Identification and characterization of lncRNAs in the human immune system: a computational approach

PhD candidate: Alberto Arrigoni

PhD supervisor: Massimiliano Pagani

Affiliation: **Dottorato di Ricerca in Medicina Traslazionale e Molecolare (DIMET), Universitá degli studi di Milano Bicocca**

Table of contents

Chapter 1

Introduction

At the beginning of this century, the notion that the biological processes are orchestrated mainly by proteins has been challenged by the discovery that a large portion of the human transcriptome codes for transcripts that are not translated.

This gave rise to a transcriptional landscape that is radically different to what was previously believed: recent works estimate that against a total of 62.1% of the human genome covered by processed transcript (74.7% by primary transcripts), exons of protein-coding genes cover only the 2.94% of the genome [1].

In particular, long non-coding RNAs (lncRNAs) are defined as transcripts having low coding potential and longer than 200 bp [2]. The choice of this length threshold is somewhat arbitrary, but it is instrumental in order to separate lncRNAs from other non-coding RNA classes, such as microRNAs (miRNAs), short interfering RNAs (siRNAs), Piwi-interacting RNAs (piRNAs), small nucleolar RNAs (snoRNAs), and other short RNAs. A more comprehensive list of different non-coding RNA categories is reported in Table 1.

The majority of lncRNAs are transcribed, capped, spliced, and polyadenylated similarly to mRNAs [32]. They are defined after the transcriptional context they are located in, and in particular their nomenclature is defined after the position they assume relative to protein coding genes (fig.1). There are - intronic lncRNAs, which originate from intronic regions, and they do not overlap any annotated exon, - antisense lncRNAs, that span at least one exon of a nearby protein coding, and are transcribed in the opposite direction, - bidirectional lncRNAs are transcripts that initiate in a divergent fashion from the promoter of a protein-coding gene and - intergenic lncRNAs, whose transcriptional units do not overlap any annotated protein coding genes [31].

In the context of this thesis, the main focus will be on this last category, as the study of lincRNAs does not suffer from the complications arising from the presence of overlapping coding genes, which make them less likely to originate from transcriptional noise or errors in RNA-seq data assembly.

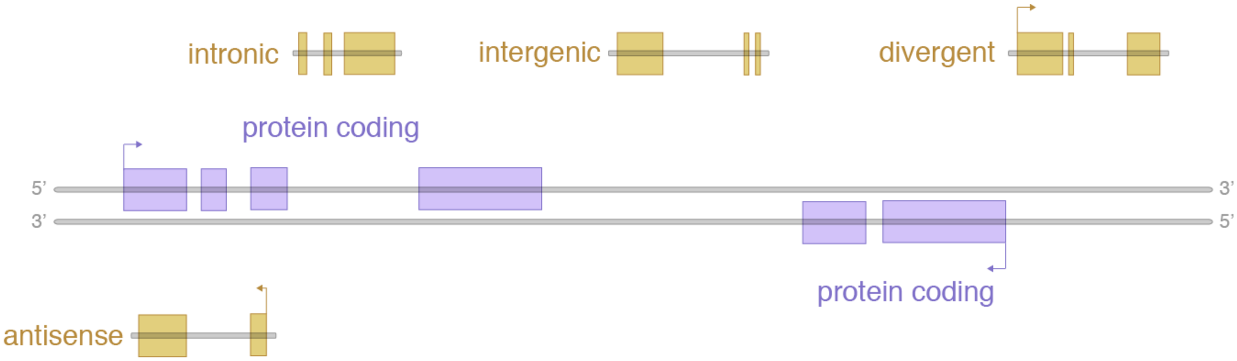


Fig.1: LncRNAs are defined after the position they assume relative to protein coding genes.

LncRNAs: mechanisms of action

The mechanisms of action of lncRNAs have been broadly categorized by Wang and Chang [33] into four archetypes that summarize what has been observed in recent years. According to this categorization, lncRNAs can act as signals, decoys, guides, and scaffolds (fig.2).

Signals: LncRNAs are transcribed by Pol II [32], as evidenced by the presence of 5' cap, polyadenylation and histone marks associated with Pol II elongation. Their expression is highly tissue/cell-specific if compared to protein coding transcripts [21,22], suggesting that lncRNAs expression is under tight transcriptional regulation. LncRNAs could therefore serve as *molecular signals*: this notion is supported by the fact that the expression of individual lncRNAs derives from the integration of molecular signals, environmental cues and regulatory processes.

LncRNAs of this archetype are regarded as markers of functionally significant biological events, such as Xist (responsible for silencing one of the two X chromosomes in female mammals) [34], HOTAIR (involved in gene-silencing of the HOXD locus by PRC2) [35] and LincRNA-p21 (promotes transcriptional repression in the canonical p53 pathway and induction of apoptosis) [36].

