the third case (miRBase versus miRBase) shows which miRNAs are conserved, in terms of sequence conservation between two different miRBase versions even after any annotation change. This function directly interrogates the integrated database populated by the miRBase versions from 10 to the most recent and the annotations files from the detection platforms considered.

Rosetta Data—Update

This function is a particular form of annotation translation in which annotation of miRNA data matrices (i.e. miRNA lists with expression values for a number of samples in tabular format) can be updated to a more recent version of the miRBase (Platform to miRBase), or converted to the annotation of a different detection platform (Platform to Platform), carrying over the corresponding data (expression values) from the original data set (Supplementary Figure S6). This function can be used to convert miRNA profiling

data sets previously generated with different platforms or referring to different miRBase versions prior to a new analysis or just for comparison with more updated data sets. The input data should be in tabular form, where rows are relative to miRNAs and columns to samples: thus the first column contains the miRNA list used as query list, the first row (the header) reports the samples' id and all other columns the expression data that will be carried over in the 'Data' column of the Result Table.

Annotations (miRNA names) of the original data set are reported in the Result Table side by side with the updated annotations (name and mature sequence) and the corresponding expression values in the original data set are carried over in the 'Data' column. The updated data matrix can be downloaded as a csv file, downloaded data values can be displayed in separate columns too, thus allowing further elaboration and/or analyses.

In case input query miRNA names cannot be translated or are not present any more in the updated annotation set: a warning is displayed. A detailed list of miRNAs that can-not be found in the destination annotation (either a more recent miRBase version or a different platform) is reported shaded in red at the bottom of the table. The single hits that may have changed, dropped out or associated with multiple names are highlighted with a yellow (missing, dead or changes) or blue (multiple names) shading in the table. Non- miRNA annotations such as those from control probes, non miRNA-probes or endogenous controls (i.e. any annotation which is by definition not included in the miRBase) can- not be recognized and thus will trigger the above mentioned warning.

Rosetta Data—Intersection

This feature allows finding common miRNAs from two profile human data sets (data matrices with miRNA expression values/rows in a number of samples/columns) regardless any change that could have occurred to the miRNA names. This function is useful when one needs to obtain a new data set from two human miRNA data sets generated by two different profiling platforms which contains common miRNAs only. The newly generated data set will be populated with all the miRNAs present in both original data sets according to the mature sequence (i.e. that have the same sequence), thus allowing to perform downstream analysis without the risk of missing any information (Supplementary Figure S7). This feature also allows the carryover of the expression values (only of the miRNAs shared between the two data sets) and the input format of the query data set tables is the same as the one previously described for the 'Update' function.

Time Warp

This is the other main function of the miRiadne application: it allows to track miRNA names evolution (name changes, survival and/or drop outs) through the miRBase versions from number 10 to the most current one. Name changes, new miRNA entries and miRNAs deleted from the miRBase are taken into consideration. Mature sequence for each entry is always appended.

User can input one or more miRNA names, directly typing them or uploading (drag and drop method) a text file with the names list. The search can be started from a selected miRBase version, or, alternatively, the default search that spans all miRBases in miRiadne (from version 10 to the latest miRBase) can be performed. Searches can be stringent (default) or relaxed: in the first case search will retrieve only miRNAs with exactly the same name as typed, in the second case all miRNA names containing the search string will be retrieved. The *Homo sapiens* is set as the default species but if the species is not selected the search will be performed on the entire miRBase and results will comprise all species in the miRBase. As an additional function it is also possible to display the profiling platforms implemented in miRiadne that can detect the searched miRNA(s). Matching miRNA(s) are displayed in a result table with relative IDs (MI and MIMAT), names, sequences, strand and length: a graphical indicator shows if that miRNA is stable, if it was deleted, at which miRBase version the searched miRNA was introduced (the name is introduced) or a name change occurred (new name displayed) (Supplementary Figure S8).

Discussion

Detection methods based on hybridization technologies such as microarrays or RTqPCR that exploit designed probes to detect mature miRNA expression are widely used and their probesets refer to and are designed on different miRBase releases. Unfortunately, miRNA probe lists in the detection platforms sometime are not explicitly associated with probe sequences and often the miRBase unique identifier (the so-called MIMAT id) is often not reported. Moreover, a substantial portion of publications on human miRNAs (more than half, as shown in a survey of 100 recent papers (14)) relies on miRNA names without disclosing sequences of investigated miRNAs, thus contributing to annotation uncertainty in the literature. Other tools have partially overlapping functions with parts of miRiadne: for instance the miRBase Tracker (14) is redundant with the miRiadne's Time Warp function and miRBase itself provides a list of previous names of miRNA sequences. Nevertheless, miRiadne offers a novel profiling platform-centric approach which protects from the confusion generated by names used in publications and experimental resources that have become out-of-date with the official names in the current version of the miRBase. While the miRBase is regularly updated and miRNA sequences can be renamed, or even deleted or added, the commercial probe- sets do not undergo a regular re-annotation (exceptions exist) thus they are usually not updated. The result is high heterogeneity in names usage when referring to list of expressed, differential expressed or selected miRNAs. Lower costs and availability of high-throughput technologies in the miRNA profiling field generated so far a body of literature suffering from a relevant degree of nomenclature inconsistencies; moreover, the lack of standard guidelines for data profiling hampers the possibility to take advantage from the wide variety of tissues and conditions in which miRNAs expression has been investigated. miRBase recently began an effort of additional annotation producing a high confidence side version of the miRBase using next-generation sequencing data (13); this is an admirable intent but, as stated by miRBase curators themselves, miRNA names remain 'entirely unsuitable to encode information about complex sequence relationships' (17).

Thus evolving nomenclature leads to inconsistencies, and as a consequence, name-driven comparisons of miRNA lists (such as signatures) from diverse sources (papers, databases) can be risky, and higher meta-analyses virtually impossible. As an example, we extracted the miRNA signature from two recent papers about miRNAs carried by human exosomes (18,19) and compared them both ‘as they are’ originally reported and ‘after translation’ using the miRiadne ‘Intersect’ function. The two works used different platforms and, as expected, slightly different experimental conditions: nonetheless if miRNAs detected both in exosomes (18) and in HDL- associated exosomes (19) were searched for, performing a name based overlap analysis between the two original signatures, discrepancies not merely due to annotations would be found. The use of miRiadne allows to perform the signatures’ comparison much faster, with sequence consistency and uncovering four name changes and one retired miRNA (Figure 2); such differences would not have been detected by a mere manual comparison of miRNA names. Even subtler changes can take place across miRBase versions, deriving from immediate replacement of dead entries when the biogenesis process is discovered and needs to be corrected; such a case is the one of hsa-miR-453 that being processed from the 5p arm of miR-323b from miRBase 15 onward became hsa-miR-325b-5p. Such changes would not be a problem for researchers focused on a few or a single miRNA, which can be individually followed in its evolution by the Time Warp function of miRiadne or other applications (14,15) but it is a major bottleneck when doing screening and automated data processing of profiling data: the use of a tool such as miRiadne will alleviate the difficulty in performing meta analyses with miRNA expression profiling data.

Acknowledgments

We wish to thank members of the INGM labs who internally used the miRiadne prototype application in their miRNA research helping us to debug it and to improve its features.

Figure and Table Legends

Figure 1. miRiadne main features. The five miRiadne's features grouped in two main classes: the first class includes the four Rosetta features used to translate miRNA names (Rosetta Stone functions) and data sets (Rosetta Data functions) produced by human miRNAs detection platforms: they are the Translation, Overlap, Intersection and Update functions. The second main feature is the Time Warp that allows one to browse any miRNA name or name lists evolution across miRBase versions, in any species.

Figure 2. Name-driven miRNA list comparison. (**a**) Overlap of miRNA signatures from two different papers (18,19). Overlap between miRNA names as they are reported in the papers (Venn diagram on the left) shows 16 commonly detected miRNAs, while overlap between miRNA signatures with updated annotations (Venn diagram on the right) shows 15 common miRNAs. Further overlap between the two commonly detected miRNAs uncovers that only 11 miRNAs are not affected by inconsistent annotations between the two papers and are correctly included in the comparison in both cases, and one miRNA (hsa-miR-923, underlined) is not present in the updated list. (**b**) The 15 miRNAs commonly detected by the two works in the result table obtained with the 'Rosetta Intersect' function. Names are automatically updated to latest miRBase version, mature sequences appended and retired miRNAs (such as hsa-miR-923) highlighted and excluded. Destiny of retired miRNAs can be further investigated with the Time Warp function.

Figure 1

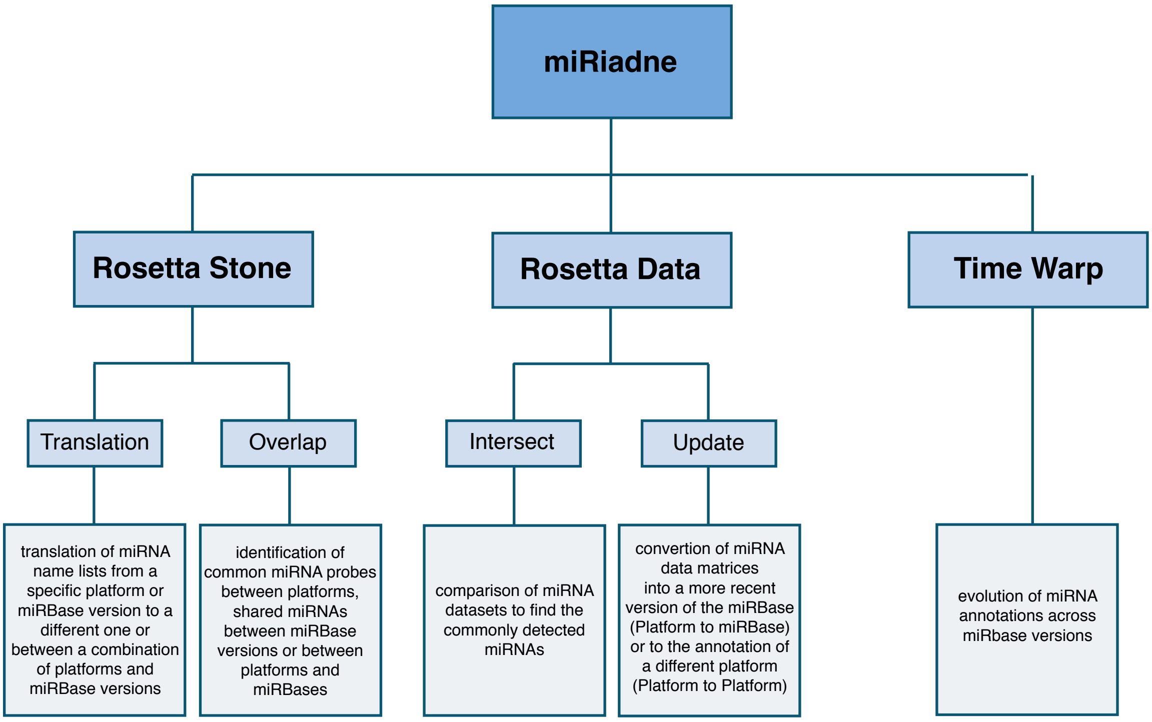
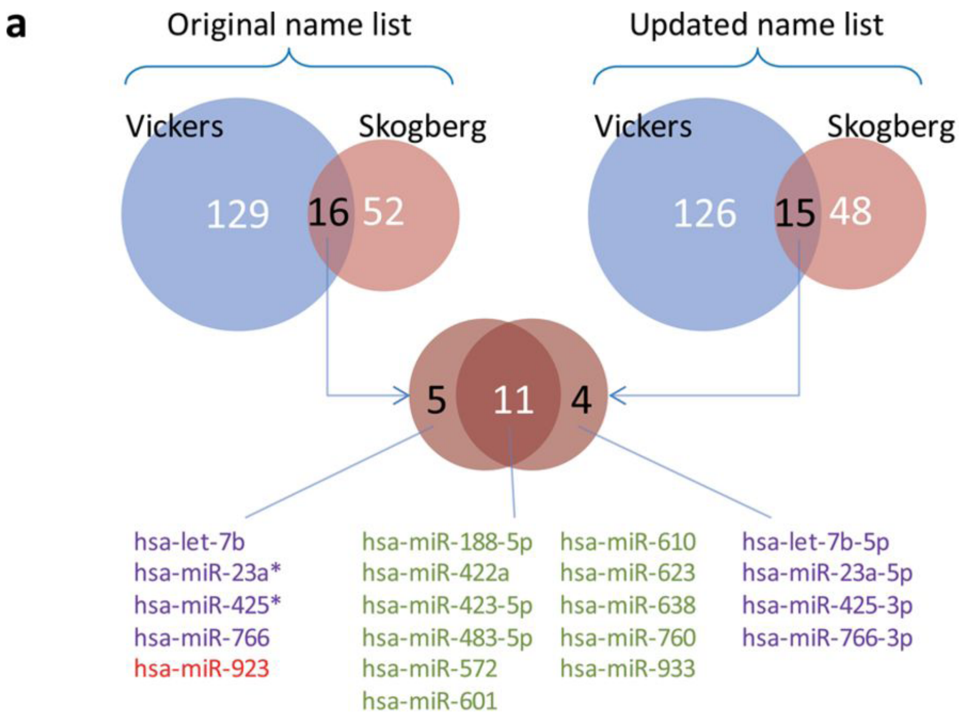
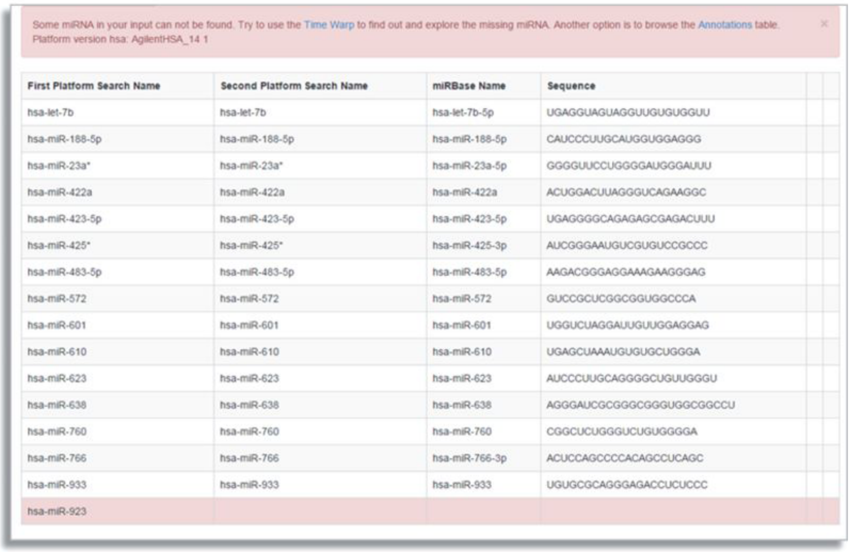

miRiadne
Rosetta Stone
Rosetta Data
Time Warp
Translation
Overlap
Intersect
Update
translation of miRNA name lists from a specific platform or miRBase version to a different one or between a combination of platforms and miRBase versions
identification of common miRNA probes between platforms, shared miRNAs between miRBase versions or between platforms and miRBases
comparison of miRNA datasets to find the commonly detected miRNAs
convertion of miRNA data matrices into a more recent version of the miRBase (Platform to miRBase) or to the annotation of a different platform (Platform to Platform)
evolution of miRNA annotations across miRbase versions

