Identification of regulatory non coding RNAs in human lymphocytes

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

Dr. Valeria RANZANI

Matr. No. 067669
..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

Introduction
The landscape of human transcription 7
Non coding RNA genes and the modern RNA world 10
Predicting long non coding RNAs using RNA sequencing 19
Functional interactions among microRNAs and long non coding RNAs
Long intergenic non coding RNAs: novel drivers in human lymphocyte differentiation

Scope of the thesis 42

References 43

LincRNAs landscape in human lymphocytes 63
highlights regulation of T cell differentiation by linc-MAF-4

Abstract 64
Introduction 65
Results 67
Discussion 73
Online methods 75
Author contribution 85
Acknowledgments 85
Figure and Table Legends 86
Figures and Tables 90
References

miRiadne: a web tool to harmonize miRNA nomenclature

Abstract
Introduction
Implementation
Features and inputs
Discussion
Acknowledgments
Figure and Table Legends
Figures and Tables

References

Conclusions and perspectives in translational medicine

References
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 re-define the gene is attempting to retrofit biological complexity into an
existing vocabulary that is understandable 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 polymorphism are commonly observed in both cancer and immortalized cells [10, 162, 16, 26; The ENCODE Project Consortium 2012] as well as the presence of polymorphism [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 coun-
terparts [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) possess regulatory functions in several cellular processes including cell [6, 20], cancer 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].
ncRNAs (lincRNA), intronic lncRNA, sense or antisense lncRNA, com-

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

<table>
<thead>
<tr>
<th>ncRNA</th>
<th>Length (nt)</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>miRNAs</td>
<td>21–23</td>
<td>In animals, associate with the mRNA-induced silencing complex (RISC) and</td>
</tr>
<tr>
<td>small</td>
<td>60–300</td>
<td>Help the chemical modification of miRNAs, thereby influencing stability,</td>
</tr>
<tr>
<td>nucleolar</td>
<td>150</td>
<td>Assist splicing of mRNAs from primary genomic transcripts</td>
</tr>
<tr>
<td>PIWI-interacting RNAs</td>
<td>25–33</td>
<td>Associate with the highly conserved PIWI family of argonaute proteins and</td>
</tr>
<tr>
<td>TAS RNAs</td>
<td>22–200</td>
<td>Enriched at the 5′ end of genes, within 2.5 kb of TSS. Can be transcribed</td>
</tr>
<tr>
<td>shRNA</td>
<td>21–23</td>
<td>Processed from a plethora of genomic sources, both foreign (virus) and</td>
</tr>
<tr>
<td>miRNA</td>
<td>15–30</td>
<td>Enriched immediately downstream transcriptional start sites (TSSs) of</td>
</tr>
<tr>
<td>LncRNA</td>
<td>&gt; 200</td>
<td>Transcribed from the same locus but opposite strand of the overlapping</td>
</tr>
<tr>
<td>PNA</td>
<td>200–1000</td>
<td>Enriched at promoters, found to regulate gene expression</td>
</tr>
<tr>
<td>P Ago 3P</td>
<td>200–600</td>
<td>Enriched at RNA-Seq, CG-rich promoter with broad TSSs. Affect promoter</td>
</tr>
<tr>
<td>UC RNA</td>
<td>&gt; 200</td>
<td>Perfectly conserved between human, rat, and mouse. Frequently located at</td>
</tr>
<tr>
<td>enhancer RNA</td>
<td>&gt; 200</td>
<td>Transcribed from introns of overlapping protein-coding sequences. Involved in</td>
</tr>
<tr>
<td>lincRNA</td>
<td>&gt; 200</td>
<td>Gene expression regulation, regulation of cellular processes</td>
</tr>
<tr>
<td>circRNA</td>
<td>&gt; 1000</td>
<td>Derive within 3′ untranslated region (3′ UTR) sequences. Function still not</td>
</tr>
<tr>
<td></td>
<td>&gt; 4000</td>
<td>Diverse, from templates for viral replication to transcriptional regulators</td>
</tr>
</tbody>
</table>

IncRNAs are commonly classified in association with annotated protein-coding mRNAs (Figure 2) and comprise the long intergenic ncRNAs (lincRNA), intronic IncRNA, 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 lincRNAs (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 lincRNAs, 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]. LincRNAs are defined bidirectional when their expression and the expression of a neighboring coding transcript is initiated in close proximity and intergenic lincRNAs if they lie between two different coding regions with no overlap [25]. Intergenic long ncRNAs are the most studied lincRNAs because they are independent transcriptional 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 recruitment 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 independent 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-coding RNAs (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 and 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) IncRNA-ES1 and IncRNA-ES2 associate with both PRC2 and SOX2 inducing the silencing of SOX2-bound developmental genes to control embryonic 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 independent 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 IncRNA function. The mechanism involved a base-pairing complementarity between IncRNAs and target RNA sequence. a) complementarity between BACE1 mRNA and its 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 IncRNA with recruitment of UPF1 RNA helicase. c) Contrariwise, STAU1-mediated mRNA stabilization with the implication of TINCR that recognizes its target mRNA through a 25 base-long motif. [75]. d) Mechanism of inhibition of translation that implicates Trp53cor1 IncRNA together with RNA helicase RCK [171]. e) Uchl1-as1 IncRNA induced translation upon stress induction [14] [34].

A peculiar mode of action of IncRNAs 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 based 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 lincRNAs 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 hypothesis 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 accurate 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 definition 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-length 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-length 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 abundance 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).

![Diagram of workflow](image)

**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 sequencing:** 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 repetitive 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 available, 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 efficiently 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 particular 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 associa-
tion 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 amino acids 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 derive 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 amino acid 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 likelihood 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 a 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 involvement 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 overexpression 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 interaction 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 thyroid cancer [33], *H19* includes 4 let-7 target sites within its sequence [64] and its down-modulation by these miRNAs promoted myotube formation in C2C12
myoblasts [64]. Finally, \textit{lincRNA-RoR} is targeted by many miRNAs including the putative regulators miR145-5p, miR-181a-5p and miR-99b-3p that specifically decrease the activity of the lncRNA in human embryonic stem cell [166].

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{image.png}
\caption{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...}
\end{figure}
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, accelerates 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].
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].

### IncRNAs 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].
**IncRNAs 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 evidences reveals that IncRNAs and microRNAs work jointly to contribute to a robust and dynamic control of gene expression through complex post-transcriptional mechanisms.

**Bioinformatic tools to investigate IncRNAs-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 IncRNAs. 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 collects ~9000 miRNA-circRNA, 16000 miRNA-pseudogene and 285 000 protein–RNA regulatory relationships. It also provides 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 dynamically 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 involvement 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 antigen recognition CD4$^+$ naïve T cells differentiate into distinct T helper subsets characterized by the expression of specific master transcription 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 flexibility and plasticity or in the maintenance of cell identity together with lineage-specific transcription factors, cytokines and other ncRNAs. In particular, ncRNAs seem 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 environment 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 occur 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.
Table 3. Studies on lncRNAs in the adaptive immune system [124].