Decoys: lncRNAs may work as ‘‘decoys’’ by sequestering miRNAs and affecting their regulation of expressed genes. An example of this mechanism of action is provided by the tumor suppressor pseudogene PTENP1, whose 3' was found to bind the same set of regulatory miRNA sequences that normally target the tumor-suppressor gene PTEN [37].

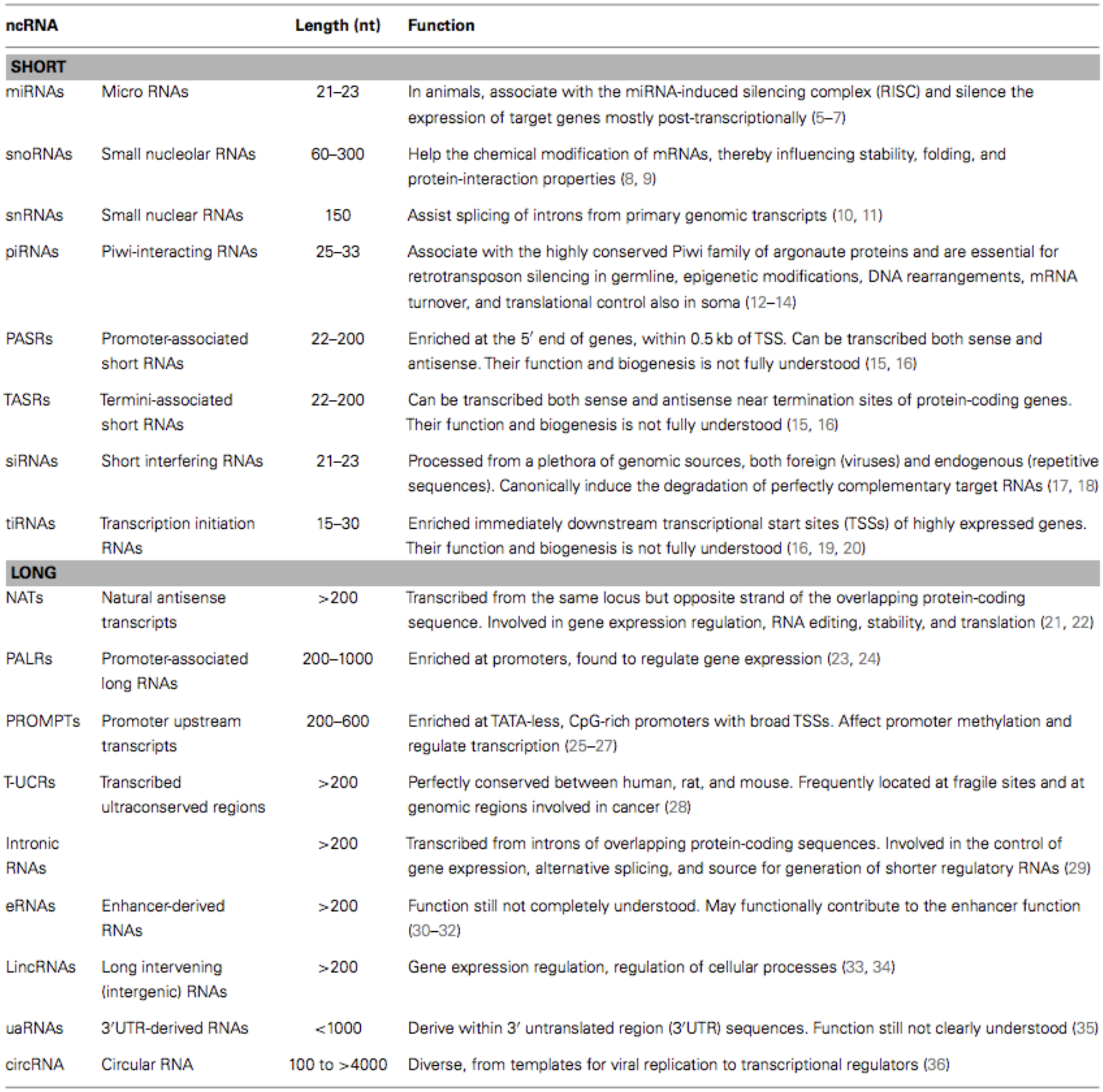
Guides: there is emerging evidence that lncRNAs can also act as 'guides' that direct the localization of ribonucleoprotein complexes to specific targets. Through the recruitment of regulatory complexes (both repressive (e.g., polycomb) and activating (MLL), lncRNAs can alter the epigenetic state of nearby regions (*cis*) (Xist [34], lnc-MAF-4 [22]), thereby influencing gene expression.