Figure 2



b



Some miRNA in your input can not be found. Try to use the Time Warp to find out and explore the missing miRNA. Another option is to browse the Annotations table.
Platform version hsa: AgilentHSA_14 1

First Platform Search Name	Second Platform Search Name	miRBase Name	Sequence
hsa-let-7b	hsa-let-7b	hsa-let-7b-5p	UGAGGUAGUAGGUUGUGUGGUU
hsa-miR-188-5p	hsa-miR-188-5p	hsa-miR-188-5p	CAUCCCUUGCAUGGUGGAGGG
hsa-miR-23a*	hsa-miR-23a*	hsa-miR-23a-5p	GGGGUUCCUGGGGAUGGGAUUU
hsa-miR-422a	hsa-miR-422a	hsa-miR-422a	ACUGGACUUAGGGUCAGAAGGC
hsa-miR-423-5p	hsa-miR-423-5p	hsa-miR-423-5p	UGAGGGGCAGAGAGCGAGACUUU
hsa-miR-425*	hsa-miR-425*	hsa-miR-425-3p	AUCGGGAAUGUCGUGUCCGCCC
hsa-miR-483-5p	hsa-miR-483-5p	hsa-miR-483-5p	AAGACGGGAGGAAAGAAGGGAG
hsa-miR-572	hsa-miR-572	hsa-miR-572	GUCCGCUCGGCGGUGGCCCA
hsa-miR-601	hsa-miR-601	hsa-miR-601	UGGUCUAGGAUUGUUGGAGGAG
hsa-miR-610	hsa-miR-610	hsa-miR-610	UGAGCUAAAUGUGUGCUGGGA
hsa-miR-623	hsa-miR-623	hsa-miR-623	AUCCCUUGCAGGGGCUGUUGGGU
hsa-miR-638	hsa-miR-638	hsa-miR-638	AGGGAUCGCGGGCGGGUGGCGGCCU
hsa-miR-760	hsa-miR-760	hsa-miR-760	CGGCUCUGGGUCUGUGGGGA
hsa-miR-766	hsa-miR-766	hsa-miR-766-3p	ACUCCAGCCCCACAGCCUCAGC
hsa-miR-933	hsa-miR-933	hsa-miR-933	UGUGCGCAGGGAGACCUCUCCC
hsa-miR-923			

Supplementary Material for

miRiadne: a web tool for consistent integration of miRNA nomenclature

SUPPLEMENTARY FIGURE AND TABLE CAPTIONS

Supplementary table 1. Human profiling platforms. Table of detection profiling platforms whose annotation information files were considered and that are used by the miRiadne translation engine. Columns of the table show, from left to right: brand and commercial name of platform, platform version (release), technology used (microarray, RTqPCR or beads counter); miRBase version to which the platform's annotations refer to, number of miRNAs detectable by the platform; platforms that disclose miRNA probes sequences, platform that do not disclose them (or for which a sequence retrieval from miRBase was necessary).

Supplementary figure 1. Number of human miRNAs per platform. Histogram showing number of human miRNAs detected by the 27 platforms taken into account by miRiadne. Control probes were excluded from the count.

Supplementary table 2. Rodent profiling platforms. Table of detection profiling platforms whose annotation information files were considered and that are used by the miRiadne translation engine. Columns of the table show, from left to right: brand and commercial name of platform, platform version (release), technology used (microarray, RTqPCR or beads counter); miRBase version to which the platform's annotations refer to, number of miRNAs detectable by the platform; platforms that disclose miRNA probes sequences, platform that do not disclose them (or for which a sequence retrieval from miRBase was necessary).

Supplementary figure 2. Number of rodent miRNAs per platform. Histogram showing number of mouse and rat miRNAs detected by the of the 13 platforms taken into account by miRiadne. Control probes were excluded from the count.

Supplementary figure 3. AIFs validation procedure. The figure shows the validation process of AIFs (Annotation Information Files, the probes' annotations supplied with each detection platform) before their inclusion into miRiadne database. If the probe/miRNA sequences were originally included in the supplied AIF they were taken into account and stored with miRNA names; if they were not, mature miRNA sequences were exrtacted from the miRBase version whose the specific platform refers to, then subjected by a manual curation to find any inconsistence between probe names and miRBase miRNA names.

Supplementary figure 4. Rosetta Stone - Translation. How to use the Rosetta translation function: select the function from menu (1-2); import the data either by typing it or dragging and dropping in the window a txt file (tab delimited or comma separated value) with the miRNA names list (3); select one of the four possible translation combination, in the screenshot the "from platform to miRBase" is selected (4); select the starting platform and the destination miRBase from the corresponding menus (5); run the conversion (6); review the result table (7). The results display the miRNA names searched, the names in the starting and destination queries (before and after translation), the alternative name (if present in the detection platform's annotation) and the mature sequence. Two cases in which a single miRNA has been annotated with more than one name (8) and one case of a retired miRNA (9) are shown (the result table in the screenshot is truncated).

Supplementary figure 5. Rosetta Stone - Overlap. How to use the Rosetta overlap function: this function does not need any input file since it is used to see overlap between platforms' probes and/or miRBase versions. Three options are available, overlap between different platforms, between different miRBase versions and between specific platforms and miRBase versions. Select the function from menu (1-2); choose one of the three options (3), select the two terms of comparison, either platforms and/or miRBases according to the option chosen (4-5); run the job (6) and review the result table (7). The result table contains all and only those miRNAs that, according to mature sequence identity, are contained in both terms of comparison.

Supplementary figure 6. Rosetta Data - Update. How to use the Rosetta update function: select the function from menu (1-2); import the data, the easier and safer way is to drag and drop in the window a txt file containing the miRNA expression dataset (tab delimited or comma separated value) with the miRNA names list in the first column (3); choose one of the two possible options, either to update the dataset from the generating platform to a specific miRBase version or to the annotation of another detection platform (4); specify the generating platform on the left and the landing translation on the right (5); run the translation job (6) and review the result table (7). The result table contains all and only those miRNAs that are present in the landing platform or miRBase (i.e. those miRNAs that can be updated given the chosen parameters), the mature sequence is displayed and the original data matrix is appended in the "Data" column. The result table can be downloaded (8) as a comma separated value file and quickly imported in spreadsheet format, the data column which contains space delimited data is also carried over and easily restored (9).

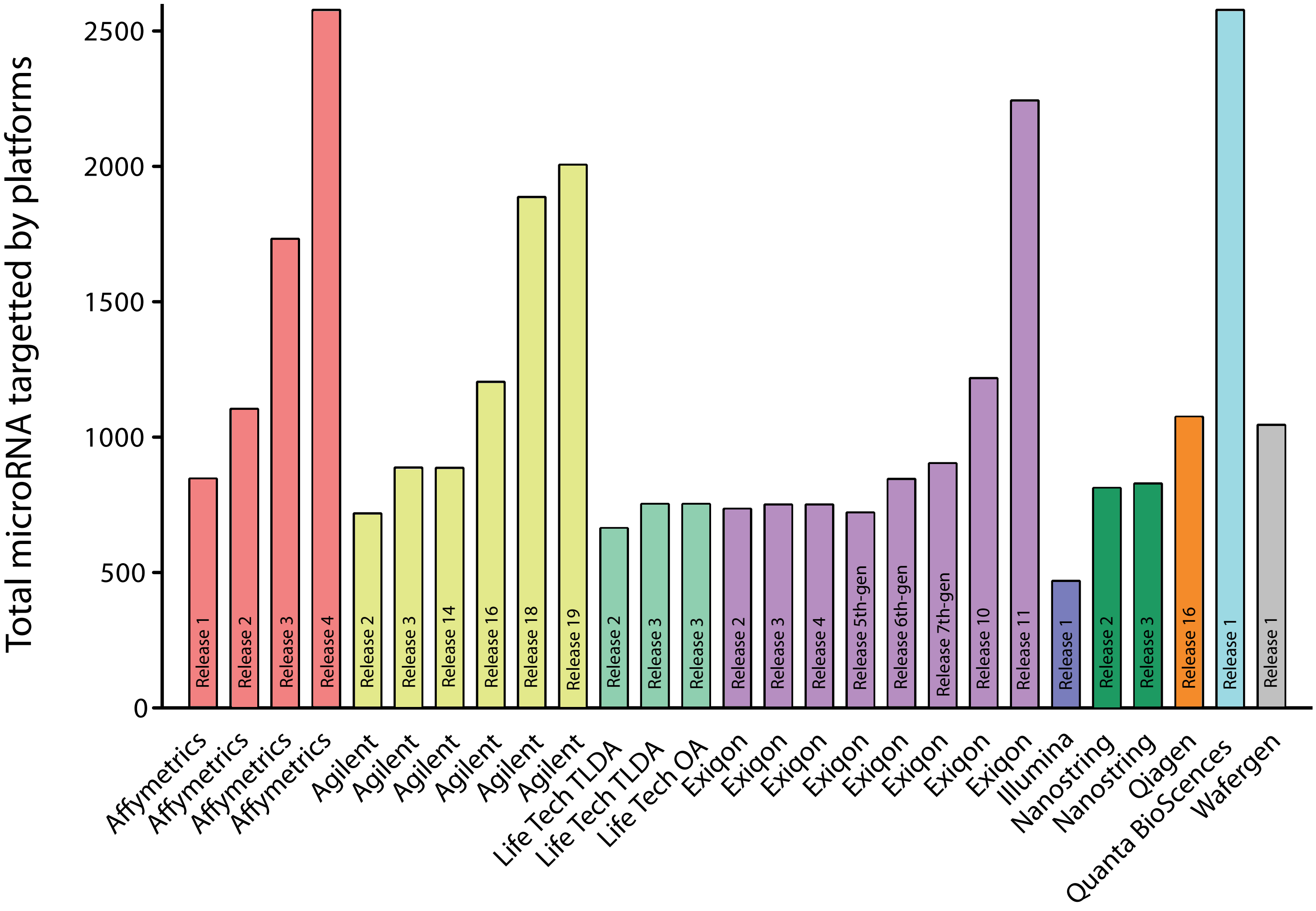
Supplementary figure 7. Rosetta Data - Intersect. How to use the Rosetta intersect function: select the function from menu (1-2); import the data either by typing it or dragging and dropping in the two windows a txt file (tab delimited or comma separated value) with the miRNA names list: a separate input window is used for the two miRNA lists or datasets that are going to be intersected. In the case of profiling datasets (with expression values) or long miRNA lists the "drag and drop" method is the easier and safer (3-4); specify the first (5) and the second (6) platform, then submit the job (7) and review the result table (8). miRNAs present in both submitted datasets or lists are displayed in the result table, their names are updated to the ltest miRBase version and mature sequence is appended. Those miRNAs in the input that cannot be found are highlighted at the bottom of the table and a warning messade is display at the top of the table.

Supplementary figure 8. Time Warp. How to use the Time Warp function: select the function from menu (1); decide whether to perform a stringent search (default) or select the "Relax search" option to search for partial or incomplete matches; you can also decide to display the platforms able to detect the searched miRNAs, if any (2); the data can be imported either by typing it directly or dragging and dropping in the window a txt file (tab delimited or comma separated value) with the miRNA names list. For short miRNA lists, miRNA names can be typed either separated by a new line (carriage return) or separated by a comma (3). Select the miRNA version to be considered as the one where to start the analysis from, the default span is all miRBases in miRiadne, i.e. from version 10 to the latest (4); select the species, the default is Homo sapiens (5): if the "blank" option is explicitly selected, search is performed over all species in miRBase, and "Relax search" should be selected too in order to facilitate matching; run the job clicking "Search" (6) and review the result table (7). The result table displays searched miRNAs with their IDs, sequence, strand positioning and length. If no specific miRBase versions were selected the table display results from miRBase 10 to the latest and for each version a colored dot is used to indicate stability of the annotation (green dot) or retirement of the miRNA (red dot). Name changes are indicated with the new miRNA name in place of the dot. If a specific miRBase version was selected (miRBase 14 as shown in the screenshot) (8) Time Warp will start analysis from that version number and any miRNA occasionally retired before that version will not displayed (as it is the case of hsa-miR-801 in the figure) (9).

Supplementary Table 1

	Platform name	Release	Tecnology	miRBase	# miRNA	Seq by vendor	Seq by miRBase
Affymetrics	GeneChip miRNA Array	1	microarray	11	847	x	
	GeneChip miRNA Array	2	microarray	15	1105	x	
	GeneChip miRNA Array	3	microarray	17	1733	x	
	GeneChip miRNA Array	4	microarray	20	2578	x	
Agilent	Human miRNA Microarray	2	microarray	10.1	719		x
	Human miRNA Microarray	3	microarray	12	888		x
	Human miRNA Microarray	14	microarray	14	887		x
	Human miRNA Microarray	16	microarray	16	1205		x
	Human miRNA Microarray	18	microarray	18	1887		x
	Human miRNA Microarray	19	microarray	19	2006		x
LifeTech (Applied Biosystems)	TaqMan Array Human microRNA	2	RT-qPCR	10	664	x	x
	TaqMan Array Human microRNA	3	RT-qPCR	14	754	x	
	OpenArray Human microRNA	3	RT-qPCR	14	754		x
Exiqon	microRNA Ready-to-Use PCR panels	2	RT-qPCR	16	736	x	
	microRNA Ready-to-Use PCR panels	3	RT-qPCR	18	752	x	
	microRNA Ready-to-Use PCR panels	4	RT-qPCR	20	752	x	
	miRCURY LNA microRNA Array	10	microarray	10	722	x	
	miRCURY LNA microRNA Array	11	microarray	11	846	x	
	miRCURY LNA microRNA Array	5th gen	microarray	14	904	x	
	miRCURY LNA microRNA Array	6th gen	microarray	16	1218	x	
	miRCURY LNA microRNA Array	7th gen	microarray	20	2244	x	
Illumina	Human microRNA expression panel	1	microarray	9	470		x
Nanostring	nCounter Human miRNA panel	2	microarray	18	812	x	
	Nanostring NCounter Human	3	counter	21	829	x	
Qiagen	miScript miRNA PCR Arrays	16	RT-qPCR	16	1075		x
Quanta BioSciences	qScript microRNA System	1	microarray	20	2578	x	
Wafergen	SmartChip Human MicroRNA Panel	1	microarray	16	1046		x