<table>
<thead>
<tr>
<th>Sample</th>
<th>LncRNAs</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Granulocytes, monocytes, NK, B, naive CD6+CD3+ and CD4+ memory human T cells, in vitro polarized precursors T helper T+ α1, Tα2, Tα3, Tα5, and Tα7 human cells</td>
<td>240 lncRNAs associated with autoimmune disease (AID) lncRNA-seq</td>
<td>Analysis of the expression profile of the AID-associated lncRNAs</td>
</tr>
<tr>
<td>CD4+CD6−, CD4+CD8−, activated CD4+ mouse Th cells</td>
<td>31423 lncRNAs (lncRNA microarray)</td>
<td>Expression analysis and prediction of function</td>
</tr>
<tr>
<td>Th17 cell in-silico cell lines</td>
<td>Thyroid expression profiling of 10 thymus-specific lncRNAs</td>
<td>Enriched in human immune cells, acts as a cytoplasmic riboregulator that reduces the level of NFκB p65</td>
</tr>
<tr>
<td>Naive, memory, activated, non-activated mouse CD4+ T cells</td>
<td>Over 1000 mouse and human lncRNAs (microarray)</td>
<td>Expression and conservation analysis</td>
</tr>
<tr>
<td>CD4+CD6−, CD4+CD8−, mouse thymic T cells, and thymus-derived Tα5 cells, in vitro differentiated Tα1, Tα2, Tα7 and induced Tα5</td>
<td>1524 lncRNA genes (RNA-seq)</td>
<td>Expression analysis and ChIP-seq data analysis to identify lncRNA genes and possible regulators, Linc0-Cxx5.5 AS is Tα5, 2-specific and it reduces the expression of Cc17, Cc02, Cc05, and Cc05. It contributes to the migration of Tα7 cells</td>
</tr>
<tr>
<td>Infected Namalwa B lymphocytes</td>
<td>IFN-1 AS</td>
<td>Cytoplasmic post-transcriptional stabilisation of IFN-1 RNA making a miRNA-binding site</td>
</tr>
<tr>
<td>Jurkat cells, primary lymphocytes, lymphoma cell lines, CD19+ B cells</td>
<td>SalRAS-AS1</td>
<td>Regulates the alternative splicing of Fas which is impaired in non-Hodgkin lymphomas associated with poor prognosis</td>
</tr>
<tr>
<td>Activated human CD4+ T cells</td>
<td>BCR RNA (EST library analysis)</td>
<td>Proto-oncogene, induced upon activation, sensitive to immunosuppressive drugs</td>
</tr>
<tr>
<td>Jurkat cells</td>
<td>NROX1 (siRNA knock-down screening)</td>
<td>Regulates NRF2 subcellular localization as part of an RNA-protein complex</td>
</tr>
<tr>
<td>CEM-CCJ CMK1, Jurkat JRF1, human primary lymphocytes</td>
<td>GAS5</td>
<td>Necessary and sufficient for growth arrest. Acts competing from GREs</td>
</tr>
<tr>
<td>Human CD4+, CD8+ cells, PBMC</td>
<td>NIT</td>
<td>Unknown, it shows a similar expression pattern to IFN γR</td>
</tr>
<tr>
<td>Thymocytes</td>
<td>TEA</td>
<td>Instruct the activity of Iκa promoters and recombination</td>
</tr>
<tr>
<td>Human Tα5 cells</td>
<td>Nuc5/SACEPG/PF-5'NG-AS1</td>
<td>Dependent on STAT4, Tβ1, and NFκB. Contributes to IFN expression by binding HDAC and after TNFα stimulation</td>
</tr>
<tr>
<td>Human primary CD4+ and CD8+ T cells, primary and polarized (from CD4+ and CD8+ T)  Tα1, Tα2, Tα7 and Tα5 cells; neutrophils, basophils, CD8+ CM, B cells</td>
<td>GAS5A3. AS1</td>
<td>Specifically expressed in Tα5 cells</td>
</tr>
</tbody>
</table>

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-$\gamma$ 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 $T_H^1$-specific lincRNA, called linc-MAF-4, that we demonstrated to be involved in the maintenance of $T_H^1$ cell identity via an epigenetic-repression of MAF gene.
References


19. Chu C, Qu K, Zhong FL, Artandi SE, Chang HY. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-


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


47. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J,


56. Heward JA, Lindsay MA. Long non-coding RNAs in the reg-


35 (Web Server issue): W345-349.


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, Gian-


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


130. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katz-


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


174. Zhang Y, Zhang XO, Chen T, Xiang JF, Yin QF, Xing YH,


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

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

\textsuperscript{1}Istituto Nazionale Genetica Molecolare “Romeo ed Enrica Invernizzi”, 20122 Milano, Italy.
\textsuperscript{2}IRCCS Ca’ Granda Ospedale Maggiore Policlinico, 20122 Milan, Italy
\textsuperscript{3}These authors contributed equally to this work
* Correspondence: pagani@ingm.org, abrignani@ingm.org

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

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 types 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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{+} T\textsubscript{H}2 specific lincRNA - LincR-Ccr2-5’AS - involved in the regulation of CD4\textsuperscript{+} T\textsubscript{H}2 lymphocytes migration\textsuperscript{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 \textit{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 lincMAF-4, a chromatin associated CD4\textsuperscript{+} T\textsubscript{H}1 specific lincRNA, correlates inversely with the transcription factor MAF and that its down-regulation skews CD4\textsuperscript{+} T cell differentiation toward T\textsubscript{H}2 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\textsuperscript{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\textsuperscript{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\textsuperscript{13, 30} (Fig. 1a). In order to identify novel lincRNAs expressed in primary human lymphocytes, we used three \textit{de novo} transcriptome reconstruction strategies that are based on the combination of two different sequence mappers, TopHat and Star\textsuperscript{31, 32}, with two different tools for \textit{de novo} transcripts assembly, Cufflinks and Trinity\textsuperscript{33, 34}. LincRNAs were identified within the newly described transcripts exploiting the following process: \textit{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); \textit{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\textsuperscript{35}; \textit{iii}) exploitation of PhyloCSF, a comparative genomics method that assesses multispecies nucleotide sequence alignment based on a formal statistical comparison of phylogenetic codon models\textsuperscript{36}, which efficiently identifies non-coding RNAs as demonstrated by ribosome profiling experiments\textsuperscript{37}. Finally we defined a stringent \textit{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_{H1} 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_{H1} lincRNAs by RT-qPCR in a new set of independent samples of primary human CD4+ naïve, T_{reg} and T_{H1} cells, as well as in naïve CD4+ T cells that were activated in vitro and induced to differentiate toward T_{H1} or T_{H2} cells. Specific subset expression was confirmed for 90% of the CD4+ T_{H1} signature lincRNAs (Fig. 2d). Moreover, 90% of CD4+ T_{H1} signature lincRNAs that are expressed in resting CD4+ T_{H1} cells purified ex vivo, are highly expressed also in naïve CD4+ T cells differentiated under T_{H1} polarizing conditions in vitro, whereas they are poorly expressed in naïve CD4+ T cells that are differentiated towards T_{H2} 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_{H0} cells in vitro, when T_{H1}, T_{H2} and T_{H17} 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 T₉2