Moreover, lncRNAs can promote epigenetic change in *trans*: in this case target recognition can occur by lncRNA binding to target DNA as a RNA:DNA heteroduplex or as RNA:DNA:DNA triplex (HOTAIR [35], LincRNA-p21 [36]).

Scaffold: The fourth archetypal class of lncRNAs is the scaffolds. Individual lncRNA transcripts of this class may bind to different effectors at the same time, generating a network of interactors that promote change in the cell and orchestrate gene expression regulation.

As mentioned above, lncRNA HOTAIR promotes gene-silencing of the HOXD locus by interacting with PRC2. Additionally, it has been demonstrated that a portion of HOTAIR (700 nct of 3') also interacts with a second complex containing LSD1, that demethylates histone H3 on K4 in a complementary suppressive action.

It is therefore possible that lncRNA transcripts can fold to create local structural domains and rearrangements that specifically interact with different regulatory complexes to bring forth specific combinations of histone modifications on target gene chromatin.



ncRNA		Length (nt)	Function
SHORT			
miRNAs	Micro RNAs	21–23	In animals, associate with the miRNA-induced silencing complex (RISC) and silence the expression of target genes mostly post-transcriptionally (5–7)
snoRNAs	Small nucleolar RNAs	60–300	Help the chemical modification of mRNAs, thereby influencing stability, folding, and protein-interaction properties (8, 9)
snRNAs	Small nuclear RNAs	150	Assist splicing of introns from primary genomic transcripts (10, 11)
piRNAs	Piwi-interacting RNAs	25–33	Associate with the highly conserved Piwi family of argonaute proteins and are essential for retrotransposon silencing in germline, epigenetic modifications, DNA rearrangements, mRNA turnover, and translational control also in soma (12–14)
PASRs	Promoter-associated short RNAs	22–200	Enriched at the 5′ end of genes, within 0.5 kb of TSS. Can be transcribed both sense and antisense. Their function and biogenesis is not fully understood (15, 16)
TASRs	Termini-associated short RNAs	22–200	Can be transcribed both sense and antisense near termination sites of protein-coding genes. Their function and biogenesis is not fully understood (15, 16)
siRNAs	Short interfering RNAs	21–23	Processed from a plethora of genomic sources, both foreign (viruses) and endogenous (repetitive sequences). Canonically induce the degradation of perfectly complementary target RNAs (17, 18)
tiRNAs	Transcription initiation RNAs	15–30	Enriched immediately downstream transcriptional start sites (TSSs) of highly expressed genes. Their function and biogenesis is not fully understood (16, 19, 20)
LONG			
NATs	Natural antisense transcripts	>200	Transcribed from the same locus but opposite strand of the overlapping protein-coding sequence. Involved in gene expression regulation, RNA editing, stability, and translation (21, 22)
PALRs	Promoter-associated long RNAs	200–1000	Enriched at promoters, found to regulate gene expression (23, 24)
PROMPTs	Promoter upstream transcripts	200–600	Enriched at TATA-less, CpG-rich promoters with broad TSSs. Affect promoter methylation and regulate transcription (25–27)
T-UCRs	Transcribed ultraconserved regions	>200	Perfectly conserved between human, rat, and mouse. Frequently located at fragile sites and at genomic regions involved in cancer (28)
Intronic RNAs		>200	Transcribed from introns of overlapping protein-coding sequences. Involved in the control of gene expression, alternative splicing, and source for generation of shorter regulatory RNAs (29)
eRNAs	Enhancer-derived RNAs	>200	Function still not completely understood. May functionally contribute to the enhancer function (30–32)
LincRNAs	Long intervening (intergenic) RNAs	>200	Gene expression regulation, regulation of cellular processes (33, 34)
uaRNAs	3′UTR-derived RNAs	<1000	Derive within 3′ untranslated region (3′UTR) sequences. Function still not clearly understood (35)
circRNA	Circular RNA	100 to >4000	Diverse, from templates for viral replication to transcriptional regulators (36)

Table 1: Major classes of short and long regulatory non-coding RNAs [2]