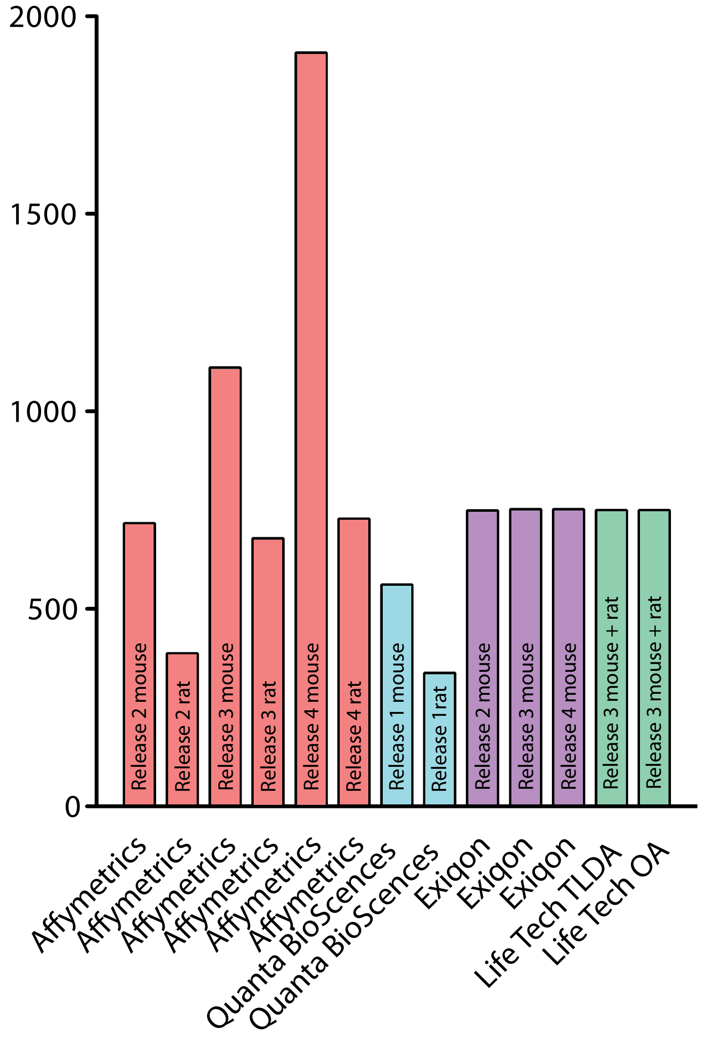
Supplementary Figure 1



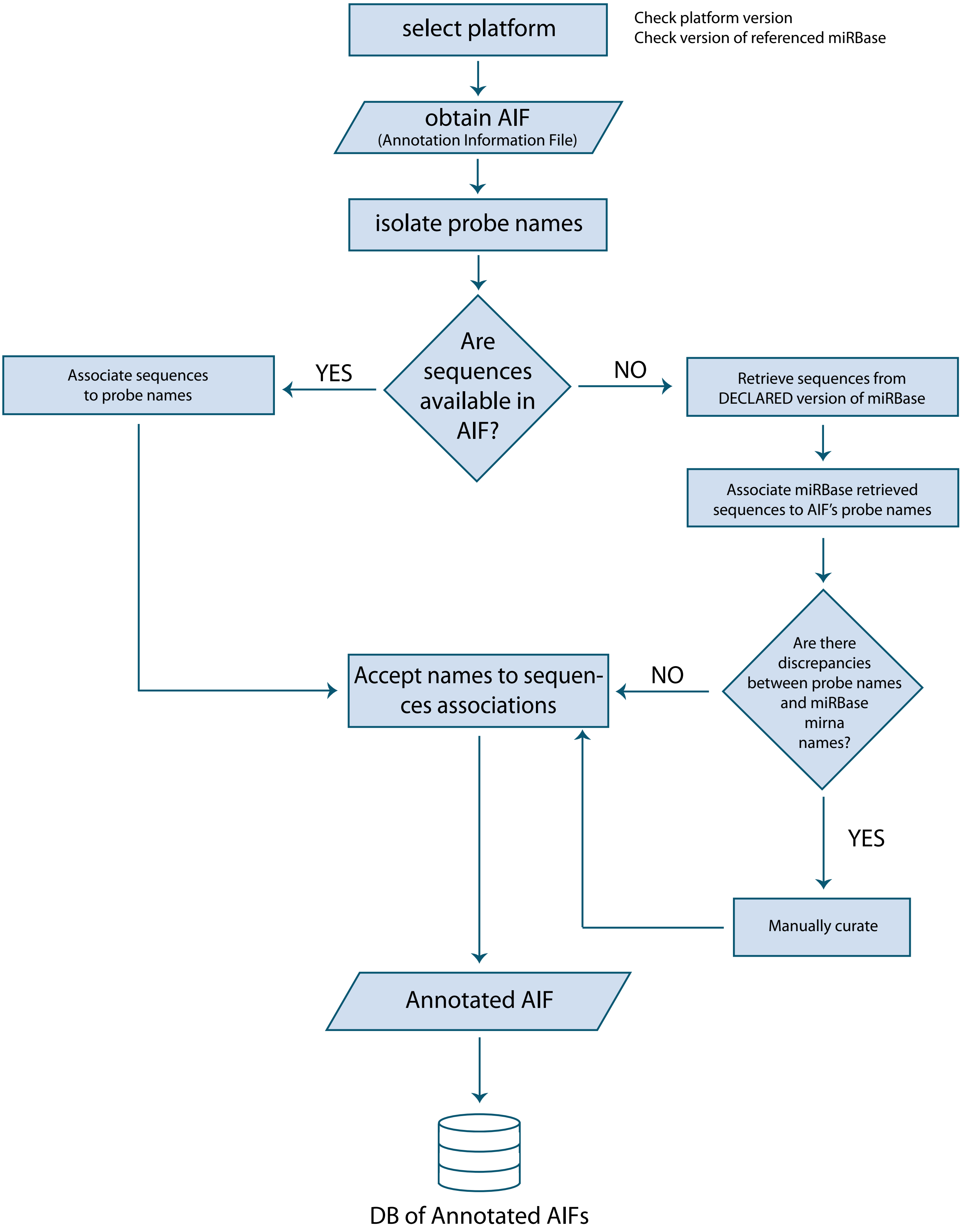
Supplementary Table 2

	Platform name	Release	Tecnology	miRBase	# miRNA	Seq by vendor	Seq by miRBase
Affymetrics	GeneChip miRNA Array mouse	2	microarray	15	717	x	
	GeneChip miRNA Array rat	2	microarray	15	387	x	
	GeneChip miRNA Array mouse	3	microarray	17	1111	x	
	GeneChip miRNA Array rat	3	microarray	17	679	x	
	GeneChip miRNA Array mouse	4	microarray	20	1908	x	
	GeneChip miRNA Array rat	4	microarray	20	728	x	
Quanta BioSciences	qScript microRNA System mouse	1	microarray	20	561		x
	qScript microRNA System rat	1	microarray	20	338		x
Exiqon	microRNA Ready-to-Use PCR panels Mouse	2	RT-qPCR	16	749	x	
	microRNA Ready-to-Use PCR panels Mouse	3	RT-qPCR	18	752	x	
	microRNA Ready-to-Use PCR panels Mouse	4	RT-qPCR	20	752	x	
LifeTech (Applied Biosystems)	TaqMan Array Rodent microRNA	3	RT-qPCR	14	750		x
	OpenArray Rodent microRNA	3	RT-qPCR	14	750		x

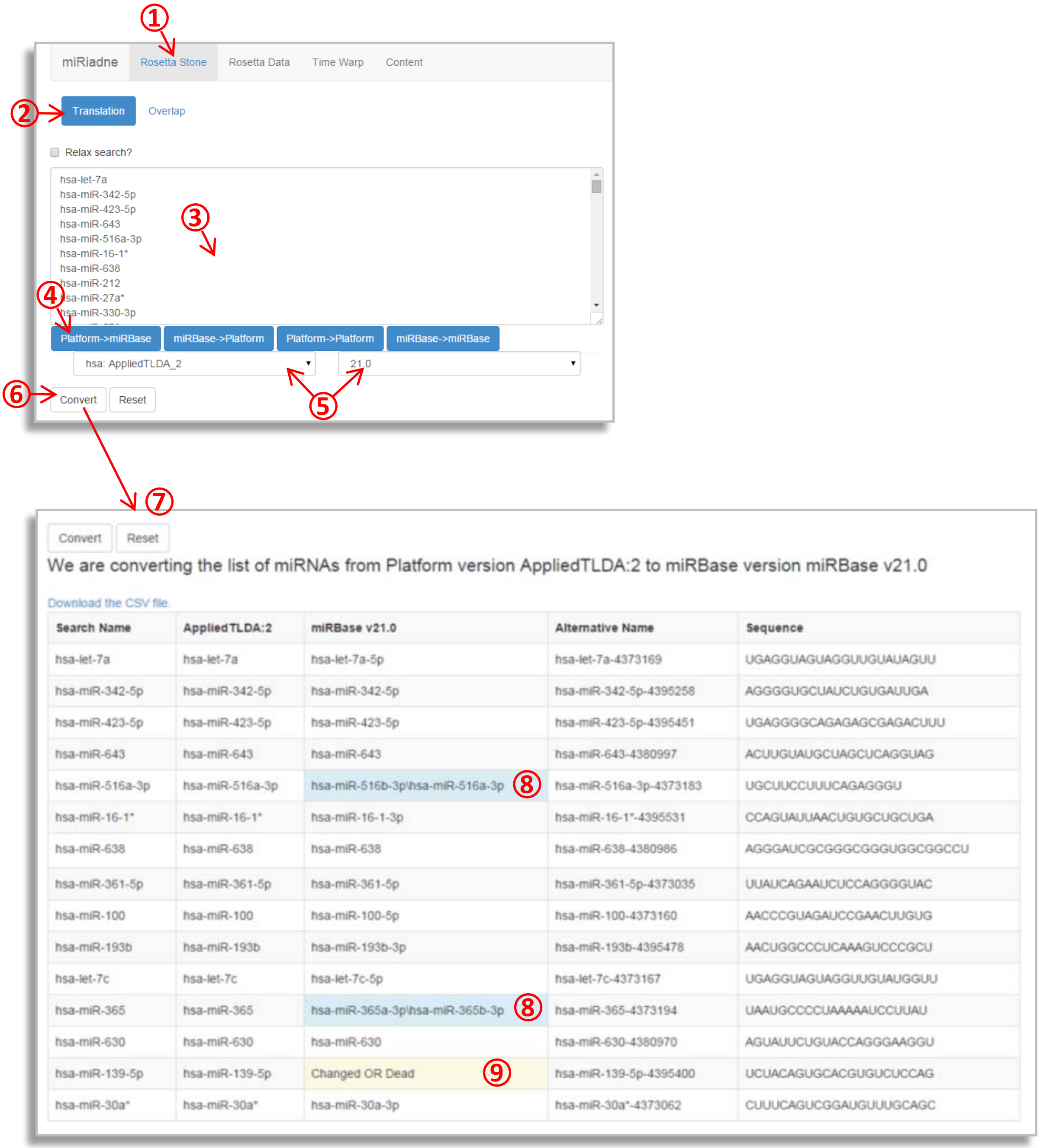
Supplementary Figure 2



Supplementary Figure 3


select platform
Check platform version
Check version of referenced miRBase
obtain AIF
(Annotation Information File)
isolate probe names
Are sequences available in AIF?
YES
Associate sequences to probe names
NO
Retrieve sequences from DECLARED version of miRBase
Associate miRBase retrieved sequences to AIF's probe names
Are there discrepancies between probe names and miRBase mirna names?
NO
Accept names to sequences associations
YES
Manually curate
Annotated AIF
DB of Annotated AIFs

Supplementary Figure 4



□ Relax search?

hsa-let-7a
hsa-miR-342-5p
hsa-miR-423-5p
hsa-miR-643
hsa-miR-516a-3p
hsa-miR-16-1*
hsa-miR-638
hsa-miR-212
hsa-miR-27a*
hsa-miR-330-3p

hsa: AppliedTLDA_2 ▾

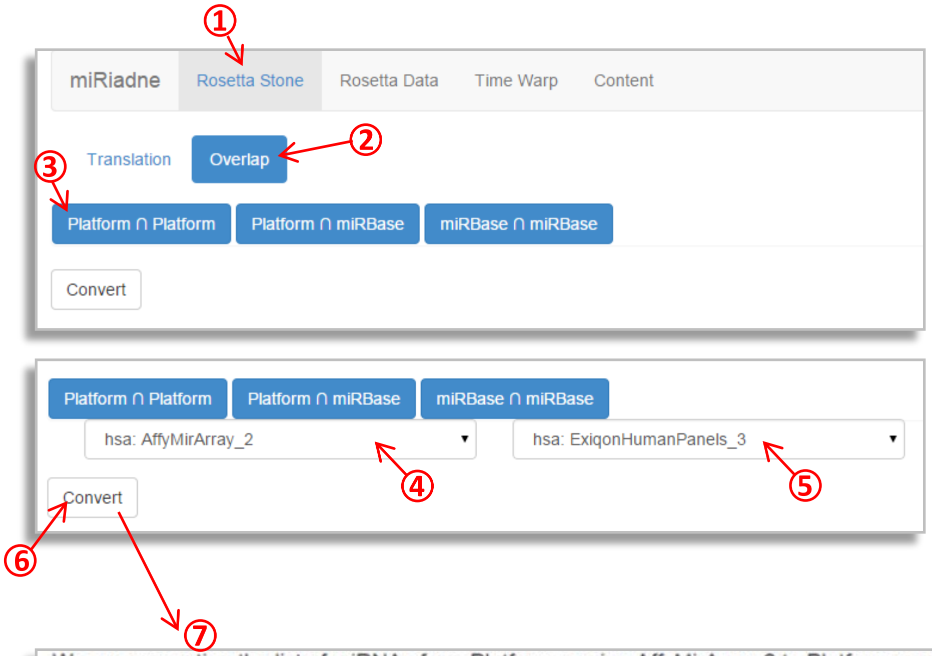
21.0 ▾

We are converting the list of miRNAs from Platform version AppliedTLDA:2 to miRBase version miRBase v21.0

Download the CSV file.