As lincRNAs have been reported to influence the expression of neighbouring genes, 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 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), a T₉1 signature lincRNA, localized 139.5 Kb upstream of the MAF gene. MAF encodes a transcription factor involved in T₉2 differentiation, which is also required for the efficient development of T₉17 cells and controls IL4 transcription in CD4⁺ T follicular helper cells. Our sequencing data showed that high expression of linc-MAF-4 correlates with low levels of MAF transcript in CD4⁺ T follicular helper cells and 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, 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 T₉1 or T₉2 cells. In details, Fig. 3b shows that in T lymphocytes differentiating towards T₉1 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 T₉2 cells, we found the opposite situation, both MAF transcript and protein levels increase constantly up to day 8 while linc-MAF4 remains constantly low (Fig. 3b and Supplementary Fig. 3c), similarly to what observed in CD4⁺ T lymphocytes differentiating towards T₉17 cells (Supplementary Fig. 3d).
We further characterized MAF transcriptional regulation by looking at H3K4 trimethylation (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_{H1} 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 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, 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 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 exploiting the LNCipedia database (http://www.lncipedia.org) 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 assess 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\textsuperscript{17,20,23,46,47}. In the mouse immune system it has been found that lincRNAs expression changes during naïve to memory CD8\textsuperscript{+} T cell differentiation\textsuperscript{24} and during naïve CD4\textsuperscript{+} T cells differentiation into distinct helper T cell lineages\textsuperscript{28}. We show in human primary lymphocytes that activation induced differentiation of CD4\textsuperscript{+} naïve T cells is associated with increased expression of lincRNAs belonging to the CD4\textsuperscript{+} T\textsubscript{H1} signature suggesting that upregulation of T\textsubscript{H1} lincRNAs is part of the cell differentiation transcriptional program. Indeed, linc-MAF-4, one of the T\textsubscript{H1} signature lincRNA, is poorly expressed in T\textsubscript{H2} cells and its experimental downregulation skews differentiating T helper cells toward a T\textsubscript{H2} 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\textsuperscript{48} and MEG3\textsuperscript{49}. We therefore provide a mechanistic proof of concept that lincRNAs can be important regulators of CD4\textsuperscript{+} 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\textsuperscript{+} 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 Ricovery 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). T₉₁ polarization was initiated with 10 ng/ml IL12 (R&D Systems) and T₉₂ neutralizing antibody anti-IL4 (2 mg/ml). T₉₂ 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 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 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) (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 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 non-coding 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 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\textsuperscript{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\textsuperscript{+}CM and CD5\textsuperscript{+}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 p-value 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 p-value 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 \(T_H\)1 or \(T_H\)2 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 \(T_H\)1 and a \(T_H\)2 cells. RNA for gene expression analysis of \(T_H\)1 and \(T_H\)2 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-\(\gamma\) and IL-13 production by \(T_H\)1 and \(T_H\)2 cells. The \(T_H\)1 and \(T_H\)2 datasets were ranked as log2 ratios of the expression values for each gene in the two conditions (\(T_H\)1/\(T_H\)2), and the most upregulated/downregulated genes (having log2 ratios ranging from |3| to |0.6|) were assigned to the \(T_H\)1 and \(T_H\)2 reference sets respectively.

Genes from the \(T_H\)1 gene list which were downregulated in a \(T_H\)1 vs. control-siRNA comparison and genes from the \(T_H\)2 gene list which were downregulated in a \(T_H\)2 vs. control-siRNA comparison were filtered out, obtaining a \(T_H\)1-specific gene set (74 genes) and a \(T_H\)2-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 log2 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+ \(T_H\)1 signature
lincRNAs specific primers were designed and 2.5 mg of CD4⁺ T₁, T₉ 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 T₁ cells were resuspended in RLN1 buffer (50 mM Tris-HCl pH 8, 140 mM NaCl; 1.5 mM MgCl₂, 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 MgCl₂, 0.5% NP-40) supplemented with RNase inhibitors for 10 minutes on ice. Chromatin was pelleted 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 T₉ cells were UV-crosslinked at 400 mJ/cm² 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<sub>H1</sub> and T<sub>H2</sub> 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 pelleted 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\textsuperscript{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.Biffò, 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 log2-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 Th1 or Th2 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 Th1 or Th2 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 Th1 and Th2 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 Th1 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.
<table>
<thead>
<tr>
<th>Subset</th>
<th>Purity (%)</th>
<th>Sorting phenotype</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4⁺ naïve</td>
<td>99.8 ± 0.1</td>
<td>CD4⁺ CCR7⁺ CD45RA⁺ CD45RO⁻</td>
<td>20061</td>
</tr>
<tr>
<td>CD4⁺ T_H1</td>
<td>99.9 ± 0.05</td>
<td>CD4⁺ CXCR3⁺</td>
<td>20855</td>
</tr>
<tr>
<td>CD4⁺ T_H2</td>
<td>99.7 ± 0.3</td>
<td>CD4⁺ CRTH2⁺ CXCR3⁻</td>
<td>19623</td>
</tr>
<tr>
<td>CD4⁺ T_H17</td>
<td>99.1 ± 1</td>
<td>CD4⁺ CCR6⁺ CD161⁺ CXCR3⁻</td>
<td>20959</td>
</tr>
<tr>
<td>CD4⁺ T_reg</td>
<td>99.0 ± 0.8</td>
<td>CD4⁺ CD127⁻ CD25⁺</td>
<td>21435</td>
</tr>
<tr>
<td>CD4⁺ T_CM</td>
<td>98.4 ± 2.8</td>
<td>CD4⁺ CCR7⁺ CD45RA⁻ CD45RO⁺</td>
<td>20600</td>
</tr>
<tr>
<td>CD4⁺ T_EM</td>
<td>95.4 ± 5.5</td>
<td>CD4⁺ CCR7⁻ CD45RA⁻ CD45RO⁺</td>
<td>19800</td>
</tr>
<tr>
<td>CD8⁺ T_CM</td>
<td>98.3 ± 0.8</td>
<td>CD8⁻ CCR7⁺ CD45RA⁻ CD45RO⁺</td>
<td>20901</td>
</tr>
<tr>
<td>CD8⁺ T_EM</td>
<td>96.8 ± 0.9</td>
<td>CD8⁻ CCR7⁻ CD45RA⁻ CD45RO⁺</td>
<td>21813</td>
</tr>
<tr>
<td>CD8⁺ naïve</td>
<td>99.3 ± 0.2</td>
<td>CD8⁺ CCR7⁺ CD45RA⁺ CD45RO⁻</td>
<td>20611</td>
</tr>
<tr>
<td>B naïve</td>
<td>99.9 ± 0.1</td>
<td>CD19⁺ CD5⁻ CD27⁻</td>
<td>21692</td>
</tr>
<tr>
<td>B memory</td>
<td>99.1 ± 0.8</td>
<td>CD19⁺ CD5⁺ CD27⁺</td>
<td>21239</td>
</tr>
<tr>
<td>B CD5⁺</td>
<td>99.1 ± 0.8</td>
<td>CD19⁺ CD5⁺</td>
<td>22499</td>
</tr>
</tbody>
</table>
**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**
  - TopHat
  - Star
  - Star
- **identification of new transcripts**
  - Cufflinks
  - Cufflinks
  - Trinity+PASA

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Selection of transcripts &gt; 200 nt</th>
<th>Selection of multiexonic transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are they long transcripts?</td>
<td>SIZE CUTOFF</td>
<td></td>
</tr>
<tr>
<td>NO transcriptional noise</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No transcripts annotated</td>
<td>Filter out all annotated transcripts</td>
<td></td>
</tr>
<tr>
<td>NO coding potential</td>
<td>Known protein domain filter: PFAM DB using HMMER-3</td>
<td>Coding potential filter using PhyloCSF</td>
</tr>
<tr>
<td>Intergenic location</td>
<td>Selection of intergenic transcripts</td>
<td></td>
</tr>
</tbody>
</table>

**RNA-seq data**

- 4201 already annotated lincRNA genes
- 4764 lincRNA genes

**Human lincRNA catalog**
- Ensembl database v.67
- GENCODE v.12
- 11170 genes

**Human lincRNA catalog**
- 2497
- 643
- 1061

**Ensembl database v.67**
- 168
- 186
- 232

**GENCODE v.12**
- 4764 lincRNA genes

**563 newly described lincRNA genes**
- identified in at least 2 out of 3 approaches
Figure 2

a

lincRNA genes

<table>
<thead>
<tr>
<th></th>
<th>CD4 naïve</th>
<th>CD4^+ T,1</th>
<th>CD4^+ T,2</th>
<th>CD4^+ T,17</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4 naïve</td>
<td>379</td>
<td>305</td>
<td>190</td>
<td>384</td>
</tr>
<tr>
<td>CD4^+ T,1</td>
<td>261</td>
<td>206</td>
<td>305</td>
<td>231</td>
</tr>
<tr>
<td>CD4^+ T,2</td>
<td>272</td>
<td>356</td>
<td>358</td>
<td>312</td>
</tr>
<tr>
<td>CD4^+ T,17</td>
<td>312</td>
<td>312</td>
<td>312</td>
<td>312</td>
</tr>
</tbody>
</table>