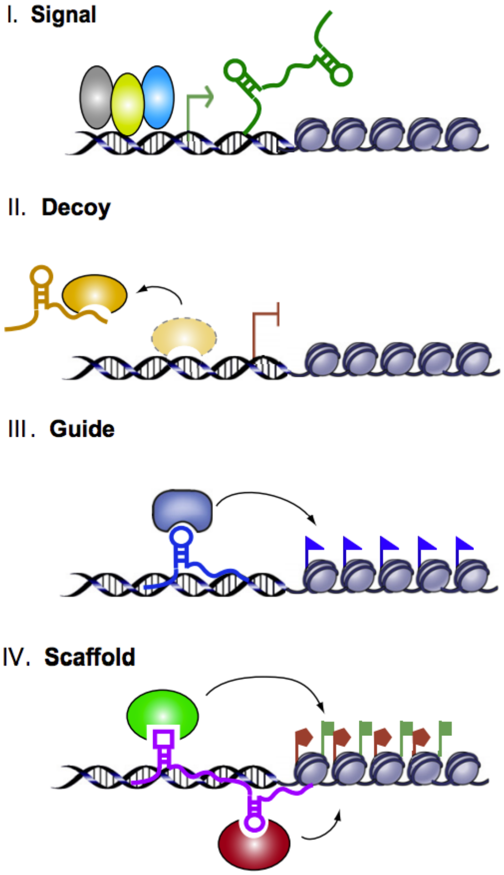


Fig.2: [33] Schematic Diagram of the Four Archetypes of LncRNA Mechanism

LncRNAs: cross-species conservation (or lack thereof)

In contrast to what has been reported for other non-coding RNA classes [42], long ncRNAs lack strong inter-species conservation. Overall, lncRNA sequences are less conserved than protein coding-genes, but more than introns or random intergenic regions [43].

Moreover, the evolutionary linkage between lncRNAs in different species is difficult to infer since the majority of approaches for conservation studies are based on primary sequence analysis. LncRNAs evolve more rapidly than protein coding genes (at a sequence level) because the evolutionary constraints at play are different: primary sequences of coding transcripts are translated to functional proteins, so that genomic mutations may alter the aminoacidic code that will shape the tertiary structure of the protein. Conversely, lncRNAs exert their functions by physically associating with proteins, and point mutations at a genomic level may not impact on the overall tridimensional structure and functionality of the molecule.

For these reasons, novel paradigms are needed to fully describe the multidimensional evolutionary linkage between lncRNAs in different species.

Sequence-structure-function relationship: Based on the assumption that most lncRNAs will fold into complex secondary and tertiary structure, conservation can be found at the structural level [44]. Algorithms that use alphabets to codify structural spatial rearrangements of RNA molecules could be used to identify evolutionary relationships when analysis of nucleotide sequence is not conclusive.

Another aspect that could be used to infer evolutionary relationships is functional conservation (for example, a lncRNA could be essential both in human and mouse), although the number of lncRNAs for which a functional characterization has been performed is too limited to exhibit a trend. Other approaches that have been proposed include conservation analysis of the promotorial regions [46] and analysis of syntenic transcription (this latter is used by LNCipedia [45] and by the Ensembl pipeline (www.ensembl.org) for lncRNAs annotation).

LncRNAs and the modulation of cellular plasticity in the immune system

LncRNAs have emerged as key regulators both for innate and adaptive immune response [41].

LincRNA-Cox2 [32] and THRIL [38] have been found to control TLR signaling in dendritic cells (DC) by activating and repressing gene transcription of critical immune response regulators and inflammatory-response genes.

In 2009, Guttman and colleagues identified 20 signal-specific lncRNAs in bone-marrow-derived *in vitro* DCs (BMDCs), 80% of which clustered with NF-kB signaling components in a systematic analysis of co-regulated genes across several datasets [32].

Furthermore, another lncRNA (lnc-DC) [39] has been found to be highly upregulated during DC differentiation and to be selectively expressed in classical antigen-presenting DCs (cDCs). Through knockdown experiments, it has been demonstrated that ln-DC controls expression levels of genes involved in T cell activation (CD40, CD80), antigen presentation (HLA-DR), and cytokine secretion.

LncRNAs have been shown to play an important role in the adaptive immune response as well. An example of this is lincR-Ccr2-50 AS [40], whose knockdown in mice leads to downregulation of key Th2 chemokine-encoding genes Ccr1, Ccr2, Ccr3, and Ccr5. Moreover, Hu et al. [40] demonstrate that the network of transcripts modulated by lincR-Ccr2-50 AS show considerable overlap with that of the ‘‘master’’ Th2 cell transcription factor GATA-3. This supports the idea that lncRNAs may represent a second layer of Th-cell-specific gene regulation downstream of master transcription factors.

Taken together, these results suggest that lineage specificity and dynamic expression of lncRNAs can be leveraged to gain insights on their role in immunological processes. Moreover, their unique characteristics would be useful in a therapeutic context, as the current picture of lncRNAs sees them as ‘drivers’ of regulatory processes that may be important to modulate in order to fight disease.

This is particularly true in the case of the adaptive immune system, where the heterogeneity of cellular populations and the complexity of immune processes require the targeting of biological ‘switches’ that are specifically expressed and that can have a significant impact on cells differentiation and survival.

Thus, the translational perspective of the work carried out in the context of this project (which is described in the ‘Conclusion’ section) is the determination of the the role of lncRNAs in peripheral tissue-resident regulatory T cells, and in particular those infiltrating the tumoral mass.