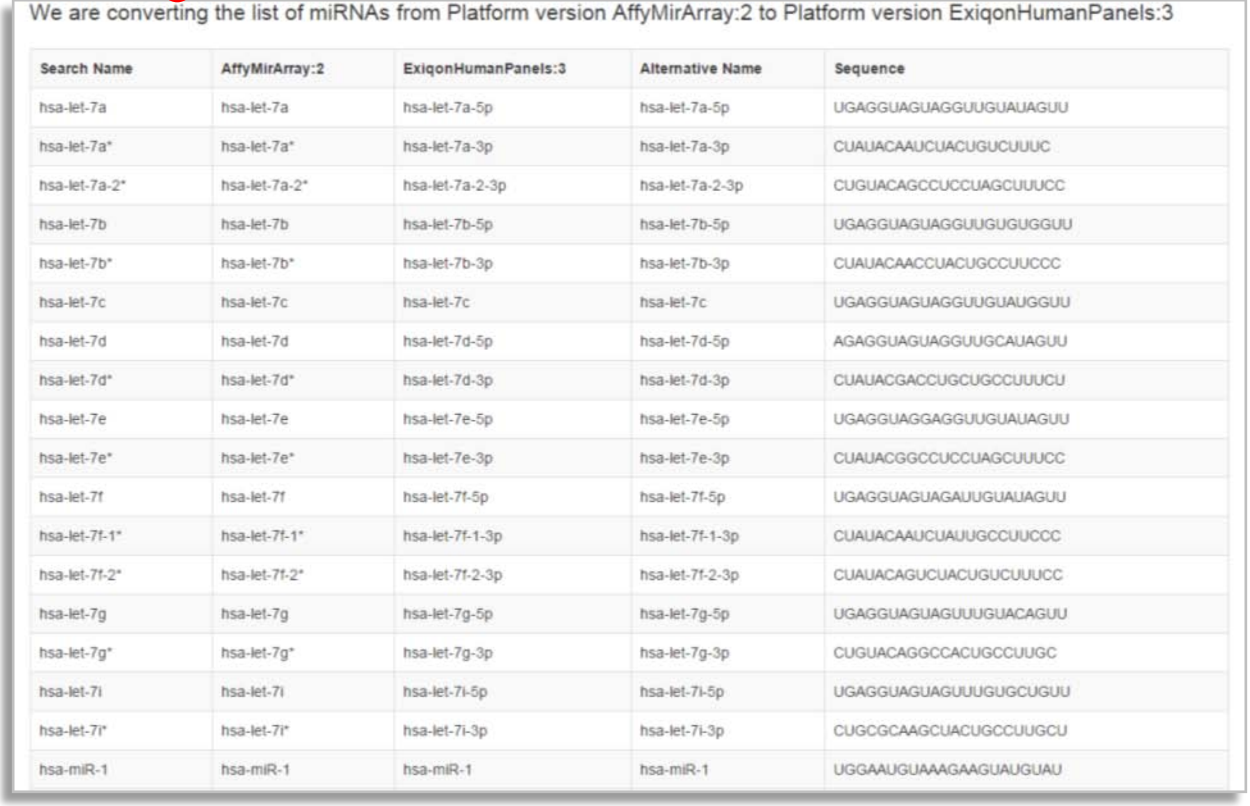
Search Name	AppliedTLDA:2	miRBase v21.0	Alternative Name	Sequence
hsa-let-7a	hsa-let-7a	hsa-let-7a-5p	hsa-let-7a-4373169	UGAGGUAGUAGGUUGUAUAGUU
hsa-miR-342-5p	hsa-miR-342-5p	hsa-miR-342-5p	hsa-miR-342-5p-4395258	AGGGGUGCUAUCUGUGAUUGA
hsa-miR-423-5p	hsa-miR-423-5p	hsa-miR-423-5p	hsa-miR-423-5p-4395451	UGAGGGGCAGAGAGCGAGACUUU
hsa-miR-643	hsa-miR-643	hsa-miR-643	hsa-miR-643-4380997	ACUUGUAUGCUAGCUCAGGUAG
hsa-miR-516a-3p	hsa-miR-516a-3p	hsa-miR-516b-3p\hsa-miR-516a-3p	hsa-miR-516a-3p-4373183	UGCUUCCUUUCAGAGGGU
hsa-miR-16-1*	hsa-miR-16-1*	hsa-miR-16-1-3p	hsa-miR-16-1*-4395531	CCAGUAUUAACUGUGCUGCUGA
hsa-miR-638	hsa-miR-638	hsa-miR-638	hsa-miR-638-4380986	AGGGAUCGCGGGCGGGUGGCGGCCU
hsa-miR-361-5p	hsa-miR-361-5p	hsa-miR-361-5p	hsa-miR-361-5p-4373035	UUAUCAGAAUCUCCAGGGGUAC
hsa-miR-100	hsa-miR-100	hsa-miR-100-5p	hsa-miR-100-4373160	AACCCGUAGAUCCGAACUUGUG
hsa-miR-193b	hsa-miR-193b	hsa-miR-193b-3p	hsa-miR-193b-4395478	AACUGGCCCUCAAAGUCCCGCU
hsa-let-7c	hsa-let-7c	hsa-let-7c-5p	hsa-let-7c-4373167	UGAGGUAGUAGGUUGUAUGGUU
hsa-miR-365	hsa-miR-365	hsa-miR-365a-3p\hsa-miR-365b-3p	hsa-miR-365-4373194	UAAUGCCCCUAAAAAUCCUUAU
hsa-miR-630	hsa-miR-630	hsa-miR-630	hsa-miR-630-4380970	AGUAUUCUGUACCAGGGAAGGU
hsa-miR-139-5p	hsa-miR-139-5p	Changed OR Dead	hsa-miR-139-5p-4395400	UCUACAGUGCACGUGUCUCCAG
hsa-miR-30a*	hsa-miR-30a*	hsa-miR-30a-3p	hsa-miR-30a*-4373062	CUUUCAGUCGGAUGUUUGCAGC

Supplementary Figure 5



hsa: AffyMirArray_2

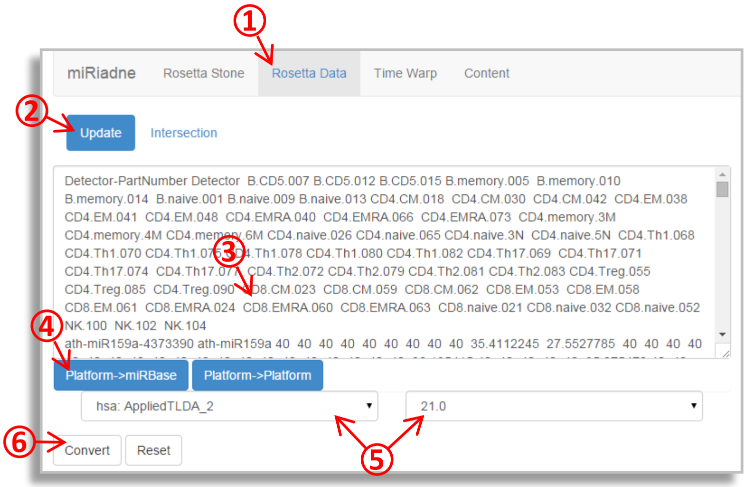
hsa: ExiqonHumanPanels_3



We are converting the list of miRNAs from Platform version AffyMirArray:2 to Platform version ExiqonHumanPanels:3

Search Name	AffyMirArray:2	ExiqonHumanPanels:3	Alternative Name	Sequence
hsa-let-7a	hsa-let-7a	hsa-let-7a-5p	hsa-let-7a-5p	UGAGGUAGUAGGUUGUAUAGUU
hsa-let-7a*	hsa-let-7a*	hsa-let-7a-3p	hsa-let-7a-3p	CUAUACAAUCUACUGUCUUUC
hsa-let-7a-2*	hsa-let-7a-2*	hsa-let-7a-2-3p	hsa-let-7a-2-3p	CUGUACAGCCUCCUAGCUUUCC
hsa-let-7b	hsa-let-7b	hsa-let-7b-5p	hsa-let-7b-5p	UGAGGUAGUAGGUUGUGUGGUU
hsa-let-7b*	hsa-let-7b*	hsa-let-7b-3p	hsa-let-7b-3p	CUAUACAACCUACUGCCUUCCC
hsa-let-7c	hsa-let-7c	hsa-let-7c	hsa-let-7c	UGAGGUAGUAGGUUGUAUGGUU
hsa-let-7d	hsa-let-7d	hsa-let-7d-5p	hsa-let-7d-5p	AGAGGUAGUAGGUUGCAUAGUU
hsa-let-7d*	hsa-let-7d*	hsa-let-7d-3p	hsa-let-7d-3p	CUAUACGACCUGCUGCCUUUCU
hsa-let-7e	hsa-let-7e	hsa-let-7e-5p	hsa-let-7e-5p	UGAGGUAGGAGGUUGUAUAGUU
hsa-let-7e*	hsa-let-7e*	hsa-let-7e-3p	hsa-let-7e-3p	CUAUACGGCCUCCUAGCUUUCC
hsa-let-7f	hsa-let-7f	hsa-let-7f-5p	hsa-let-7f-5p	UGAGGUAGUAGAUUGUAUAGUU
hsa-let-7f-1*	hsa-let-7f-1*	hsa-let-7f-1-3p	hsa-let-7f-1-3p	CUAUACAAUCUAUUGCCUUCCC
hsa-let-7f-2*	hsa-let-7f-2*	hsa-let-7f-2-3p	hsa-let-7f-2-3p	CUAUACAGUCUACUGUCUUUCC
hsa-let-7g	hsa-let-7g	hsa-let-7g-5p	hsa-let-7g-5p	UGAGGUAGUAGUUUGUACAGUU
hsa-let-7g*	hsa-let-7g*	hsa-let-7g-3p	hsa-let-7g-3p	CUGUACAGGCCACUGCCUUGC
hsa-let-7i	hsa-let-7i	hsa-let-7i-5p	hsa-let-7i-5p	UGAGGUAGUAGUUUGUGCUGUU
hsa-let-7i*	hsa-let-7i*	hsa-let-7i-3p	hsa-let-7i-3p	CUGCGCAAGCUACUGCCUUGCU
hsa-miR-1	hsa-miR-1	hsa-miR-1	hsa-miR-1	UGGAAUGUAAAGAAGUAUGUAU

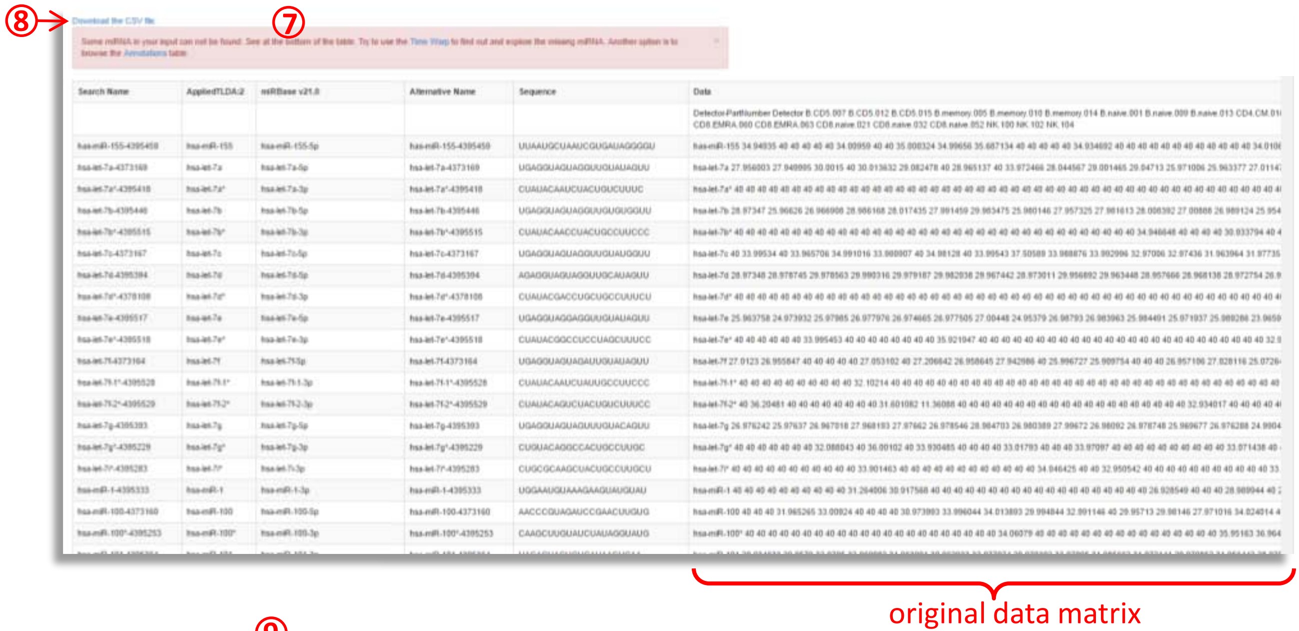
Supplementary Figure 6



Detector-PartNumber Detector B.CD5.007 B.CD5.012 B.CD5.015 B.memory.005 B.memory.010 B.memory.014 B.naive.001 B.naive.009 B.naive.013 CD4.CM.018 CD4.CM.030 CD4.CM.042 CD4.EM.038 CD4.EM.041 CD4.EM.048 CD4.EMRA.040 CD4.EMRA.066 CD4.EMRA.073 CD4.memory.3M CD4.memory.4M CD4.memory.6M CD4.naive.026 CD4.naive.065 CD4.naive.3N CD4.naive.5N CD4.Th1.068 CD4.Th1.070 CD4.Th1.075 CD4.Th1.078 CD4.Th1.080 CD4.Th1.082 CD4.Th17.069 CD4.Th17.071 CD4.Th17.074 CD4.Th17.077 CD4.Th2.072 CD4.Th2.079 CD4.Th2.081 CD4.Th2.083 CD4.Treg.055 CD4.Treg.085 CD4.Treg.090 CD8.CM.023 CD8.CM.059 CD8.CM.062 CD8.EM.053 CD8.EM.058 CD8.EM.061 CD8.EMRA.024 CD8.EMRA.060 CD8.EMRA.063 CD8.naive.021 CD8.naive.032 CD8.naive.052 NK.100 NK.102 NK.104
ath-miR159a-4373390 ath-miR159a 40 40 40 40 40 40 40 40 40 35.4112245 27.5527785 40 40 40 40

hsa: AppliedTLDA_2 ▾

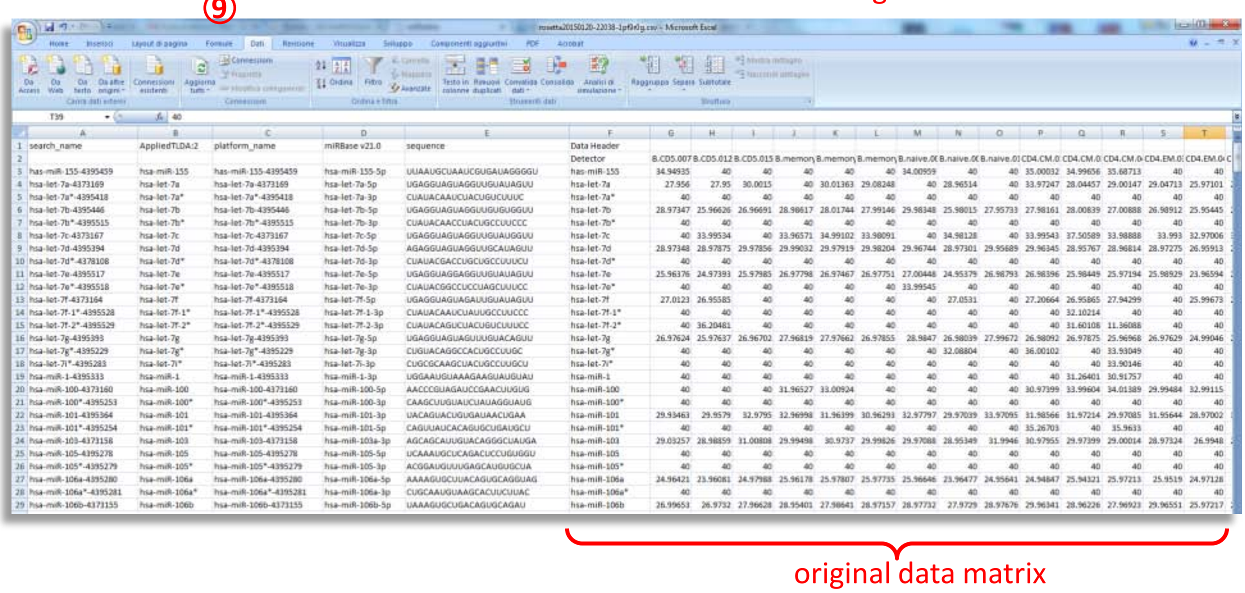
21.0 ▾



Download the CSV file

Some miRNA in your input can not be found. See at the bottom of the table. Try to use the Time Warp to find out and explore the missing miRNA. Another option is to browse the Annotations table ×