“Protein coding” genes

<table>
<thead>
<tr>
<th></th>
<th>CD4 naïve</th>
<th>CD4^+ T,1</th>
<th>CD4^+ T,2</th>
<th>CD4^+ T,17</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4 naïve</td>
<td>543</td>
<td>446</td>
<td>780</td>
<td>407</td>
</tr>
<tr>
<td>CD4^+ T,1</td>
<td>504</td>
<td>462</td>
<td>323</td>
<td>475</td>
</tr>
<tr>
<td>CD4^+ T,2</td>
<td>504</td>
<td>504</td>
<td>323</td>
<td>475</td>
</tr>
<tr>
<td>CD4^+ T,17</td>
<td>504</td>
<td>504</td>
<td>323</td>
<td>475</td>
</tr>
</tbody>
</table>

Normalized gene expression

<table>
<thead>
<tr>
<th></th>
<th>B naïve</th>
<th>B memory</th>
<th>B naïve</th>
<th>B memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>B naïve</td>
<td>458</td>
<td>345</td>
<td>483</td>
<td>345</td>
</tr>
<tr>
<td>B memory</td>
<td>588</td>
<td>1558</td>
<td>1558</td>
<td>514</td>
</tr>
<tr>
<td>pan B</td>
<td>1508</td>
<td>816</td>
<td>635</td>
<td>674</td>
</tr>
<tr>
<td>cluster A</td>
<td>1558</td>
<td>816</td>
<td>635</td>
<td>674</td>
</tr>
<tr>
<td>cluster B</td>
<td>1558</td>
<td>816</td>
<td>635</td>
<td>674</td>
</tr>
<tr>
<td>cluster C</td>
<td>1558</td>
<td>816</td>
<td>635</td>
<td>674</td>
</tr>
<tr>
<td>cluster D</td>
<td>1558</td>
<td>816</td>
<td>635</td>
<td>674</td>
</tr>
<tr>
<td>cluster E</td>
<td>1558</td>
<td>816</td>
<td>635</td>
<td>674</td>
</tr>
<tr>
<td>cluster F</td>
<td>1558</td>
<td>816</td>
<td>635</td>
<td>674</td>
</tr>
<tr>
<td>cluster G</td>
<td>1558</td>
<td>816</td>
<td>635</td>
<td>674</td>
</tr>
<tr>
<td>cluster H</td>
<td>1558</td>
<td>816</td>
<td>635</td>
<td>674</td>
</tr>
</tbody>
</table>

b

lincRNA genes

<table>
<thead>
<tr>
<th></th>
<th>CD4 naïve</th>
<th>CD4^+ T,1</th>
<th>CD4^+ T,2</th>
<th>CD4^+ T,17</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4 naïve</td>
<td>543</td>
<td>446</td>
<td>780</td>
<td>407</td>
</tr>
<tr>
<td>CD4^+ T,1</td>
<td>504</td>
<td>462</td>
<td>323</td>
<td>475</td>
</tr>
<tr>
<td>CD4^+ T,2</td>
<td>504</td>
<td>504</td>
<td>323</td>
<td>475</td>
</tr>
<tr>
<td>CD4^+ T,17</td>
<td>504</td>
<td>504</td>
<td>323</td>
<td>475</td>
</tr>
</tbody>
</table>

“Protein coding” genes

<table>
<thead>
<tr>
<th></th>
<th>CD4 naïve</th>
<th>CD4^+ T,1</th>
<th>CD4^+ T,2</th>
<th>CD4^+ T,17</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4 naïve</td>
<td>543</td>
<td>446</td>
<td>780</td>
<td>407</td>
</tr>
<tr>
<td>CD4^+ T,1</td>
<td>504</td>
<td>462</td>
<td>323</td>
<td>475</td>
</tr>
<tr>
<td>CD4^+ T,2</td>
<td>504</td>
<td>504</td>
<td>323</td>
<td>475</td>
</tr>
<tr>
<td>CD4^+ T,17</td>
<td>504</td>
<td>504</td>
<td>323</td>
<td>475</td>
</tr>
</tbody>
</table>

Legend

- 73%
- 31%
- 40%
- 24%

4764 lincRNA genes

15911 “Protein coding” genes

2.5 -2.5

b

lincRNA genes

<table>
<thead>
<tr>
<th></th>
<th>CD4 naïve</th>
<th>CD4^+ T,1</th>
<th>CD4^+ T,2</th>
<th>CD4^+ T,17</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4 naïve</td>
<td>543</td>
<td>446</td>
<td>780</td>
<td>407</td>
</tr>
<tr>
<td>CD4^+ T,1</td>
<td>504</td>
<td>462</td>
<td>323</td>
<td>475</td>
</tr>
<tr>
<td>CD4^+ T,2</td>
<td>504</td>
<td>504</td>
<td>323</td>
<td>475</td>
</tr>
<tr>
<td>CD4^+ T,17</td>
<td>504</td>
<td>504</td>
<td>323</td>
<td>475</td>
</tr>
</tbody>
</table>

“Protein coding” genes

<table>
<thead>
<tr>
<th></th>
<th>CD4 naïve</th>
<th>CD4^+ T,1</th>
<th>CD4^+ T,2</th>
<th>CD4^+ T,17</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4 naïve</td>
<td>543</td>
<td>446</td>
<td>780</td>
<td>407</td>
</tr>
<tr>
<td>CD4^+ T,1</td>
<td>504</td>
<td>462</td>
<td>323</td>
<td>475</td>
</tr>
<tr>
<td>CD4^+ T,2</td>
<td>504</td>
<td>504</td>
<td>323</td>
<td>475</td>
</tr>
<tr>
<td>CD4^+ T,17</td>
<td>504</td>
<td>504</td>
<td>323</td>
<td>475</td>
</tr>
</tbody>
</table>

Legend

- 73%
- 31%
- 40%
- 24%

4764 lincRNA genes

15911 “Protein coding” genes

2.5 -2.5

C

1051 Receptors genes

6375 Metabolic process genes
Figure 3

(a) Heatmap showing gene expression levels in different human lymphocyte subsets (CD4+ naive, CD4+ Treg, CD8+ T, CD8+ naïve, CD4+ TH1, CD4+ TH2, CD4+ T cell memory, B naïve, B memory, B CD5+), lymph node, and various human non-lymphoid tissues. The expression levels are color-coded from green (low expression) to red (high expression).

(b) Similar heatmap to (a) focusing on specific lincRNAs: 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, and linc-LRRC49-4. The expression levels are color-coded from green (low expression) to red (high expression).

(c) Bar graph showing gene expression levels in human lymphocyte subsets (CD4+ naive, CD4+ Treg, CD4+ TH1), human lymphoid tissues, and human non-lymphoid tissues. The expression levels are log2 transformed and normalized. The error bars indicate the standard error.

(d) Log2 gene expression levels for the same lincRNAs as in (b) but across different time points (0, 4, 8, 12 days) for CD4+ naive and CD4+ Treg cells. The expression levels are color-coded from green (low expression) to red (high expression).

(e) Heatmap showing fold changes of gene expression for the same lincRNAs across different tissues and time points. The fold change is shown in red and green, with red indicating a significant increase and green indicating a decrease. The time points are 0, 4, 8, 12 days.
Figure 4
Figure 5

a

b

c

d

e

f

[Diagrams and graphs showing expression levels of MAF and linc-MAF-4, H3K4me3, RNA polymerase II, and gene expression changes in CD4+ TH1 and CD4+ TH2 cells after siRNA treatment.]
**Figure 6**

(a) Genomic location of linc-MAF-4, MAF, and their respective control regions. The chr16 bp range is depicted from 79,800,000 to 79,600,000. The frequency of interactions is shown across different locations.

(b) Relative percentage of total relative expression levels for linc-MAF-4, Linc00339, Malat1, and RNU2.1.

(c) Relative expression levels of LSD1 and EZH2 (fold enrichment) for ACTB, RNK21, Linc-MAF-4 control, and Linc-MAF-4.