Regulatory T cells in nonlymphoid tissues

The presence of regulatory T cells has been documented in several nonlymphoid tissues both in human and mice: skin, intestinal mucosa, lung, liver, adipose tissue, autoimmune target tissues, infected tissues, grafts, placenta, tumors, atherosclerotic plaques and injured muscle are just some examples [refs]. The scarcity of studies addressing peripheral Treg cells recruitment and the role/function they play in these districts makes difficult to generalize and derive common characteristics for them. Nevertheless, what emerges from recent studies is that tissue-infiltrating Tregs retain overall suppressive functional characteristics (as evidenced by suppression assays) [48], and they are characterized by the expression of specific transcription factors, chemokine receptors or effector molecules which render them unique.

The set of tissue-resident Treg cells that have been characterized is limited to the ones found in mice Visceral Adipose Tissue (VAT) [48] and the regenerating skeletal muscle [49].

Peripheral Tregs studied from VAT exhibit upregulation of IL-10 and CTLA-4, suggesting that they can control conventional CD4+ T cell and CD8+ T cell populations in the adipose tissue. Moreover, they may also exert control on co-resident myeloid cells, as suggested by an inverse correlation between the frequency of Treg cells and that of proinflammatory myeloid populations.

Interestingly, VAT Treg cells activity is not limited to immunological processes. There is an increasing amount of evidence that VAT Tregs can exert control on metabolic indices, inhibiting local and systemic insulin resistance and glucose intolerance. This effect is partly exerted by direct action of Treg cells on adipocytes, as the IL-10 they produce can engage receptors on adipocytes to downregulate proinflammatory cytokines and increase glucose uptake.

Similarly, peripheral Treg cells accumulation in skeletal muscle after injuries seems to be instrumental to achieve a full recovery due to the non-immunological actions performed by Tregs, which overexpress amphiregulin, a growth factor that induces *in vitro* differentiation of satellite cells.

TCR repertoire of peripheral Treg cells: it has been suggested that the localization of Treg cells in nonlymphoid tissues may be facilitated by the fact that their TCR repertoire is biased toward the recognition of self antigens. The pool of peripheral Treg cells would then be considerably expanded in pathological situations such as inflammation or cancer in which an enhanced presentation of self antigen can occur in periphery [49]. It makes sense to think that tumors may leverage this 'failsafe' mechanism to build up defenses against the attack of effector lymphocytes by recruiting Treg cells inside the tumoral mass.

Transcriptomic analyses performed on peripheral Tregs (from the works referenced in this section) are all based on data produced with microarray technology, since no published RNA-seq dataset on peripheral Tregs is available at the time of writing (See Conclusion and future work)

Transcriptome profiling of tissue-resident Treg cells will represent an invaluable addition to the mass of experimental data that has already been produced on the subject, as it will uncover the features that make peripheral (and tumor infiltrating) Treg cells unique, and will pave the way to the development of *ad hoc* immunological treatments.

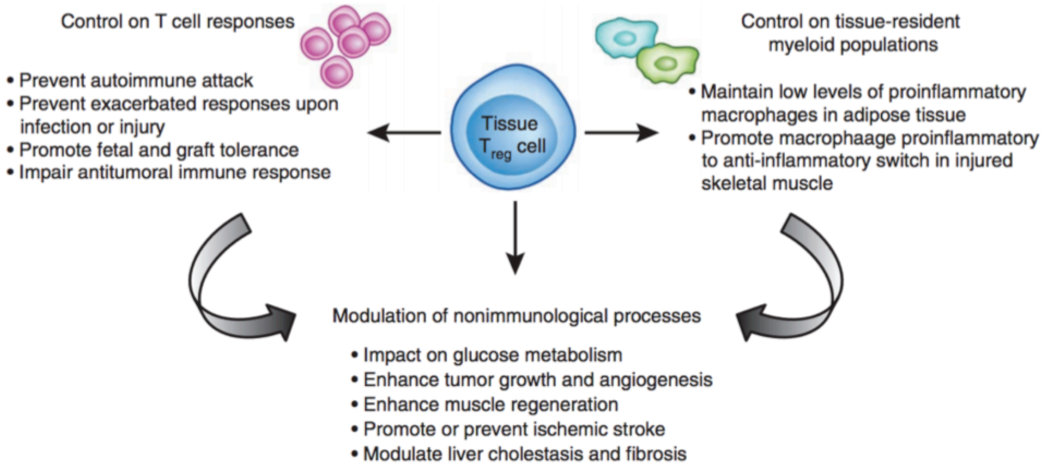

Control on T cell responses
• Prevent autoimmune attack
• Prevent exacerbated responses upon infection or injury
• Promote fetal and graft tolerance
• Impair antitumoral immune response
Tissue T_{reg} cell
Control on tissue-resident myeloid populations
• Maintain low levels of proinflammatory macrophages in adipose tissue
• Promote macrophaage proinflammatory to anti-inflammatory switch in injured skeletal muscle
Modulation of nonimmunological processes
• Impact on glucose metabolism
• Enhance tumor growth and angiogenesis
• Enhance muscle regeneration
• Promote or prevent ischemic stroke
• Modulate liver cholestasis and fibrosis