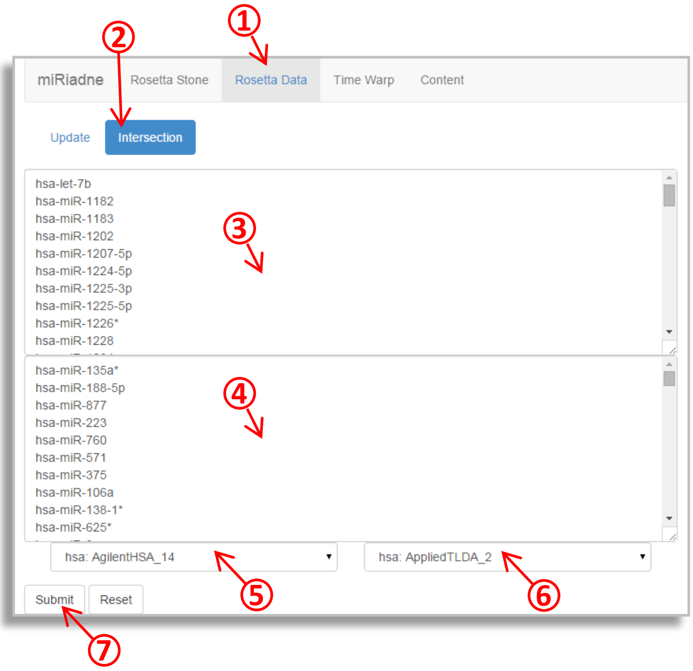
Search Name	AppliedTLDA:2	miRBase v21.0	Alternative Name	Sequence	Data
					Detector-PartNumber Detector B.CD5.007 B.CD5.012 B.CD5.015 B.memory.005 B.memory.010 B.memory.014 B.naive.001 B.naive.009 B.naive.013 CD4.CM.01 CD8.EMRA.060 CD8.EMRA.063 CD8.naive.021 CD8.naive.032 CD8.naive.052 NK.100 NK.102 NK.104
has-miR-155-4395459	hsa-miR-155	hsa-miR-155-5p	has-miR-155-4395459	UUAAUGCUAAUCGUGAUAGGGGU	has-miR-155 34.94935 40 40 40 40 40 34.00959 40 40 35.000324 34.99656 35.687134 40 40 40 40 40 34.934692 40 40 40 40 40 40 40 40 40 40 40 40 34.010
hsa-let-7a-4373169	hsa-let-7a	hsa-let-7a-5p	hsa-let-7a-4373169	UGAGGUAGUAGGUUGUAUAGUU	hsa-let-7a 27.956003 27.949995 30.0015 40 30.013632 29.082478 40 28.965137 40 33.972466 28.044567 29.001465 29.04713 25.971006 25.963377 27.0114
hsa-let-7a*-4395418	hsa-let-7a*	hsa-let-7a-3p	hsa-let-7a*-4395418	CUAUACAAUCUACUGUCUUUC	hsa-let-7a* 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 4
hsa-let-7b-4395446	hsa-let-7b	hsa-let-7b-5p	hsa-let-7b-4395446	UGAGGUAGUAGGUUGUGUGGUU	hsa-let-7b 28.97347 25.96626 26.966908 28.986168 28.017435 27.991459 29.983475 25.980146 27.957325 27.981613 28.008392 27.00888 26.989124 25.954
hsa-let-7b*-4395515	hsa-let-7b*	hsa-let-7b-3p	hsa-let-7b*-4395515	CUAUACAACCUACUGCCUUCCC	hsa-let-7b* 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 34.946648 40 40 40 40 30.933794 40 4
hsa-let-7c-4373167	hsa-let-7c	hsa-let-7c-5p	hsa-let-7c-4373167	UGAGGUAGUAGGUUGUAUGGUU	hsa-let-7c 40 33.99534 40 33.965706 34.991016 33.980907 40 34.98128 40 33.99543 37.50589 33.988876 33.992996 32.97006 32.97436 31.963964 31.97735
hsa-let-7d-4395394	hsa-let-7d	hsa-let-7d-5p	hsa-let-7d-4395394	AGAGGUAGUAGGUUGCAUAGUU	hsa-let-7d 28.97348 28.978745 29.978563 29.990316 29.979187 29.982036 29.967442 28.973011 29.956892 29.963448 28.957666 28.968138 28.972754 26.9
hsa-let-7d*-4378108	hsa-let-7d*	hsa-let-7d-3p	hsa-let-7d*-4378108	CUAUACGACCUGCUGCCUUUCU	hsa-let-7d* 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 4
hsa-let-7e-4395517	hsa-let-7e	hsa-let-7e-5p	hsa-let-7e-4395517	UGAGGUAGGAGGUUGUAUAGUU	hsa-let-7e 25.963758 24.973932 25.97985 26.977976 26.974665 26.977505 27.00448 24.95379 26.98793 26.983963 25.984491 25.971937 25.989286 23.9659
hsa-let-7e*-4395518	hsa-let-7e*	hsa-let-7e-3p	hsa-let-7e*-4395518	CUAUACGGCCUCCUAGCUUUCC	hsa-let-7e* 40 40 40 40 40 40 33.995453 40 40 40 40 40 40 40 40 40 35.921047 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 32.9
hsa-let-7f-4373164	hsa-let-7f	hsa-let-7f-5p	hsa-let-7f-4373164	UGAGGUAGUAGAUUGUAUAGUU	hsa-let-7f 27.0123 26.955847 40 40 40 40 40 27.053102 40 27.206642 26.958645 27.942986 40 25.996727 25.909754 40 40 40 26.957106 27.028116 25.0726
hsa-let-7f-1*-4395528	hsa-let-7f-1*	hsa-let-7f-1-3p	hsa-let-7f-1*-4395528	CUAUACAAUCUAUUGCCUUCCC	hsa-let-7f-1* 40 40 40 40 40 40 40 40 40 40 32.10214 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
hsa-let-7f-2*-4395529	hsa-let-7f-2*	hsa-let-7f-2-3p	hsa-let-7f-2*-4395529	CUAUACAGUCUACUGUCUUUCC	hsa-let-7f-2* 40 36.20481 40 40 40 40 40 40 40 40 31.601082 11.36088 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 32.934017 40 40 40 40 4
hsa-let-7g-4395393	hsa-let-7g	hsa-let-7g-5p	hsa-let-7g-4395393	UGAGGUAGUAGUUUGUACAGUU	hsa-let-7g 26.976242 25.97637 26.967018 27.968193 27.97662 26.978546 28.984703 26.980389 27.99672 26.98092 26.978748 25.969677 26.976288 24.9904
hsa-let-7g*-4395229	hsa-let-7g*	hsa-let-7g-3p	hsa-let-7g*-4395229	CUGUACAGGCCACUGCCUUGC	hsa-let-7g* 40 40 40 40 40 40 40 32.088043 40 36.00102 40 33.930485 40 40 40 40 33.01793 40 40 40 33.97097 40 40 40 40 40 40 40 40 40 40 33.071438 40
hsa-let-7i*-4395283	hsa-let-7i*	hsa-let-7i-3p	hsa-let-7i*-4395283	CUGCGCAAGCUACUGCCUUGCU	hsa-let-7i* 40 40 40 40 40 40 40 40 40 40 40 40 33.901463 40 40 40 40 40 40 40 40 40 40 40 40 40 34.946425 40 40 32.950542 40 40 40 40 40 40 40 40 40 40 40 33.
hsa-miR-1-4395333	hsa-miR-1	hsa-miR-1-3p	hsa-miR-1-4395333	UGGAAUGUAAAGAAGUAUGUAU	hsa-miR-1 40 40 40 40 40 40 40 40 40 40 40 31.264006 30.917568 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 26.928549 40 40 40 28.989944 40 2
hsa-miR-100-4373160	hsa-miR-100	hsa-miR-100-5p	hsa-miR-100-4373160	AACCCGUAGAUCCGAACUUGUG	hsa-miR-100 40 40 40 31.965265 33.00924 40 40 40 40 30.973993 33.996044 34.013893 29.994844 32.991146 40 29.95713 29.98146 27.971016 34.024014 4
hsa-miR-100*-4395253	hsa-miR-100*	hsa-miR-100-3p	hsa-miR-100*-4395253	CAAGCUUGUAUCUAUAGGUAUG	hsa-miR-100* 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 34.06079 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 35.95163 36.964



T39 fx 40

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	search_name	AppliedTLDA:2	platform_name	miRBase v21.0	sequence	Data Header														
2						Detector	B.CD5.007	B.CD5.012	B.CD5.015	B.memory	B.memory	B.memory	B.naive.0	B.naive.0	B.naive.01	CD4.CM.0	CD4.CM.0	CD4.CM.0	CD4.EM.0	CD4.EM.0
3	has-miR-155-4395459	hsa-miR-155	has-miR-155-4395459	hsa-miR-155-5p	UUAAUGCUAAUCGUGAUAGGGGU	has-miR-155	34.94935	40	40	40	40	40	34.00959	40	40	35.00032	34.99656	35.68713	40	40
4	hsa-let-7a-4373169	hsa-let-7a	hsa-let-7a-4373169	hsa-let-7a-5p	UGAGGUAGUAGGUUGUAUAGUU	hsa-let-7a	27.956	27.95	30.0015	40	30.01363	29.08248	40	28.96514	40	33.97247	28.04457	29.00147	29.04713	25.97101
5	hsa-let-7a*-4395418	hsa-let-7a*	hsa-let-7a*-4395418	hsa-let-7a-3p	CUAUACAAUCUACUGUCUUUC	hsa-let-7a*	40	40	40	40	40	40	40	40	40	40	40	40	40	40
6	hsa-let-7b-4395446	hsa-let-7b	hsa-let-7b-4395446	hsa-let-7b-5p	UGAGGUAGUAGGUUGUGUGGUU	hsa-let-7b	28.97347	25.96626	26.96691	28.98617	28.01744	27.99146	29.98348	25.98015	27.95733	27.98161	28.00839	27.00888	26.98912	25.95445
7	hsa-let-7b*-4395515	hsa-let-7b*	hsa-let-7b*-4395515	hsa-let-7b-3p	CUAUACAACCUACUGCCUUCCC	hsa-let-7b*	40	40	40	40	40	40	40	40	40	40	40	40	40	40
8	hsa-let-7c-4373167	hsa-let-7c	hsa-let-7c-4373167	hsa-let-7c-5p	UGAGGUAGUAGGUUGUAUGGUU	hsa-let-7c	40	33.99534	40	33.96571	34.99102	33.98091	40	34.98128	40	33.99543	37.50589	33.98888	33.993	32.97006
9	hsa-let-7d-4395394	hsa-let-7d	hsa-let-7d-4395394	hsa-let-7d-5p	AGAGGUAGUAGGUUGCAUAGUU	hsa-let-7d	28.97348	28.97875	29.97856	29.99032	29.97919	29.98204	29.96744	28.97301	29.95689	29.96345	28.95767	28.96814	28.97275	26.95913
10	hsa-let-7d*-4378108	hsa-let-7d*	hsa-let-7d*-4378108	hsa-let-7d-3p	CUAUACGACCUGCUGCCUUUCU	hsa-let-7d*	40	40	40	40	40	40	40	40	40	40	40	40	40	40
11	hsa-let-7e-4395517	hsa-let-7e	hsa-let-7e-4395517	hsa-let-7e-5p	UGAGGUAGGAGGUUGUAUAGUU	hsa-let-7e	25.96376	24.97393	25.97985	26.97798	26.97467	26.97751	27.00448	24.95379	26.98793	26.98396	25.98449	25.97194	25.98929	23.96594
12	hsa-let-7e*-4395518	hsa-let-7e*	hsa-let-7e*-4395518	hsa-let-7e-3p	CUAUACGGCCUCCUAGCUUUCC	hsa-let-7e*	40	40	40	40	40	40	33.99545	40	40	40	40	40	40	40
13	hsa-let-7f-4373164	hsa-let-7f	hsa-let-7f-4373164	hsa-let-7f-5p	UGAGGUAGUAGAUUGUAUAGUU	hsa-let-7f	27.0123	26.95585	40	40	40	40	40	27.0531	40	27.20664	26.95865	27.94299	40	25.99673
14	hsa-let-7f-1*-4395528	hsa-let-7f-1*	hsa-let-7f-1*-4395528	hsa-let-7f-1-3p	CUAUACAAUCUAUUGCCUUCCC	hsa-let-7f-1*	40	40	40	40	40	40	40	40	40	40	32.10214	40	40	40
15	hsa-let-7f-2*-4395529	hsa-let-7f-2*	hsa-let-7f-2*-4395529	hsa-let-7f-2-3p	CUAUACAGUCUACUGUCUUUCC	hsa-let-7f-2*	40	36.20481	40	40	40	40	40	40	40	40	31.60108	11.36088	40	40
16	hsa-let-7g-4395393	hsa-let-7g	hsa-let-7g-4395393	hsa-let-7g-5p	UGAGGUAGUAGUUUGUACAGUU	hsa-let-7g	26.97624	25.97637	26.96702	27.96819	27.97662	26.97855	28.9847	26.98039	27.99672	26.98092	26.97875	25.96968	26.97629	24.99046
17	hsa-let-7g*-4395229	hsa-let-7g*	hsa-let-7g*-4395229	hsa-let-7g-3p	CUGUACAGGCCACUGCCUUGC	hsa-let-7g*	40	40	40	40	40	40	40	32.08804	40	36.00102	40	33.93049	40	40
18	hsa-let-7i*-4395283	hsa-let-7i*	hsa-let-7i*-4395283	hsa-let-7i-3p	CUGCGCAAGCUACUGCCUUGCU	hsa-let-7i*	40	40	40	40	40	40	40	40	40	40	40	33.90146	40	40
19	hsa-miR-1-4395333	hsa-miR-1	hsa-miR-1-4395333	hsa-miR-1-3p	UGGAAUGUAAAGAAGUAUGUAU	hsa-miR-1	40	40	40	40	40	40	40	40	40	40	31.26401	30.91757	40	40
20	hsa-miR-100-4373160	hsa-miR-100	hsa-miR-100-4373160	hsa-miR-100-5p	AACCCGUAGAUCCGAACUUGUG	hsa-miR-100	40	40	40	31.96527	33.00924	40	40	40	40	30.97399	33.99604	34.01389	29.99484	32.99115
21	hsa-miR-100*-4395253	hsa-miR-100*	hsa-miR-100*-4395253	hsa-miR-100-3p	CAAGCUUGUAUCUAUAGGUAUG	hsa-miR-100*	40	40	40	40	40	40	40	40	40	40	40	40	40	40
22	hsa-miR-101-4395364	hsa-miR-101	hsa-miR-101-4395364	hsa-miR-101-3p	UACAGUACUGUGAUAACUGAA	hsa-miR-101	29.93463	29.9579	32.9795	32.96998	31.96399	30.96293	32.97797	29.97039	33.97095	31.98566	31.97214	29.97085	31.95644	28.97002
23	hsa-miR-101*-4395254	hsa-miR-101*	hsa-miR-101*-4395254	hsa-miR-101-5p	CAGUUAUCACAGUGCUGAUGCU	hsa-miR-101*	40	40	40	40	40	40	40	40	40	35.26703	40	35.9633	40	40
24	hsa-miR-103-4373158	hsa-miR-103	hsa-miR-103-4373158	hsa-miR-103a-3p	AGCAGCAUUGUACAGGGCUAUGA	hsa-miR-103	29.03257	28.98859	31.00808	29.99498	30.9737	29.99826	29.97088	28.95349	31.9946	30.97955	29.97399	29.00014	28.97324	26.9948
25	hsa-miR-105-4395278	hsa-miR-105	hsa-miR-105-4395278	hsa-miR-105-5p	UCAAAUGCUCAGACUCCUGUGGU	hsa-miR-105	40	40	40	40	40	40	40	40	40	40	40	40	40	40
26	hsa-miR-105*-4395279	hsa-miR-105*	hsa-miR-105*-4395279	hsa-miR-105-3p	ACGGAUGUUUGAGCAUGUGCUA	hsa-miR-105*	40	40	40	40	40	40	40	40	40	40	40	40	40	40
27	hsa-miR-106a-4395280	hsa-miR-106a	hsa-miR-106a-4395280	hsa-miR-106a-5p	AAAAGUGCUUACAGUGCAGGUAG	hsa-miR-106a	24.96421	23.96081	24.97988	25.96178	25.97807	25.97735	25.96646	23.96477	24.95641	24.94847	25.94321	25.97213	25.9519	24.97128
28	hsa-miR-106a*-4395281	hsa-miR-106a*	hsa-miR-106a*-4395281	hsa-miR-106a-3p	CUGCAAUGUAAGCACUUCUUAC	hsa-miR-106a*	40	40	40	40	40	40	40	40	40	40	40	40	40	40
29	hsa-miR-106b-4373155	hsa-miR-106b	hsa-miR-106b-4373155	hsa-miR-106b-5p	UAAAGUGCUGACAGUGCAGAU	hsa-miR-106b	26.99653	26.9732	27.96628	28.95401	27.98641	28.97157	28.97732	27.9729	28.97676	29.96341	28.96226	27.96923	29.96551	25.97217