(d) Percentage of EZH2, H3K27me3, and LSD1 across MAF with siRNA ctrl and siRNA linc-MAF-4.

(e) Molecular interactions and localization of EZH2, LSD1, and MAF in the context of the linc-MAF-4 region.
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_{H1}, T_{H2} and T_{H17} signature lincRNAs trends in CD4+ naïve T cells differentiated in T_{H0} 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 and 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 Th1, Th2 and Th0 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 Th17 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 naïve 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.
Supplementary Figure 5

Chromosome-conformation capture on \textit{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 Cufflinks, 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 specific 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 \( e^i \):

\[
S_i(e^i) = \frac{b(e^i) - a(e^i)}{\max(a(e^i), b(e^i))}
\]

where:

\[
a(e^i) = E(\text{Dist}(e^i, e^j)|e^i \in c^x \text{ and } e^j \in c^x),
\]

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:

\[
b(e^i) = \min_{c^x} E(\text{Dist}(e^i, e^j)|e^i \not \in c^x \text{ and } e^j \in c^x)
\]

\( b(e^i) \) 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 $p_1, p_2$, 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:

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

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

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 $e^2$, $e^i = (e_{i1}^1 \ldots e_{in}^i)$, was defined as

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

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.
### CD4+ naïve signature

<table>
<thead>
<tr>
<th>gene id</th>
<th>locus</th>
<th>strand</th>
<th>n° isoforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>INGME_00014</td>
<td>1:125394375-1254058</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>INGME_00006</td>
<td>17:18962904-1907615</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>INGME_00089</td>
<td>11:11228928-1127258</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>INGME_00048</td>
<td>2:97147907-2971409</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>INGME_00056</td>
<td>17:19038405-1905434</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>INGME_00034</td>
<td>5:91498669-9150054</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>INGME_00060</td>
<td>3:97147907-3971409</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>INGME_00033</td>
<td>17:20178811-2018808</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>INGME_00027</td>
<td>7:24943450-7494685</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>INGME_00015</td>
<td>1:229114082-229116130</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>INGME_00039</td>
<td>6:27187611-7218870</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>INGME_00009</td>
<td>12:125510477-125513897</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>INGME_00005</td>
<td>5:91498669-9150054</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>INGME_00027</td>
<td>7:24943450-7494685</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>INGME_00015</td>
<td>1:229114082-229116130</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>INGME_00039</td>
<td>6:27187611-7218870</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>INGME_00009</td>
<td>12:125510477-125513897</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>INGME_00005</td>
<td>5:91498669-9150054</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>INGME_00027</td>
<td>7:24943450-7494685</td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>

### CD4+ TH2 signature

<table>
<thead>
<tr>
<th>gene id</th>
<th>locus</th>
<th>strand</th>
<th>n° isoforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>INGME_00026</td>
<td>12:125510477-125513897</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>XLOC_001683</td>
<td>15:126567724-126618000</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>XLOC_00268</td>
<td>2:7512184-7513642</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>XLOC_00268</td>
<td>2:7860237-7865579</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>XLOC_00268</td>
<td>2:7797732-7811547</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>XLOC_00268</td>
<td>2:7860237-7865579</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>XLOC_00268</td>
<td>2:7797732-7811547</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>XLOC_00268</td>
<td>2:7860237-7865579</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>XLOC_00268</td>
<td>2:7797732-7811547</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>XLOC_00268</td>
<td>2:7860237-7865579</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>XLOC_00268</td>
<td>2:7797732-7811547</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

### CD4+ TH17 signature

<table>
<thead>
<tr>
<th>gene id</th>
<th>locus</th>
<th>strand</th>
<th>n° isoforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>INGME_00015</td>
<td>12:125510477-125513897</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>INGME_00015</td>
<td>12:125510477-125513897</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>INGME_00015</td>
<td>12:125510477-125513897</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>INGME_00015</td>
<td>12:125510477-125513897</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>INGME_00015</td>
<td>12:125510477-125513897</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>INGME_00015</td>
<td>12:125510477-125513897</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>INGME_00015</td>
<td>12:125510477-125513897</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>INGME_00015</td>
<td>12:125510477-125513897</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>INGME_00015</td>
<td>12:125510477-125513897</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>INGME_00015</td>
<td>12:125510477-125513897</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>INGME_00015</td>
<td>12:125510477-125513897</td>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>

Nature Immunology: doi:10.1038/ni.3093
Heatmaps of signature lncRNAs expression for each lymphocytes subset. For each lincRNA gene id, locus, strand prediction and number of isoforms are also reported. Right panel represents signature lncRNAs relative expression values in a panel of 16 human tissues (Human BodyMap 2.0 project).

Nature Immunology: doi:10.1038/ni.3093
<table>
<thead>
<tr>
<th>PRIMER ID</th>
<th>TARGET</th>
<th>SEQUENCE (5'-&gt; 3')</th>
<th></th>
<th>REFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>linc-MAF-4 F</td>
<td>XLOC_012017</td>
<td>GGCTACGTCTCCATTGTTTT</td>
<td></td>
<td>Rani et al., 2011</td>
</tr>
<tr>
<td>linc-MAF-4 R</td>
<td>XLOC_012017</td>
<td>TGGTGTTTGGGAGCATTTGTT</td>
<td></td>
<td>Rani et al., 2011</td>
</tr>
<tr>
<td>T6F</td>
<td>ENSG00000257860</td>
<td>TTTTCATGGTGAGGAGAATGG</td>
<td></td>
<td>Rani et al., 2011</td>
</tr>
<tr>
<td>T6R</td>
<td>ENSG00000257860</td>
<td>CTGGGTCTTGGCCTCTTAAATG</td>
<td></td>
<td>Rani et al., 2011</td>
</tr>
<tr>
<td>T8F2</td>
<td>INGMG_000772</td>
<td>AGCCTGGGCTTTTGGAGTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T8R3</td>
<td>INGMG_000772</td>
<td>GGCTTTGCCAGGATCTCCACA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T21F2</td>
<td>ENSG00000234535</td>
<td>GAAATCCGCAATGAAGCGAGAAGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T21R2</td>
<td>ENSG00000234535</td>
<td>GTGACAAAGATAGAGAGGTTTGGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T24F1</td>
<td>XLOC_002906</td>
<td>GATCCTGGTCTTGGCTGGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T24R1</td>
<td>XLOC_002906</td>
<td>ACCTCTGTTATGTGTGATT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T27F1</td>
<td>ENSG00000253988</td>
<td>ACATAGGATGAGCTGGAAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T27R1</td>
<td>ENSG00000253988</td>
<td>TGAGAACATGCGCTCTTCTTGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T28F4</td>
<td>XLOC_013498</td>
<td>TACAGCCTGCCACCTATGGATT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T28R4</td>
<td>XLOC_013498</td>
<td>ATGCTTACAGGTAGAGTTT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T30F3</td>
<td>XLOC_012199</td>
<td>CTGGTGATACACTGTCTA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T30R3</td>
<td>XLOC_012199</td>
<td>GCTCAAGATCAAAGGGCTAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T31F1</td>
<td>XLOC_011294</td>
<td>TCGTGTTGGCGAGAGAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T31R1</td>
<td>XLOC_011294</td>
<td>AGTGTAGGAGGGCGAGTGT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>siRNA ID</th>
<th>TARGET</th>
<th>SEQUENCE (5'-&gt; 3')</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>T2_si1</td>
<td>XLOC_012017</td>
<td>GGACCAACCTCTTCTTTA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2_si2</td>
<td>XLOC_012017</td>
<td>GTACTGCAAAGGTCTAATA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2_si3</td>
<td>XLOC_012017</td>
<td>CCGCATACTTTTCAGACTT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2_si4</td>
<td>XLOC_012017</td>
<td>GCTTGAACACTCACAAAGAAA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PRIMER ID</th>
<th>TARGET</th>
<th>SEQUENCE (5'-&gt; 3')</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GAS1f</td>
<td>MAF-promoter</td>
<td>TTAAGTGCGATGCTATAAGTTT</td>
<td>Rani et al., 2011</td>
<td></td>
</tr>
<tr>
<td>GAS1r</td>
<td>MAF-promoter</td>
<td>GGGGAAGACCATTCTGAAGTG</td>
<td>Rani et al., 2011</td>
<td></td>
</tr>
<tr>
<td>IFNgf</td>
<td>IFN-promoter</td>
<td>AAGTATACCAGCAGCCAGAGGA</td>
<td>Rani et al., 2011</td>
<td></td>
</tr>
<tr>
<td>IFNgr</td>
<td>IFN-promoter</td>
<td>AGCTGATAGGTCAGAAAGGA</td>
<td>Rani et al., 2011</td>
<td></td>
</tr>
<tr>
<td>ILCRf</td>
<td>Internal loop control region</td>
<td>TGAGCAGAGAAATGCTAGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ILCRr</td>
<td>Internal loop control region</td>
<td>TCACAGGCTTTCTGTACC</td>
<td>Rani et al., 2011</td>
<td></td>
</tr>
<tr>
<td>MyoD1f</td>
<td>MyoD1 5’ regulatory region</td>
<td>ACGTGAGATTGATGTGAGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MyoD1r</td>
<td>MyoD1 5’ regulatory region</td>
<td>ATCGAAGATTGTGCTAAGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACTBcf</td>
<td>ACTB-promoter</td>
<td>AAAGAGGAGAGAGAGAGAGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACTBcr</td>
<td>ACTB-promoter</td>
<td>AAGCAGAGAGGAGAGGAGAGA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### 3C PRIMERS