Fig.3 Functions of tissue-resident Treg cells [49]

RNA Next Generation Sequencing (NGS) and lncRNAs identification and characterization

The recent developments of 'omics' technologies for the study of gene expression profiles have made available new tools for the accurate identification and characterization of lncRNAs.

RNA-Seq is an approach to transcriptome profiling that uses deep-sequencing technologies to detect and accurately quantify RNA molecules originating from a genome at a given moment in time.

For this reason, RNA-seq (and in general techniques using short reads to attempt transcriptome reconstruction) is quantitative in nature, and it has been used in recent years to accurately identify differentially expressed genes for a vast range of applications (transcriptome reconstruction, variants identification, definition of splicing isoforms...). With the advent of RNA-seq, it is now possible to virtually reconstruct an entire transcriptome, with a greater dynamic range than microarrays and with the possibility to discover new loci and new transcripts [3].

A standard protocol for the identification and the characterization of novel lncRNAs from RNA-seq data is reported in fig.4.

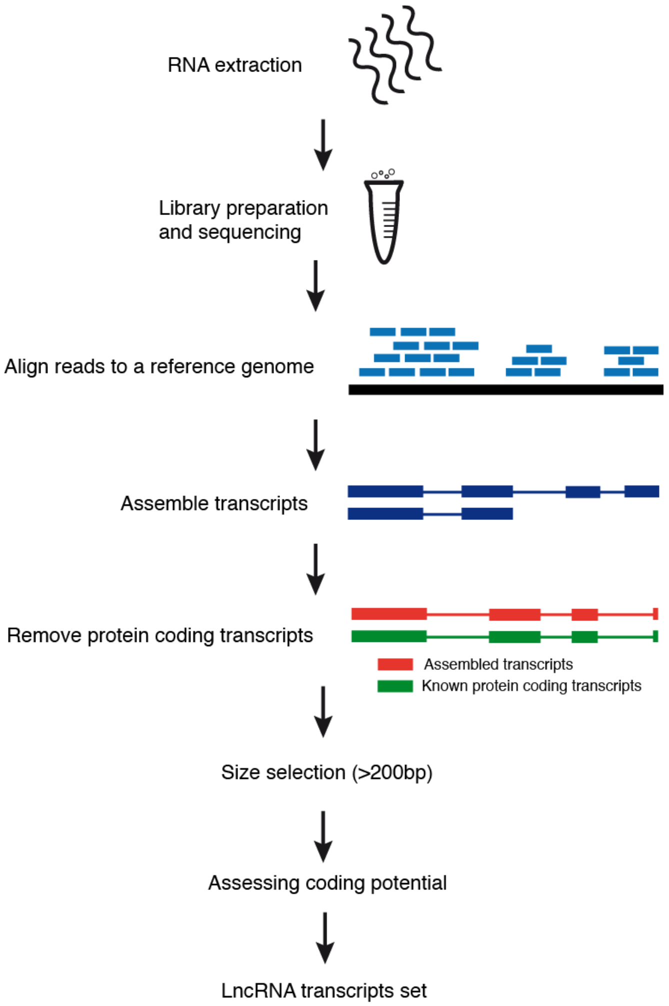

RNA extraction
Library preparation and sequencing
Align reads to a reference genome
Assemble transcripts
Remove protein coding transcripts
🟥 Assembled transcripts
🟩 Known protein coding transcripts
Size selection (>200bp)
Assessing coding potential
LncRNA transcripts set

Fig.4: Standard protocol for the identification of lncRNAs from RNA-seq data.

The following description of the protocol will not dwell on experimental considerations, as the main focus in the context of this thesis is the subsequent analysis of RNA-seq data and the computational strategies that must be implemented to reconstruct and characterize lncRNAs.

As listed in fig.4, the first step of the protocol is the selection of the polyA fraction from total purified RNA and the depletion of ribosomal RNA, followed by RNA fragmentation and (paired-end or single-end) sequencing.

Paired-end vs. single-end sequencing: Short-reads sequencing technologies are limited by the length of the reads that are generated (a range that goes from 50 to 130 bp is common for Illumina technology). This means that transcripts (or more precisely, the cDNA molecules resulting from the fragmentation of the entire transcriptome) are not sequenced in their entirety, but instead they need to be reconstructed by the means of 'aligner' software that leverages the sequence information contained in the genome reference (hg38 assembly is currently used for human) to stitch back the transcripts to their original form.