Supplementary Figure 7

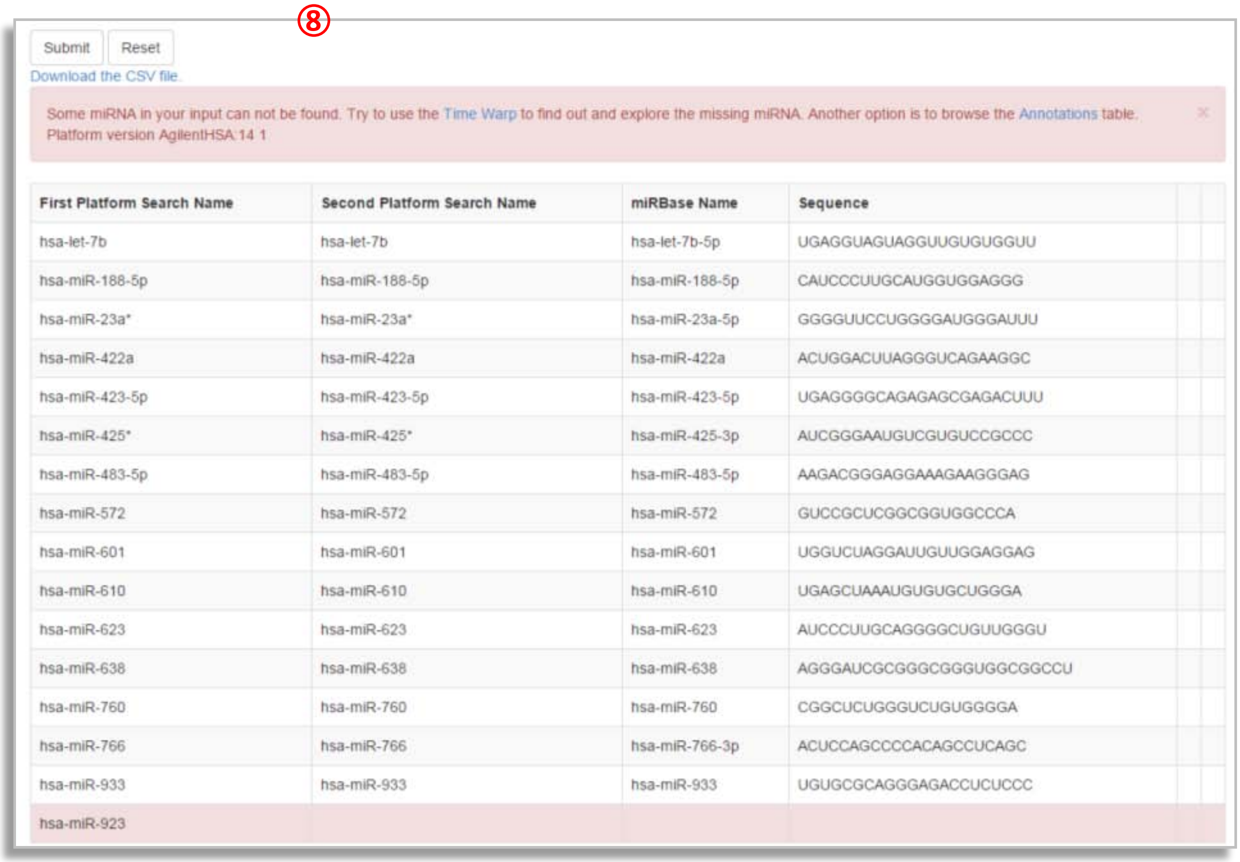


hsa-let-7b
hsa-miR-1182
hsa-miR-1183
hsa-miR-1202
hsa-miR-1207-5p
hsa-miR-1224-5p
hsa-miR-1225-3p
hsa-miR-1225-5p
hsa-miR-1226*
hsa-miR-1228

hsa-miR-135a*
hsa-miR-188-5p
hsa-miR-877
hsa-miR-223
hsa-miR-760
hsa-miR-571
hsa-miR-375
hsa-miR-106a
hsa-miR-138-1*
hsa-miR-625*

hsa: AgilentHSA_14

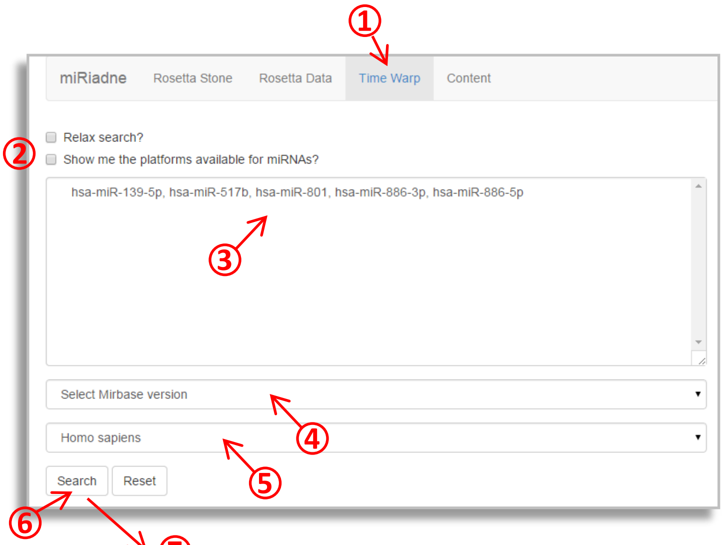
hsa: AppliedTLDA_2



Some miRNA in your input can not be found. Try to use the Time Warp to find out and explore the missing miRNA. Another option is to browse the Annotations table. Platform version AgilentHSA:14 1

First Platform Search Name	Second Platform Search Name	miRBase Name	Sequence
hsa-let-7b	hsa-let-7b	hsa-let-7b-5p	UGAGGUAGUAGGUUGUGUGGUU
hsa-miR-188-5p	hsa-miR-188-5p	hsa-miR-188-5p	CAUCCCUUGCAUGGUGGAGGG
hsa-miR-23a*	hsa-miR-23a*	hsa-miR-23a-5p	GGGGUUCCUGGGGAUGGGAUUU
hsa-miR-422a	hsa-miR-422a	hsa-miR-422a	ACUGGACUUAGGGUCAGAAGGC
hsa-miR-423-5p	hsa-miR-423-5p	hsa-miR-423-5p	UGAGGGGCAGAGAGCGAGACUUU
hsa-miR-425*	hsa-miR-425*	hsa-miR-425-3p	AUCGGGAAUGUCGUGUCCGCCC
hsa-miR-483-5p	hsa-miR-483-5p	hsa-miR-483-5p	AAGACGGGAGGAAAGAAGGGAG
hsa-miR-572	hsa-miR-572	hsa-miR-572	GUCCGCUCGGCGGUGGCCCA
hsa-miR-601	hsa-miR-601	hsa-miR-601	UGGUCUAGGAUUGUUGGAGGAG
hsa-miR-610	hsa-miR-610	hsa-miR-610	UGAGCUAAAUGUGUGCUGGGA
hsa-miR-623	hsa-miR-623	hsa-miR-623	AUCCCUUGCAGGGGCUGUUGGGU
hsa-miR-638	hsa-miR-638	hsa-miR-638	AGGGAUCGCGGGCGGGUGGCGGCCU
hsa-miR-760	hsa-miR-760	hsa-miR-760	CGGCUCUGGGUCUGUGGGGA
hsa-miR-766	hsa-miR-766	hsa-miR-766-3p	ACUCCAGCCCCACAGCCUCAGC
hsa-miR-933	hsa-miR-933	hsa-miR-933	UGUGCGCAGGGAGACCUCUCCC
hsa-miR-923			

Supplementary Figure 8

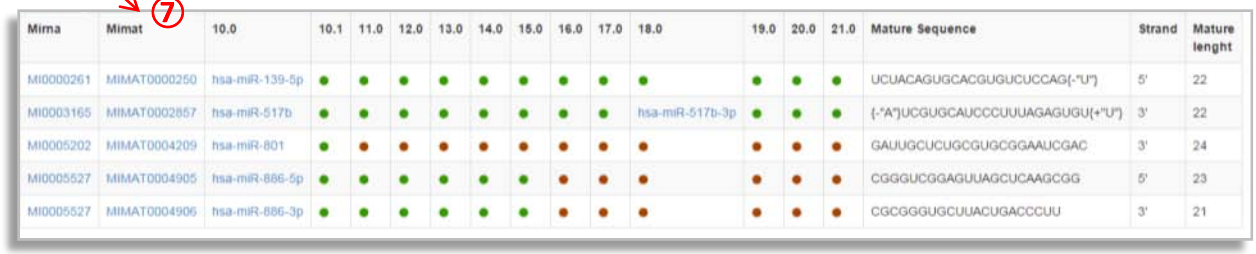


□ Relax search?
□ Show me the platforms available for miRNAs?

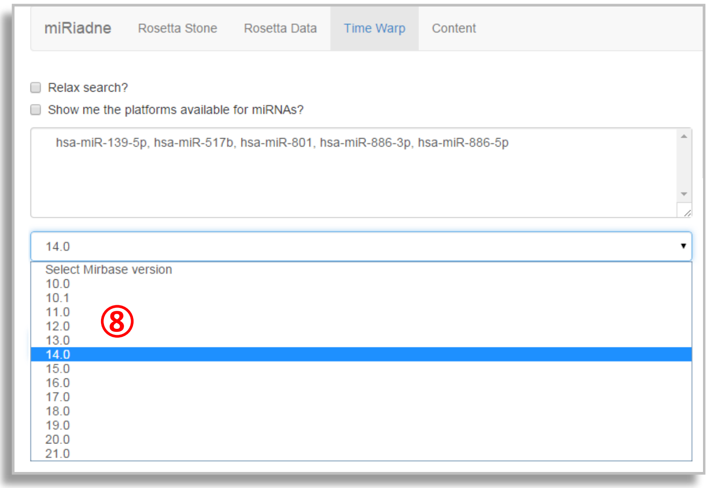
hsa-miR-139-5p, hsa-miR-517b, hsa-miR-801, hsa-miR-886-3p, hsa-miR-886-5p

Select Mirbase version ▾

Homo sapiens ▾



Mirna	Mimat	10.0	10.1	11.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0	19.0	20.0	21.0	Mature Sequence	Strand	Mature lenght
MI0000261	MIMAT0000250	hsa-miR-139-5p	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	UCUACAGUGCACGUGUCUCCAG(-"U")	5'	22
MI0003165	MIMAT0002857	hsa-miR-517b	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	hsa-miR-517b-3p	🟢	🟢	🟢	(-"A")UCGUGCAUCCCUUUAGAGUGU(+"U")	3'	22
MI0005202	MIMAT0004209	hsa-miR-801	🟢	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	🟠	GAUUGCUCUGCGUGCGGAAUCGAC	3'	24
MI0005527	MIMAT0004905	hsa-miR-886-5p	🟢	🟢	🟢	🟢	🟢	🟢	🟠	🟠	🟠	🟠	🟠	🟠	CGGGUCGGAGUUAGCUCAAGCGG	5'	23
MI0005527	MIMAT0004906	hsa-miR-886-3p	🟢	🟢	🟢	🟢	🟢	🟢	🟠	🟠	🟠	🟠	🟠	🟠	CGCGGGUGCUUACUGACCCUU	3'	21

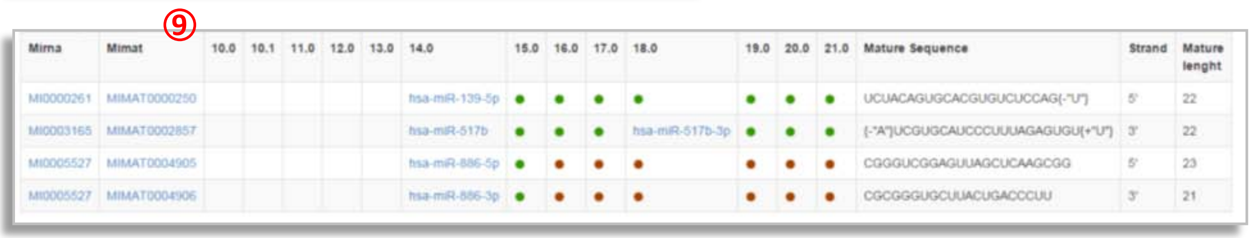


□ Relax search?
□ Show me the platforms available for miRNAs?

hsa-miR-139-5p, hsa-miR-517b, hsa-miR-801, hsa-miR-886-3p, hsa-miR-886-5p

14.0 ▾

Select Mirbase version
10.0
10.1
11.0
12.0
13.0
14.0
15.0
16.0
17.0
18.0
19.0
20.0
21.0



Mirna	Mimat	10.0	10.1	11.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0	19.0	20.0	21.0	Mature Sequence	Strand	Mature lenght
MI0000261	MIMAT0000250						hsa-miR-139-5p	🟢	🟢	🟢	🟢	🟢	🟢	🟢	UCUACAGUGCACGUGUCUCCAG(-"U")	5'	22
MI0003165	MIMAT0002857						hsa-miR-517b	🟢	🟢	🟢	hsa-miR-517b-3p	🟢	🟢	🟢	(-"A")UCGUGCAUCCCUUUAGAGUGU(+"U")	3'	22
MI0005527	MIMAT0004905						hsa-miR-886-5p	🟢	🟠	🟠	🟠	🟠	🟠	🟠	CGGGUCGGAGUUAGCUCAAGCGG	5'	23
MI0005527	MIMAT0004906						hsa-miR-886-3p	🟢	🟠	🟠	🟠	🟠	🟠	🟠	CGCGGGUGCUUACUGACCCUU	3'	21

References

1. Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.