<table>
<thead>
<tr>
<th>PRIMER ID</th>
<th>SEQUENCE (5'-&gt; 3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>GCAGAACTCGCCTAATGG</td>
</tr>
<tr>
<td>L1</td>
<td>TGATTAATGCTGGGTAAGG</td>
</tr>
<tr>
<td>L2</td>
<td>TTCAAGCTCCTTCCTCC</td>
</tr>
<tr>
<td>L3</td>
<td>GGTCTTCAATTTCAATAGCC</td>
</tr>
<tr>
<td>L4</td>
<td>CCAATTGGGAAGTCGAAGGC</td>
</tr>
<tr>
<td>L5</td>
<td>ACTGCCCTTCAAGTCCTTGC</td>
</tr>
<tr>
<td>L6</td>
<td>ACAGGGAGAGCTGAGCCTT</td>
</tr>
<tr>
<td>L7</td>
<td>ATTGAAGCCATGTTTTTAAG</td>
</tr>
<tr>
<td>L8</td>
<td>ACTGCATGGCATTGGTCTGG</td>
</tr>
<tr>
<td>L9</td>
<td>CTTTTTCCGTAGTAGAGCC</td>
</tr>
<tr>
<td>L10</td>
<td>TCTCTGCTGACAGTGCTACC</td>
</tr>
<tr>
<td>L11</td>
<td>GTACAGCGAGCTCCACAAAG</td>
</tr>
<tr>
<td>L12</td>
<td>ATACATATTGGGAGGCCCTGG</td>
</tr>
<tr>
<td>L13</td>
<td>GCTGCAAATCTTGGGATTG</td>
</tr>
<tr>
<td>L14</td>
<td>GCTGAGGTCACAGAGCTAGG</td>
</tr>
<tr>
<td>L15</td>
<td>TGCAGGCTCACAATAAAC</td>
</tr>
<tr>
<td>L16</td>
<td>AGTACAGTGGCCTCCTT</td>
</tr>
<tr>
<td>L17</td>
<td>TTTGGGTGTTCGTGGSATC</td>
</tr>
<tr>
<td>L18</td>
<td>TGCCTTGTAGTGCTAGAG</td>
</tr>
<tr>
<td>L19</td>
<td>AGGCCCTGCAATATGCAAC</td>
</tr>
<tr>
<td>L20</td>
<td>TCCAGCCAGGGCATCAAATC</td>
</tr>
<tr>
<td>L21</td>
<td>ACACCCACCAAATTTTG</td>
</tr>
<tr>
<td>L22</td>
<td>ATAGCAGCTGTCTGTATC</td>
</tr>
<tr>
<td>L23</td>
<td>CCTATCACGTCTTCTGAG</td>
</tr>
<tr>
<td>L24</td>
<td>AGGCCAAACGTTAGTGTT</td>
</tr>
</tbody>
</table>

### RIP PRIMERS

<table>
<thead>
<tr>
<th>PRIMER ID</th>
<th>TARGET</th>
<th>SEQUENCE (5'-&gt; 3')</th>
<th>REFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actin_sy-F2</td>
<td>β-actin</td>
<td>CATCCTCACCTCAAGTACC</td>
<td></td>
</tr>
<tr>
<td>Actin_sy-R2</td>
<td>β-actin</td>
<td>CACGCAGCTTCATTGAG</td>
<td></td>
</tr>
<tr>
<td>LincM_pr-F1</td>
<td>linc-MAF-4 (control)</td>
<td>AGGTCATGAGCGAGAAAG</td>
<td></td>
</tr>
<tr>
<td>LincM_pr-R1</td>
<td>linc-MAF-4 (control)</td>
<td>TCCCTTGGGAGTAAAAAC</td>
<td></td>
</tr>
<tr>
<td>HOTAIR/H2-F</td>
<td>HOTAIR/H2</td>
<td>GGTAGAAAAAGCAACCAGAAAG</td>
<td>Tsai et al., 2010</td>
</tr>
<tr>
<td>HOTAIR/H2-R</td>
<td>HOTAIR/H2</td>
<td>ACATAAACCTCGTCTAGTGGG</td>
<td>Tsai et al., 2010</td>
</tr>
</tbody>
</table>
References


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


miRiadne: a web tool for consistent integration of miRNA nomenclature

Raoul J. P. Bonnal1,*,†, Riccardo L. Rossi1,*,†, Donatella Carpi1, Valeria Ranzani1, Sergio Abrignani1 and Massimiliano Pagani1,2,*

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 platforms, even from the same vendor. Published miRNA lists are thus generated with different profiling platforms 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 convert 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 translate 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 platforms 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 according 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 cannot 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) cannot 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.
miRiadne

Rosetta Stone
- Translation
  - 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
- Overlap
  - identification of common miRNA probes between platforms, shared miRNAs between miRBase versions or between platforms and miRBase versions

Rosetta Data
- Intersect
  - comparison of miRNA datasets to find the commonly detected miRNAs
- Update
  - conversion 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)

Time Warp
- evolution of miRNA annotations across miRBase versions

Figure 1
Figure 2

(a) Original name list

Vickers

129

Skogberg

16 52

hsa-let-7b
hsa-miR-23a*
hsa-miR-425*
hsa-miR-766
hsa-miR-923

hsa-miR-188-5p
hsa-miR-422a
hsa-miR-423-5p
hsa-miR-483-5p
hsa-miR-572
hsa-miR-601