For this reason, the advantage of paired-end sequencing vs. single-end sequencing lies in the fact that both fragment ends are sequenced: this produces more alignable data and results in an overall higher probability to align to a reference genome. Moreover, paired-ends reads give precious information of the reciprocal distance of the same couple of reads, so that alternative splicing studies and the accurate detection of exon boundaries can be carried out in great detail.

Align reads to a reference genome: Since the number of reads that are typically produced in a RNA-seq experiment ranges from 10 to 100 millions, *ad hoc* software has been developed to perform mapping to a reference genome, such as Bowtie [5], Stampy [6], and GEM [8]. These tools achieve high performance by storing genomic coordinates of short oligomers as an indexed genome using hash table indexing and compression with a Burrows-Wheeler transform. The accuracy of the final alignments is influenced by several factors which need to be taken into account. For instance, some read matches may be missed as the genome search performed by this class of algorithms is not exhaustive; moreover, the intrinsic variability of RNA deriving from single nucleotide polymorphisms (SNPs) or small indels can compromise the detection of a read-reference match.

A number of reads derived from RNA fragments can map to two genomic regions with large gaps that span spliced introns. The accurate detection of novel intron-spanning junctions is important in the discovery of lncRNAs, and new programs that can adequately model the complexity of these transcripts have been released. These include STAR [7], MapSplice [9], GSNAP [10] and TopHat [11].

Though the overall mapping performances of these tools are comparable with one another, they are characterized by different requirements in terms of memory usage and execution time. For instance, GSNAP has been found [7] to have a poor mapped reads / hour ratio and large memory consumption, while STAR is the fastest mapper but also shows poor memory usage.

Although TopHat exhibits the least accurate performance, it maintains the advantages of its speed and its lack of requirement for large amounts of memory: it is therefore the mapper of choice when computing power is limited.

With the increase of computational capabilities of research institutes and the advent of cloud computing solutions for genomic studies, STAR has been increasingly adopted for the analysis of a vast number of RNA-seq datasets.

Reconstructing transcript models: In recent years, thousands of lncRNAs have been identified and characterized using RNA-seq [21,22,24]. The methods used belong to two broad categories: *reference-based* and *assembly-first* algorithms. With both these methods, it is possible to characterize alternative splicing events by analyzing the distribution of sequenced reads across splice junctions.

In a *reference-based* strategy, RNA-seq reads are first aligned to a reference genome using a splice-aware aligner, such as Tophat [11], GSNAP [10] or STAR [7]. Then, the overlapping reads from each locus are clustered to build a graph representing all possible isoforms, and in the final step the graph is traversed to resolve individual isoforms (fig.). Examples of software using this strategy are Cufflinks [18] and Scripture [24].

In particular, Cufflinks creates a graph with all the reads that align to a locus, and traverses it to assemble isoforms by finding the minimum set of transcripts that 'explain' the intron junctions found within the reads. Conversely, Scripture creates a splice graph containing each base of a chromosome and adds edges between bases if there is a read joining these two, and then traverses the graph in order to identify paths that are statistically significant (transcripts).

From this follows that Cufflinks is more conservative in the choice of transcripts to reconstruct, while Scripture may generate a larger set of novel transcripts.

The main advantage of a *reference-based* approach lies on the fact that the computational complexity of the transcript reconstruction problem is greatly reduced by the presence of a reference genome, so that this class of algorithms can be run on machines with only a few gigabytes of RAM.