2. Zhang,B., Wang,Q. and Pan,X. (2007) MicroRNAs and their regulatory roles in animals and plants. *J. Cell. Physiol.*, **210**, 279–289.

3. Lu,M., Zhang,Q., Deng,M., Miao,J., Guo,Y., Gao,W. and Cui,Q. (2008) An analysis of human microRNA and disease associations. *PLoS One*, **3**, e3420.

4. Nelson,K.M. and Weiss,G.J. (2008) MicroRNAs and cancer: past, present, and potential future. *Mol. Cancer Ther.*, **7**, 3655–3660.

5. Pagani,M., Rossetti,G., Panzeri,I., de Candia,P., Bonnal,R.J.P., Rossi,R.L., Geginat,J. and Abrignani,S. (2013) Role of microRNAs and long-non-coding RNAs in CD4(+) T-cell differentiation. *Immunol. Rev.*, **253**, 82–96.

6. Rossi,R.L., Rossetti,G., Wenandy,L., Curti,S., Ripamonti,A., Bonnal,R.J.P., Birolo,R.S., Moro,M., Crosti,M.C., Gruarin,P. *et al.* (2011) Distinct microRNA signatures in human lymphocyte subsets and enforcement of the naive state in CD4+ T cells by the microRNA miR-125b. *Nat. Immunol.*, **12**, 796–803.

7. De Candia,P., Torri,A., Pagani,M. and Abrignani,S. (2014) Serum microRNAs as biomarkers of human lymphocyte activation in health and disease. *Front. Immunol.*, **5**, 43.

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

9. Lee,R., Feinbaum,R. and Ambros,V. (2004) A short history of a short RNA. *Cell*, **116**, S89–S92.

10. Lagos-Quintana,M., Rauhut,R., Lendeckel,W. and Tuschl,T. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.

11. Ambros,V., Bartel,B., Bartel,D.P., Burge,C.B., Carrington,J.C., Chen,X., Dreyfuss,G., Eddy,S.R., Griffiths-Jones,S., Marshall,M. *et al.* (2003) A uniform system for microRNA annotation. *RNA*, **9**, 277–279.

12. Griffiths-Jones,S. (2004) The microRNA Registry. *Nucleic Acids Res.*, **32**, D109–D111.

13. Kozomara,A. and Griffiths-Jones,S. (2013) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.

14. Van Peer,G., Lefever,S., Anckaert,J., Beckers,A., Rihani,A., Van Goethem,A., Volders,P.-J., Zeka,F., Ongenaert,M., Mestdagh,P. *et al.* (2014) miRBase Tracker: keeping track of microRNA annotation changes. *Database (Oxford).*, bau080.

15. Lu,T.-P., Lee,C.-Y., Tsai,M.-H., Chiu,Y.-C., Hsiao,C.K., Lai,L.-C. and Chuang,E.Y. (2012) miRSystem: an integrated system for characterizing enriched functions and pathways of microRNA targets. *PLoS One*, **7**, e42390.

16. Brown,M., Suryawanshi,H., Hafner,M., Farazi,T.A. and Tuschl,T. (2013) Mammalian miRNA curation through next-generation sequencing. *Front. Genet.*, **4**, 145.

17. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.

18. Skogberg,G., Gudmundsdottir,J., van der Post,S., Sandstro¨ m,K., Bruhn,S., Benson,M., Mincheva-Nilsson,L., Baranov,V.,

Telemo,E. and Ekwall,O. (2013) Characterization of human thymic exosomes. *PLoS One*, **8**, e67554.

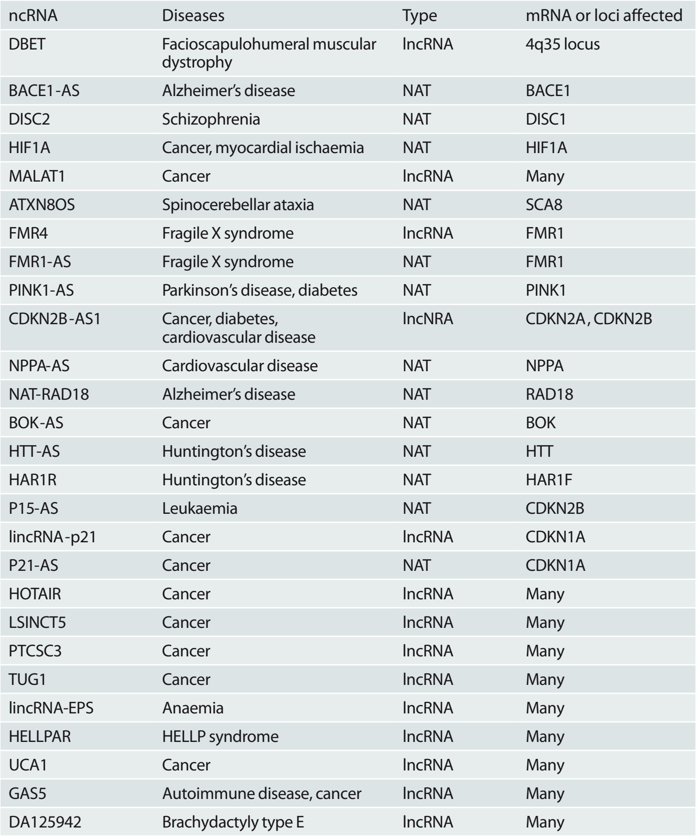
19. Vickers,K.C., Palmisano,B.T., Shoucri,B.M., Shamburek,R.D. and Remaley,A.T. (2011) MicroRNAs are transported in plasma and delivered to recipient cells by high-density lipoproteins. *Nat. Cell Biol.*, **13**, 423–433.

Chapter 4

Conclusions and perspectives in translational medicine

Thanks to new sequencing technologies, it has become clear that the non-coding part of the genome plays a fundamental role in different cellular functions. The number of non-coding transcripts grows steadily every day but relatively few examples of functional lncRNAs have been described in adaptive immune system. With our study, we not only investigated the expression landscape of annotated lincRNAs in 13 subsets of human primary lymphocytes, but we also identified novel lincRNA specifically expressed in these cells. The identification of new lincRNAs by *de novo* approaches is one of the most important features of the study because these transcripts, thanks to their high specificity and restricted expression, define cellular identity better than protein-coding genes.

Given these observations, it is not surprising that an altered expression of lncRNAs could be linked to many different pathologies (Table 4). Indeed, it has been found that the >90% of disease-associated genetic variants identified by Genome Wide Association Studies (GWAS) are located outside protein coding regions implicated in transcription control (promoters and enhancers) or in gene expression (non coding genes) [12, 26]. In particular, it has been estimated that 7% of SNPs associated to immuno-mediated diseases are in intergenic non coding region (lincRNAs). A powerful method to elucidate the genetic component underlying altered gene expression is mapping of expression quantitative trait loci (eQTLs): some GWAS-SNPs have eQTLs effects on these lncRNA transcripts [17]. If the majority of lncRNAs is involved in the regulation of protein coding genes expression, the SNPs associated with these transcripts may in turn indirectly influence the protein genes related to a given disease.



ncRNA	Diseases	Type	mRNA or loci affected
DBET	Facioscapulohumeral muscular dystrophy	lncRNA	4q35 locus
BACE1-AS	Alzheimer's disease	NAT	BACE1
DISC2	Schizophrenia	NAT	DISC1
HIF1A	Cancer, myocardial ischaemia	NAT	HIF1A
MALAT1	Cancer	lncRNA	Many
ATXN8OS	Spinocerebellar ataxia	NAT	SCA8
FMR4	Fragile X syndrome	lncRNA	FMR1
FMR1-AS	Fragile X syndrome	NAT	FMR1
PINK1-AS	Parkinson's disease, diabetes	NAT	PINK1
CDKN2B-AS1	Cancer, diabetes, cardiovascular disease	lncNRA	CDKN2A , CDKN2B
NPPA-AS	Cardiovascular disease	NAT	NPPA
NAT-RAD18	Alzheimer's disease	NAT	RAD18
BOK-AS	Cancer	NAT	BOK
HTT-AS	Huntington's disease	NAT	HTT
HAR1R	Huntington's disease	NAT	HAR1F
P15-AS	Leukaemia	NAT	CDKN2B
lincRNA-p21	Cancer	lncRNA	CDKN1A
P21-AS	Cancer	NAT	CDKN1A
HOTAIR	Cancer	lncRNA	Many
LSINCT5	Cancer	lncRNA	Many
PTCSC3	Cancer	lncRNA	Many
TUG1	Cancer	lncRNA	Many
lincRNA-EPS	Anaemia	lncRNA	Many
HELLPAR	HELLP syndrome	lncRNA	Many
UCA1	Cancer	lncRNA	Many
GAS5	Autoimmune disease, cancer	lncRNA	Many
DA125942	Brachydactyly type E	lncRNA	Many

Table 4. LncRNAs with potential roles in human diseases [37].

The potential role and involvment of lncRNAs related to immune-mediated diseases is still unclear, reflecting the poor knowledge on this field. Emerging lines of evidence suggest that expression dysregulation and large or small-scale mutations in primary sequence of lncRNAs are strongly linked with phenotype changes and disease susceptibility. In particular, if genetic mutations are located within regulatory motifs of lncRNAs, when these molecules are transcribed their functions could be compromised. Indeed, alterations of the lncRNA functions could potentially induce changes in their folding patterns, secondary structure stability and affect expression [22]. For example, a systematic study of all annotated IBD (inflammatory bowel disease) and T1D (type 1 diabetes) loci-associated lncRNAs showed that several SNPs within these loci have a significant propensity to disrupt lncRNA secondary structures. These SNPs induce structural perturbations that influence the molecular function of lncRNAs, altering their transcription levels, their associated protein coding genes and contributing towards disease phenotype [22]. Another lncRNA called *ANRIL* has been identified within p15/CDKN2B-p16/CDKN2A-p14/ARF gene cluster as a major hotspot in GWAS studies associated to coronary disease, intracranial aneurysm and type 2 diabetes. This lncRNA has been shown to regulate its neighbor tumor suppressors CDKN2A/B by epigenetic mechanisms, through the binding to two polycomb proteins CBX7 (PCR1) and SUZ12 (PCR2), which result in the regulation of histone modification in CDKN2A/B locus. More recently *ANRIL* was linked to several cancers, glaucoma, endometriosis and diabetes. For example, it is over-expressed in leukemia patients leukocytes compared with normal controls, while CDKN2B showed the opposite pattern of expression [43]. Remarkably, GWAS studies have detected a risk allele for acute lymphoblastic leukemia in the first exon of CDKN2A [28], coinciding with the promoter region of *ANRIL*. Moreover, *ANRIL* is also up-regulated in prostate cancer tissues in comparison with normal epithelial cells, accompanied by down-regulation of CDKN2A [41].

A genome-wide profiling of mRNA and lincRNA in human primary NT cells, CD4$^+$ T central memory cells, and T$_H$1, T$_H$2, T$_H$17, and Treg cells was performed by Zhang et al., to identified T$_H$2-specific markers relevant for allergy. They found that lncRNA *GATA3-AS1* is specifically expressed in primary T$_H$2, it is significantly correlated with GATA3 and it shares the same promoter. This lncRNA showed an increase expression in patients with T$_H$2-associated disease (seasonal allergic rhinitis) compared to healthy control subjects. Then, lncRNA *GATA3-AS1* is now considered a specific indicator of T$_H$2-response and T$_H$2-associated diseases [44].

Also ulcera colitis and Chron's disease (CD) are connected to lncRNAs. In particular LRRK2 gene, encoded by a major susceptibility gene for CD and ulcera colitis, is a component of a complex that directly inhibits the traslocation of the transcription factor NFAT to the nucleus, altering the immune responses. This complex includes the lncRNA *NRON* (NFAT repressor) that is directly involved in the binding of NFAT and LRRK2, in fact the knockdown of *NRON* results in the release of LRRK2 and translocation of NFAT to the nucleus [21]. Another lncRNA, *DQ786243*, correlates with the severity of CD [25]. This lncRNA affects the expression of CREB and Foxp3 through which it regulates Treg cells function [45]. *DQ786243* is upregulated in the blood of patients and it seems to be involved in the regulation of CREB phosphorylation, contributing to the development of CD [25].

LncRNA play also a special role in the control of autoimmune diseases. The increase risk of autoimmune thyroid disease (AITD) is correlated to the SNP Ex9b-SNP10 [29] that resides in intron 9 of the protein-coding gene ZFAT and the promoter region of the antisense lncRNA *SAS-ZFAT*. When the variation is present, *SAS-ZFAT* is upregulated and in turn downregulates the expression level of its antisense counterpart. SAS-ZFAT is exclusively expressed in CD19^{+}B cells in peripheral blood lymphocytes, therefore this lncRNA play a critical role in B cell function and determine susceptibility to AITD [18]. Another example is lncRNA *PRINS* (Psoriasis Susceptibility-related non-coding RNA gene) that harbors two repeat elements [30]. This lncRNA is expressed at different levels in several human tissues but shows a higher expression in the uninvolved epidermis of psoriatic patients than in both psoriatic lesional and healthy epidermis, suggesting a role in psoriatic susceptibility. The RNA expression level of *PRINS* is decreased in the uninvolved psoriatic, but not healthy, epidermis with treatment of T-lymphokines that are known to precipitate psoriatic symptoms. Moreover, downregulating the RNA level of PRINS by RNAi can impair cell viability after serum starvation, but not under normal serum conditions. Then, *PRINS* acts as "riboregulator" involved in the proliferation and survival of cells [18].

A recent study bu Hrdlickova and collegues, investigated the shared genetics of autoimmune and immune-related diseases (AID) considering eight different AID pathologies [14]: autoimmune thyroid disease [5], celiac disease (CeD) [35], inflammatory bowel disease (IBD) [16], juvenile idiopathic arthritis (JIA) [13], primary biliary cirrhosis (PBC) [20], psoriasis (PS) [36], primary sclerosing cholangitis (PsCh) [19] and rheumatoid arthritis (RA) [7]. They identified 284 loci related to these pathologies, 119 of which overlapped partly or completely in two or more AID (defined "AID shared loci"). These loci harbour 240 lncRNAs and 626 protein coding genes. Comparing the mean expression levels of lncRNAs versus protein coding genes, only a two-fold lower expression of AID lncRNAs clearly emerged, suggesting that these transcripts in AID are expressed to higher levels than previously assumed in particular in cell types functionally involved in the disease. This observation highlights an important role of lncRNAs located at AID loci, as descriptors more specific than AID protein coding genes [14].

Also bioinformatic tools are used to integrate and update information about lncRNAs and diseases.

LncRNADisease (http://www.cuilab.cn/lncrnadisease) is a manually curated database that collects all "lncRNA–disease" relations experimentally reported in the literature. Of all the included pathologies, cancer (39.8%), cardiovascular disease (10.8%) and neurodegeneration disease (8.4%) are the top three classes. Moreover novel lncRNA–disease associations are predicted based on the genomic context of a given lncRNA and integrated into the database [3].