Vickers

126

Skogberg

15 48

hsa-let-7b-5p
hsa-miR-23a-5p
hsa-miR-425-3p
hsa-miR-766-3p

(b) Updated name list

<table>
<thead>
<tr>
<th>First Platform Name</th>
<th>Second Platform Name</th>
<th>miRBase Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-miR-7b</td>
<td>hsa-miR-7b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa-miR-188-5p</td>
<td>hsa-miR-188-5p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa-miR-23a*</td>
<td>hsa-miR-23a*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa-miR-425*</td>
<td>hsa-miR-425*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa-miR-483-5p</td>
<td>hsa-miR-483-5p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa-miR-572</td>
<td>hsa-miR-572</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa-miR-601</td>
<td>hsa-miR-601</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa-miR-766</td>
<td>hsa-miR-766</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa-miR-923</td>
<td>hsa-miR-923</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa-let-7b-5p</td>
<td>hsa-let-7b-5p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa-miR-23a-5p</td>
<td>hsa-miR-23a-5p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa-miR-425-3p</td>
<td>hsa-miR-425-3p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa-miR-766-3p</td>
<td>hsa-miR-766-3p</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 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 extracted 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 latest 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 message is displayed 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 displays 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 be displayed (as it is the case of hsa-miR-801 in the figure) (9).
<table>
<thead>
<tr>
<th>Platform name</th>
<th>Release</th>
<th>Tecnology</th>
<th>miRBase</th>
<th># miRNA</th>
<th>Seq by vendor</th>
<th>Seq by miRBase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affymetrics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneChip miRNA Array</td>
<td>1</td>
<td>microarray</td>
<td>11</td>
<td>847</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>GeneChip miRNA Array</td>
<td>2</td>
<td>microarray</td>
<td>15</td>
<td>1105</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>GeneChip miRNA Array</td>
<td>3</td>
<td>microarray</td>
<td>17</td>
<td>1733</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>GeneChip miRNA Array</td>
<td>4</td>
<td>microarray</td>
<td>20</td>
<td>2578</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Agilent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human miRNA Microarray</td>
<td>2</td>
<td>microarray</td>
<td>10,1</td>
<td>719</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Human miRNA Microarray</td>
<td>3</td>
<td>microarray</td>
<td>12</td>
<td>888</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Human miRNA Microarray</td>
<td>14</td>
<td>microarray</td>
<td>14</td>
<td>887</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Human miRNA Microarray</td>
<td>16</td>
<td>microarray</td>
<td>16</td>
<td>1205</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Human miRNA Microarray</td>
<td>18</td>
<td>microarray</td>
<td>18</td>
<td>1887</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Human miRNA Microarray</td>
<td>19</td>
<td>microarray</td>
<td>19</td>
<td>2006</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>LifeTech (Applied Biosystems)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TaqMan Array Human microRNA</td>
<td>2</td>
<td>RT-qPCR</td>
<td>10</td>
<td>664</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>TaqMan Array Human microRNA</td>
<td>3</td>
<td>RT-qPCR</td>
<td>14</td>
<td>754</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>OpenArray Human microRNA</td>
<td>3</td>
<td>RT-qPCR</td>
<td>14</td>
<td>754</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>microRNA Ready-to-Use PCR panels</td>
<td>2</td>
<td>RT-qPCR</td>
<td>16</td>
<td>736</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>microRNA Ready-to-Use PCR panels</td>
<td>3</td>
<td>RT-qPCR</td>
<td>18</td>
<td>752</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>microRNA Ready-to-Use PCR panels</td>
<td>4</td>
<td>RT-qPCR</td>
<td>20</td>
<td>752</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>miRCURY LNA microRNA Array</td>
<td>10</td>
<td>microarray</td>
<td>10</td>
<td>722</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>miRCURY LNA microRNA Array</td>
<td>11</td>
<td>microarray</td>
<td>11</td>
<td>846</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>miRCURY LNA microRNA Array</td>
<td>5th gen</td>
<td>microarray</td>
<td>14</td>
<td>904</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>miRCURY LNA microRNA Array</td>
<td>6th gen</td>
<td>microarray</td>
<td>16</td>
<td>1218</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>miRCURY LNA microRNA Array</td>
<td>7th gen</td>
<td>microarray</td>
<td>20</td>
<td>2244</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Exiqon</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human microRNA expression panel</td>
<td>1</td>
<td>microarray</td>
<td>9</td>
<td>470</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Nanostring</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nCounter Human miRNA panel</td>
<td>2</td>
<td>microarray</td>
<td>18</td>
<td>812</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Nanostring NCouler Human</td>
<td>3</td>
<td>counter</td>
<td>21</td>
<td>829</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Illumina</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quanta BioSciences</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>miScript microRNA PCR Arrays</td>
<td>16</td>
<td>RT-qPCR</td>
<td>16</td>
<td>1075</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Qiagen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>qScript microRNA System</td>
<td>1</td>
<td>microarray</td>
<td>20</td>
<td>2578</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Wafergen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SmartChip Human MicroRNA Panel</td>
<td>1</td>
<td>microarray</td>
<td>16</td>
<td>1046</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Platform name</td>
<td>Release</td>
<td>Technology</td>
<td>miRBase</td>
<td># miRNA</td>
<td>Seq by vendor</td>
<td>Seq by miRBase</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>---------</td>
<td>------------</td>
<td>---------</td>
<td>---------</td>
<td>---------------</td>
<td>---------------</td>
</tr>
<tr>
<td><strong>Affymetrics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneChip miRNA Array mouse</td>
<td>2</td>
<td>microarray</td>
<td>15</td>
<td>717</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>GeneChip miRNA Array rat</td>
<td>2</td>
<td>microarray</td>
<td>15</td>
<td>387</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>GeneChip miRNA Array mouse</td>
<td>3</td>
<td>microarray</td>
<td>17</td>
<td>1111</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>GeneChip miRNA Array rat</td>
<td>3</td>
<td>microarray</td>
<td>17</td>
<td>679</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>GeneChip miRNA Array mouse</td>
<td>4</td>
<td>microarray</td>
<td>20</td>
<td>1908</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>GeneChip miRNA Array rat</td>
<td>4</td>
<td>microarray</td>
<td>20</td>
<td>728</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td><strong>Quanta BioSciences</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>qScript microRNA System mouse</td>
<td>1</td>
<td>microarray</td>
<td>20</td>
<td>561</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>qScript microRNA System rat</td>
<td>1</td>
<td>microarray</td>
<td>20</td>
<td>338</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td><strong>Exiqon</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>microRNA Ready-to-Use PCR panels Mouse</td>
<td>2</td>
<td>RT-qPCR</td>
<td>16</td>
<td>749</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>microRNA Ready-to-Use PCR panels Mouse</td>
<td>3</td>
<td>RT-qPCR</td>
<td>18</td>
<td>752</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>microRNA Ready-to-Use PCR panels Mouse</td>
<td>4</td>
<td>RT-qPCR</td>
<td>20</td>
<td>752</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td><strong>LifeTech</strong> (Applied Biosystems)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TaqMan Array Rodent microRNA</td>
<td>3</td>
<td>RT-qPCR</td>
<td>14</td>
<td>750</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>OpenArray Rodent microRNA</td>
<td>3</td>
<td>RT-qPCR</td>
<td>14</td>
<td>750</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Figure 2
Supplementary Figure 3

1. Select platform
2. Obtain AIF (Annotation Information File)
3. Isolate probe names
4. 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
5. Are there discrepancies between probe names and miRBase mirna names?
   - YES: Manually curate
   - NO: Accept names to sequences associations
6. Annotated AIF
7. DB of Annotated AIFs
Supplementary Figure 4