Long noncoding RNAs could also be promising novel target for therapies or biomarkers for diagnosis and prognosis. The currently available drugs and tool compounds are focused on inhibitory mechanism of action and only very few pharmaceutical agents are able to increase the activity of effectors or pathways for therapeutic benefit. In specific contexts, though, the upregulation of key genes, as transcription factors, tumor suppressors, growth factors or more recently discovered lncRNA regulators, would be desired in specific context [37].

Different strategies for the therapeutical manipulation of NATs (Natural Antisense Transcripts) and other lncRNAs are evolving in the last years. Single-stranded oligonucleotides (also called "antago-NATs") are designed to specifically block the interaction between NAT transcript and sense gene mRNA and/or degrade the NAT transcript itself. The outcome of this approach is transcriptional derepression of the gene (Figure 11).

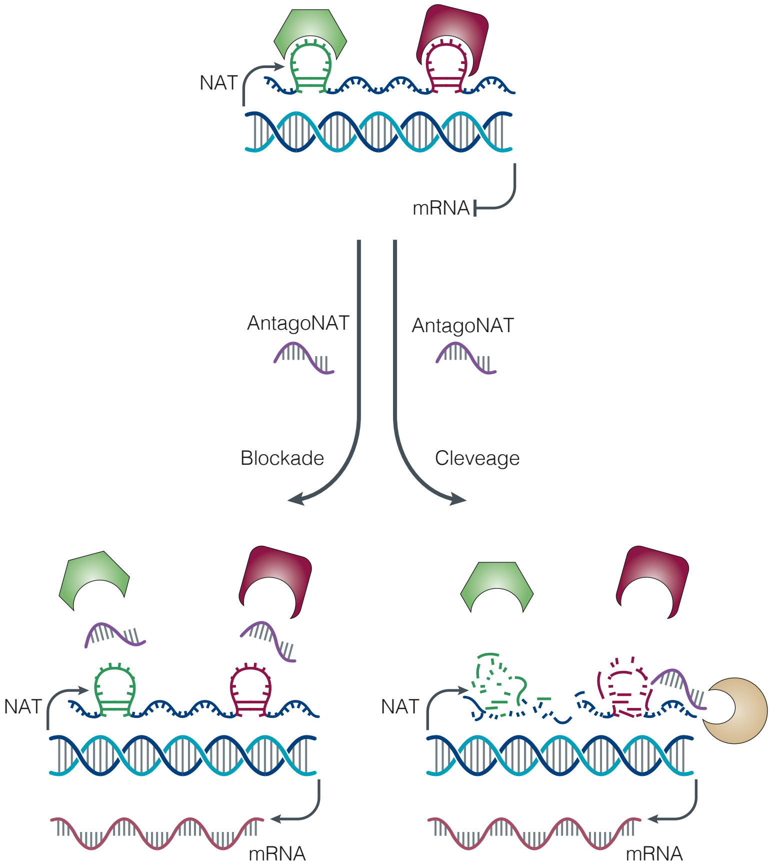


Figure 11. Simplified scheme of antagoNAT design. Natural antisense transcripts (NATs) can repress mRNA gene expression at the transcriptional level. Targeting NATs with single-stranded oligonucleotides named antagoNATs can result in mRNA derepression. AntagoNATs can act by blocking the interactions of NATs with effector proteins (bottom left) and/or by causing RNAase H-mediated degradation of the antisense transcript (bottom right). LSD1, lysine-specific histone demethylase 1; PRC2, Polycomb repressive complex 2 [37].

Several characteristics should be considered to develop an antagoNAT with potential *in vivo* applications, including modifications to promote metabolic stability and to minimize the oligonucleotide length for cellular uptake and off-target activities. This method induces locus-specific upregulation of the gene of interest without affecting unrelated or neighboring genes. Also double-stranded small interfering RNAs (siRNAs) are exploited *in vitro* to induce upregulation of genes, targeting a region that does not directly overlap with sense transcript [37]. Finally, promoter region targeting by small double-stranded RNAs has also been considered as a tool to increase gene expression [37].

It is also important to underline the differences between inhibition of NATs and miRNAs. First of all, NATs are involved in the repression of transcription acting as chromatin regulators, whereas miRNAs are translational repressors that affect mRNA stability in the cytoplasm. Second, the action of cis-acting NATs is gene-locus specific, whereas the number of miRNAs targets is larger and therefore the effects of these molecules less specific. Third, the number of NATs discovered is steadily growing whereas the miRNAs counterpart is less abundant and already extensively studied. Finally, NAT inhibition by an antagoNAT can be achieved by the cleavage or blockade of RNA or a protein target, whereas steric blockade is the primary option for the inhibition of a mature miRNA.

There are important aspects in targeting lncRNAs that cannot be ignored: the toxicity and off-target effects. These effects, observed with all oligonucleotide antisense technologies, may be due to protein binding [1]. Although the oligonucleotide chemistry can have an impact on adverse events, both the prevalence and incidence of these adverse events are sequence-dependent. However, in most instances it remains unclear which sequence motifs underlie such liabilities, so preclinical experiments in animal models are required to eliminate those sequences that appear to be most toxic. Unwanted and unanticipated events can results also by hybridization of the oligonucleotides with off-targets. This effect is difficult to control, but it is possible to predict some of nonspecific hybridization events using bioinformatics tools. It is emerging for antagoNATs, that only 10-50% of these molecules will reduce the expression of the target, these numbers are in line with what has been observed for other oligonucleotide antisense technologies. Also RNA structure and association with various proteins could potentially influence the effects of antisense technologies. Safe and efficient *in vivo* delivery is another crucial hurdle that lncRNA targeting technologies have in common with oligonucleotide-based therapies. Systemic delivery by once-weekly intravenous treatment was approved in January 2013 by the US Food and Drug Administration (FDA) for a NAT targeting the mRNA for apolipoprotein B, but denied by the European Medicines Agency panel due to potential adverse effect. Targeted or highly localized delivery was used to treat the central nervous system (CNS) with relative success [39, 40, 32, 42, 31], especially through the intrathecal route [11, 27]. Acting on chemical modifications [6] or using various delivery approaches (viral or non-viral vectors) can help to decrease immune activation and to achieve a proper dosing control of treatments. Notably, a major advantage of antagoNATs is their ability to be administered systemically without requirement for any delivery vehicles [40, 33].

The immune system is an interesting model for lncRNAs-mediated therapies. Given that lncRNAs act as fine-tuners of cell differentiation, we could modulate the differentiation network acting through lncRNAs, ensuring a proper immune response to different external clues and challenges in a way that is advantageous in terms of host defense. These therapies could provide a way to modulate the balance between effector lymphocytes reprogramming already differentiated cells in a less invasive and more specific way. Indeed, given that lncRNAs are not major hubs within cell networks, their overexpression or downregulation would cause a more physiological cascade with minor perturbations if compared to the overexpression or downregulation of a key regulatory hub such as a master gene. As in our case, understanding the mechanisms of function of lncRNAs in driving the differentiation events in the human immune system is of central importance for the identification of novel and more specific therapeutic targets for immune-related diseases.

References

1. Bennett CF, Swayze EE. RNA targeting therapeutics: molecular mechanisms of antisense oligonucleotides as a therapeutic platform. *Annual review of pharmacology and toxicology* 2010, 50: 259-293.

2. Burd CE, Jeck WR, Liu Y, Sanoff HK, Wang Z, Sharpless NE. Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. *PLoS genetics* 2010, 6(12): e1001233.

3. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, et al. LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic acids research* 2013, 41(Database issue): D983-986.

4. Congrains A, Kamide K, Ohishi M, Rakugi H. ANRIL: Molecular Mechanisms and Implications in Human Health. *International journal of molecular sciences* 2013, 14(1): 1278-1292.

5. Cooper JD, Simmonds MJ, Walker NM, Burren O, Brand OJ, Guo H, et al. Seven newly identified loci for autoimmune thyroid disease. *Human molecular genetics* 2012, 21(23): 5202-5208.

6. Dirin M, Winkler J. Influence of diverse chemical modifications on the ADME characteristics and toxicology of antisense oligonucleotides. *Expert opinion on biological therapy* 2013, 13(6): 875-888.

7. Eyre S, Bowes J, Diogo D, Lee A, Barton A, Martin P, et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature* genetics 2012, 44(12): 1336-1340.

8. Gutschner T, Hammerle M, Diederichs S. MALAT1 – a paradigm for long noncoding RNA function in cancer. *Journal of molecular medicine* 2013, 91(7): 791-801.

9. Gutschner T, Hammerle M, Eissmann M, Hsu J, Kim Y, Hung

G, et al. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer research* 2013, 73(3): 1180-1189.

10. Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. *Nature* 2011, 470(7333): 264-268.

11. Hayek SM, Deer TR, Pope JE, Panchal SJ, Patel VB. Intrathecal therapy for cancer and non-cancer pain. *Pain physician* 2011, 14(3): 219-248.

12. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* 2009, 106(23): 9362-9367.

13. Hinks A, Cobb J, Marion MC, Prahalad S, Sudman M, Bowes J, et al. Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nature genetics* 2013, 45(6): 664-669.

14. Hrdlickova B, Kumar V, Kanduri K, Zhernakova DV, Tripathi S, Karjalainen J, et al. Expression profiles of long non-coding RNAs located in autoimmune disease-associated regions reveal immune cell-type specificity. *Genome medicine* 2014, 6(10): 88.

15. Ji P, Diederichs S, Wang W, Boing S, Metzger R, Schneider PM, et al. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 2003, 22(39): 8031-8041.

16. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012, 491(7422):

119-124.

17. Kumar V, Westra HJ, Karjalainen J, Zhernakova DV, Esko T, Hrdlickova B, et al. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS genetics* 2013, 9(1): e1003201.

18. Li J, Xuan Z, Liu C. Long non-coding RNAs and complex human diseases. *International journal of molecular sciences* 2013, 14(9): 18790-18808.

19. Liu JZ, Almarri MA, Gaffney DJ, Mells GF, Jostins L, Cordell HJ, et al. Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nature genetics* 2012, 44(10): 1137-1141.

20. Liu JZ, Hov JR, Folseraas T, Ellinghaus E, Rushbrook SM, Doncheva NT, et al. Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nature genetics* 2013, 45(6): 670-675.

21. Liu Z, Lee J, Krummey S, Lu W, Cai H, Lenardo MJ. The kinase LRRK2 is a regulator of the transcription factor NFAT that modulates the severity of inflammatory bowel disease. *Nature immunology* 2011, 12(11): 1063-1070.

22. Mirza AH, Kaur S, Brorsson CA, Pociot F. Effects of GWAS-associated genetic variants on lncRNAs within IBD and T1D candidate loci. *PloS one* 2014, 9(8): e105723.

23. Panzeri I, Rossetti G, Abrignani S, Pagani M. Long Intergenic Non-Coding RNAs: Novel Drivers of Human Lymphocyte Differentiation. *Frontiers in immunology* 2015, 6: 175.

24. Pasmant E, Sabbagh A, Vidaud M, Bieche I. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 2011, 25(2): 444-448.

25. Qiao YQ, Huang ML, Xu AT, Zhao D, Ran ZH, Shen J.

LncRNA DQ786243 affects Treg related CREB and Foxp3 expression in Crohn's disease. *Journal of biomedical science* 2013, 20: 87.

26. Ricano-Ponce I, Wijmenga C. Mapping of immune-mediated disease genes. *Annual review of genomics and human genetics* 2013, 14: 325-353.

27. Rigo F, Hua Y, Krainer AR, Bennett CF. Antisense-based therapy for the treatment of spinal muscular atrophy. *The Journal of cell biology* 2012, 199(1): 21-25.

28. Sherborne AL, Hosking FJ, Prasad RB, Kumar R, Koehler R, Vijayakrishnan J, et al. Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk. *Nature genetics* 2010, 42(6): 492-494.

29. Shirasawa S, Harada H, Furugaki K, Akamizu T, Ishikawa N, Ito K, et al. SNPs in the promoter of a B cell-specific antisense transcript, SAS-ZFAT, determine susceptibility to autoimmune thyroid disease. *Human molecular genetics* 2004, 13(19): 2221-2231.

30. Sonkoly E, Bata-Csorgo Z, Pivarcsi A, Polyanka H, Kenderessy-Szabo A, Molnar G, et al. Identification and characterization of a novel, psoriasis susceptibility-related noncoding RNA gene, PRINS. *The Journal of biological chemistry* 2005, 280(25): 24159-24167.

31. Southwell AL, Skotte NH, Bennett CF, Hayden MR. Antisense oligonucleotide therapeutics for inherited neurodegenerative diseases. *Trends in molecular medicine* 2012, 18(11): 634-643.

32. Standifer KM, Chien CC, Wahlestedt C, Brown GP, Pasternak GW. Selective loss of delta opioid analgesia and binding by antisense oligodeoxynucleotides to a delta opioid receptor. *Neuron* 1994, 12(4): 805-810.

33. Stein CA. The experimental use of antisense oligonucleotides: a guide for the perplexed. *The Journal of clinical investigation* 2001, 108(5): 641-644.

34. Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, et al. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular cell* 2010, 39(6): 925-938.

35. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature genetics* 2011, 43(12): 1193-1201.

36. Tsoi LC, Spain SL, Knight J, Ellinghaus E, Stuart PE, Capon F, et al. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nature genetics* 2012, 44(12): 1341-1348.

37. Wahlestedt C. Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nature reviews Drug discovery* 2013, 12(6): 433-446.

38. Wahlestedt C, Golanov E, Yamamoto S, Yee F, Ericson H, Yoo H, et al. Antisense oligodeoxynucleotides to NMDA-R1 receptor channel protect cortical neurons from excitotoxicity and reduce focal ischaemic infarctions. *Nature* 1993, 363(6426): 260-263.

39. Wahlestedt C, Pich EM, Koob GF, Yee F, Heilig M. Modulation of anxiety and neuropeptide Y-Y1 receptors by antisense oligodeoxynucleotides. *Science* 1993, 259(5094): 528-531.

40. Wahlestedt C, Salmi P, Good L, Kela J, Johnsson T, Hokfelt T, et al. Potent and nontoxic antisense oligonucleotides containing locked nucleic acids. *Proceedings of the National Academy of Sciences of the United States of America* 2000, 97(10): 5633-5638.

41. Yap KL, Li S, Munoz-Cabello AM, Raguz S, Zeng L, Mujtaba S, et al. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Molecular cell* 2010, 38(5): 662-674.

42. Yee F, Ericson H, Reis DJ, Wahlestedt C. Cellular uptake of intracerebroventricularly administered biotin- or digoxigenin-labeled antisense oligodeoxynucleotides in the rat. *Cellular and molecular neurobiology* 1994, 14(5): 475-486.

43. Yu W, Gius D, Onyango P, Muldoon-Jacobs K, Karp J, Feinberg AP, et al. Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* 2008, 451(7175): 202-206.

44. Zhang H, Nestor CE, Zhao S, Lentini A, Bohle B, Benson M, et al. Profiling of human CD4+ T-cell subsets identifies the TH2-specific noncoding RNA GATA3-AS1. *The Journal of allergy and clinical immunology* 2013, 132(4): 1005-1008.

45. Zhang L, Zhao Y. The regulation of Foxp3 expression in regulatory CD4(+)CD25(+)T cells: multiple pathways on the road. *ss* 2007, 211(3): 590-597.