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

<table>
<thead>
<tr>
<th>Search Name</th>
<th>AppliedTLDA:2</th>
<th>miRBase v21.0</th>
<th>Alternative Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-let-7a</td>
<td>hsa-let-7a</td>
<td>hsa-let-7a-5p</td>
<td>hsa-let-7a-5p</td>
<td>UGAGGUAGUAGGUAGUAGUAGU</td>
</tr>
<tr>
<td>hsa-miR-342-5p</td>
<td>hsa-miR-342-5p</td>
<td>hsa-miR-342-5p</td>
<td></td>
<td>AGGGGCUACUCUGAUAGA</td>
</tr>
<tr>
<td>hsa-miR-423-5p</td>
<td>hsa-miR-423-5p</td>
<td>hsa-miR-423-5p</td>
<td></td>
<td>UGAGGGGCAGAGAGAGACAG</td>
</tr>
<tr>
<td>hsa-miR-516a-3p</td>
<td>hsa-miR-516a-3p</td>
<td>hsa-miR-516a-3p</td>
<td></td>
<td>ACUUCAGUAGUAUCGGAGAC</td>
</tr>
<tr>
<td>hsa-miR-16-1*</td>
<td>hsa-miR-16-1*</td>
<td>hsa-miR-16-1*</td>
<td></td>
<td>CCAGUAAUACCGUAGCGUA</td>
</tr>
<tr>
<td>hsa-let-7c</td>
<td>hsa-let-7c</td>
<td>hsa-let-7c-5p</td>
<td>hsa-let-7c-5p</td>
<td>UGAGGUAGUAGGUAGUAGU</td>
</tr>
<tr>
<td>hsa-miR-365</td>
<td>hsa-miR-365</td>
<td>hsa-miR-365a-3p</td>
<td>hsa-miR-365b-3p</td>
<td>UGAGGUAGUAGGUAGUAGU</td>
</tr>
<tr>
<td>hsa-let-100</td>
<td>hsa-let-100</td>
<td>hsa-let-100-5p</td>
<td>hsa-let-100-5p</td>
<td>UGAGGUAGUAGGUAGUAGU</td>
</tr>
<tr>
<td>hsa-miR-93b</td>
<td>hsa-miR-93b</td>
<td>hsa-miR-93b-3p</td>
<td>hsa-miR-93b-3p</td>
<td>UGAGGUAGUAGGUAGUAGU</td>
</tr>
<tr>
<td>hsa-miR-139-5p</td>
<td>hsa-miR-139-5p</td>
<td>Changed OR Dead</td>
<td></td>
<td>UGAGGUAGUAGGUAGUAGU</td>
</tr>
<tr>
<td>hsa-miR-30a*</td>
<td>hsa-miR-30a*</td>
<td>hsa-miR-30a-3p</td>
<td>hsa-miR-30a-3p</td>
<td>UGAGGUAGUAGGUAGUAGU</td>
</tr>
</tbody>
</table>

Download the CSV file.
Supplementary Figure 5

We are converting the list of miRNAs from Platform version AffyMirArray:2 to Platform version ExiqonHumanPanels:3.
Supplementary Figure 6

Original data matrix

Original data matrix
Supplementary Figure 7

The figure shows a screenshot of a web interface with a grid of miRNA entries. The interface includes a search bar and a list of miRNAs with their corresponding sequences. The table below the interface contains columns for the First Platform Search Name, Second Platform Search Name, miRBase Name, and Sequence. The table lists various miRNA entries, each with a two-letter code followed by a number and a complementary sequence.
References


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) \cite{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 com-
ponent 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.

<table>
<thead>
<tr>
<th>lncRNA</th>
<th>Diseases</th>
<th>Type</th>
<th>mRNA or loci affected</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBET</td>
<td>Facioscapulohumeral muscular dystrophy</td>
<td>IncRNA</td>
<td>4q35 locus</td>
</tr>
<tr>
<td>BACE1-AS</td>
<td>Alzheimer's disease</td>
<td>NAT</td>
<td>BACE1</td>
</tr>
<tr>
<td>DISC2</td>
<td>Schizophrenia</td>
<td>NAT</td>
<td>DISC1</td>
</tr>
<tr>
<td>HIF1A</td>
<td>Cancer, myocardi al ischaemia</td>
<td>NAT</td>
<td>HIF1A</td>
</tr>
<tr>
<td>MALAT1</td>
<td>Cancer</td>
<td>IncRNA</td>
<td>Many</td>
</tr>
<tr>
<td>ATXN8OS</td>
<td>Spinocerebellar ataxia</td>
<td>NAT</td>
<td>SCA8</td>
</tr>
<tr>
<td>FMR1</td>
<td>Fragile X syndrome</td>
<td>IncRNA</td>
<td>FMR1</td>
</tr>
<tr>
<td>FMR1-AS</td>
<td>Fragile X syndrome</td>
<td>NAT</td>
<td>FMR1</td>
</tr>
<tr>
<td>PINK1-AS</td>
<td>Parkinson's disease, diabetes</td>
<td>NAT</td>
<td>PINK1</td>
</tr>
<tr>
<td>CDKN2B-AS1</td>
<td>Cancer, diabetes, cardiovascular disease</td>
<td>IncRNA</td>
<td>CDKN2A, CDKN2B</td>
</tr>
<tr>
<td>NPPA-AS</td>
<td>Cardiovascular disease</td>
<td>NAT</td>
<td>NPPA</td>
</tr>
<tr>
<td>NAT-RAD18</td>
<td>Alzheimer's disease</td>
<td>NAT</td>
<td>RAD18</td>
</tr>
<tr>
<td>BOK-AS</td>
<td>Cancer</td>
<td>NAT</td>
<td>BOK</td>
</tr>
<tr>
<td>HTT-AS</td>
<td>Huntington's disease</td>
<td>NAT</td>
<td>HTT</td>
</tr>
<tr>
<td>HAP1R</td>
<td>Huntington's disease</td>
<td>NAT</td>
<td>HAP1R</td>
</tr>
<tr>
<td>P15-AS</td>
<td>Leukaemia</td>
<td>NAT</td>
<td>CDKN2B</td>
</tr>
<tr>
<td>lncRNA-p21</td>
<td>Cancer</td>
<td>IncRNA</td>
<td>CDKN1A</td>
</tr>
<tr>
<td>P21-AS</td>
<td>Cancer</td>
<td>NAT</td>
<td>CDKN1A</td>
</tr>
<tr>
<td>HOTAIR</td>
<td>Cancer</td>
<td>IncRNA</td>
<td>Many</td>
</tr>
<tr>
<td>LSINCT5</td>
<td>Cancer</td>
<td>IncRNA</td>
<td>Many</td>
</tr>
<tr>
<td>PTSC3</td>
<td>Cancer</td>
<td>IncRNA</td>
<td>Many</td>
</tr>
<tr>
<td>TUG1</td>
<td>Cancer</td>
<td>IncRNA</td>
<td>Many</td>
</tr>
<tr>
<td>lncRNA-EPS</td>
<td>Anaemia</td>
<td>IncRNA</td>
<td>Many</td>
</tr>
<tr>
<td>HELLPAR</td>
<td>HELLP syndrome</td>
<td>IncRNA</td>
<td>Many</td>
</tr>
<tr>
<td>UCA1</td>
<td>Cancer</td>
<td>IncRNA</td>
<td>Many</td>
</tr>
<tr>
<td>GAS5</td>
<td>Autoimmune disease, cancer</td>
<td>IncRNA</td>
<td>Many</td>
</tr>
<tr>
<td>DA125942</td>
<td>Brachydactyly type E</td>
<td>IncRNA</td>
<td>Many</td>
</tr>
</tbody>
</table>

Table 4. LncRNAs with potential roles in human diseases [37].
The potential role and involvement 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 TH1, TH2, TH17, and Treg cells was performed by Zhang et al., to identified TH2-specific markers relevant for allergy. They found that lncRNA GATA3-AS1 is specifically expressed in primary TH2, it is significantly correlated with GATA3 and it shares the same promoter. This lncRNA showed an increase expression in patients with TH2-associated disease (seasonal allergic rhinitis) compared to healthy control subjects. Then, lncRNA GATA3-AS1 is now considered a specific indicator of TH2-response and TH2-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 anti-sense 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 IncRNA play a critical role in B cell function and determine susceptibility to AITD [18]. Another example is IncRNA PRINS (Psoriasis Susceptibility-related non-coding RNA gene) that harbors two repeat elements [30]. This IncRNA 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 by Hrdlickova and colleagues, 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 IncRNAs and 626 protein coding genes. Comparing the mean expression levels of IncRNAs versus protein coding genes, only a two-fold lower expression of AID IncRNAs 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 de-repression. 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 IncRNAs 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 emerg-
ing 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 \textit{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 downregu-
lation 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


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


42. Yee F, Ericson H, Reis DJ, Wahlestedt C. Cellular uptake