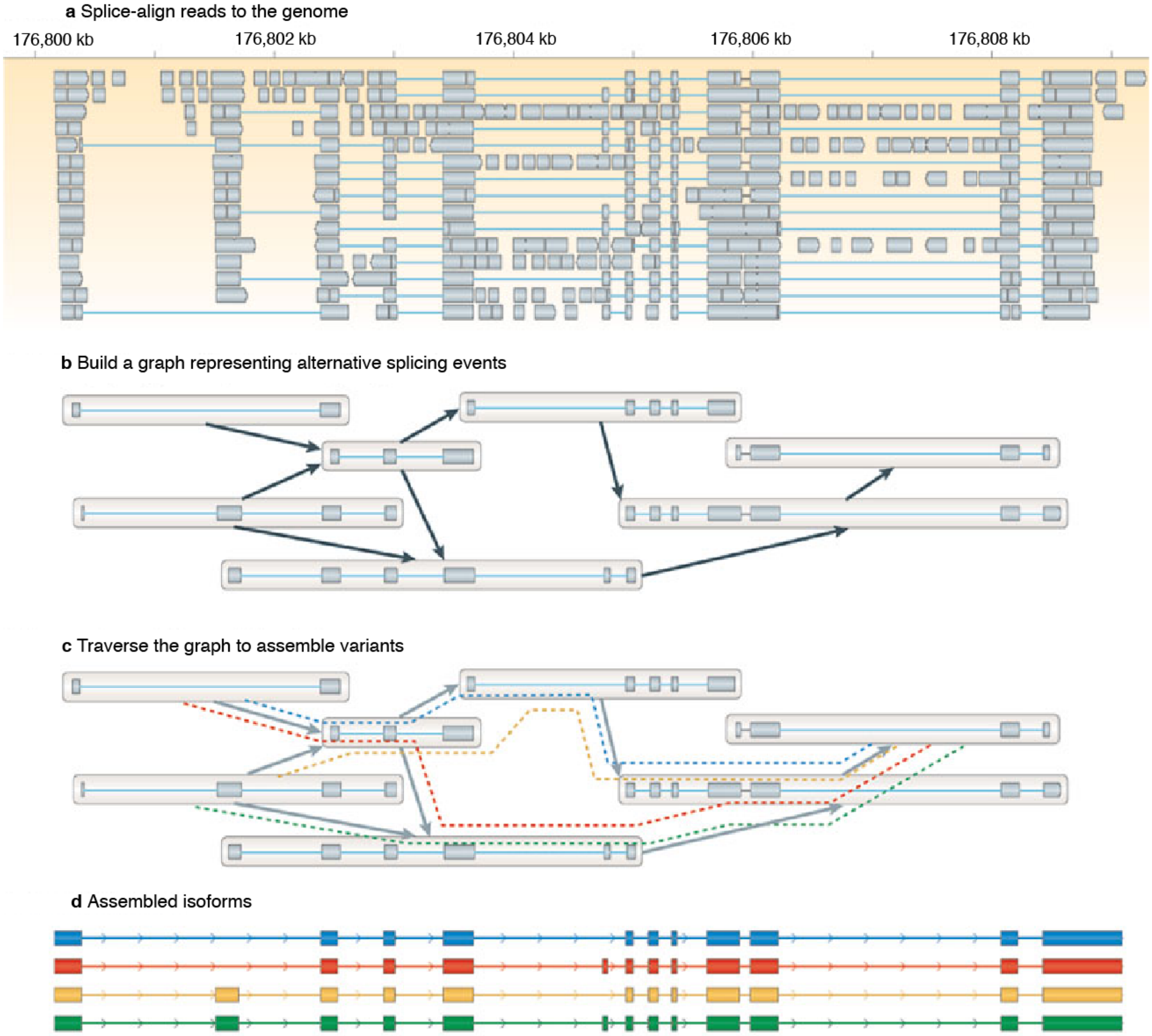

a Splice-align reads to the genome
176,800 kb
176,802 kb
176,804 kb
176,806 kb
176,808 kb
b Build a graph representing alternative splicing events
c Traverse the graph to assemble variants
d Assembled isoforms

Fig. 5: Overview of the reference-based transcriptome assembly. The steps of the reference-based transcriptome strategy are shown using as an example the maize gene (GRMZM2G060216) [23].

The *assembly-first* strategy (used by Trinity [25] and Oases [26] among the others) does not use the reference genome, but leverages the redundancy of short-read sequencing to find overlaps between the reads and assembles them into transcripts. This class of algorithms assembles the dataset multiple times using an approach based on the De Bruijn graph. Once constructed, the graph can be traversed directly to assemble each isoform.

In this approach, overlapping kmers of length *k* are generated from the reads: each kmer is assigned to a node in the De Brujin graph and edges are added between nodes if shifting a k-mer by one character creates an exact k-1 overlap between the two k-mers. Chains of adjacent nodes in the graph are collapsed into a single node when the first node has an out degree of one and the second node has an in degree on one. In the final step of the procedure the graph is traversed to assemble isoforms.

LncRNAs tend to be lowly expressed and *de novo* assemblers tend not to perform well across their expression range. It has been reported [20] that a map-first strategy using Cufflinks outperforms *de novo* assembly in terms of the number of Ensembl-annotated full-length transcripts that are reconstructed.

Nevertheless, assembly-first methodologies can be applied to cases where a genome reference is available in order to achieve higher sensitivity in the detection of novel transcripts [22]. A combined approach that leverages the strengths of both mapping-first and assembly-first algorithms may lead to a better characterization of the transcriptional landscape of the samples being analyzed.

Coding potential evaluation: a number of strategies have been developed in order to separate protein coding genes from lncRNAs in the pool of transcripts that are reconstructed by *de novo* methodologies on RNA-seq data. The performance and the accuracy of such strategies varies (as they are based on slightly different assumptions and features), so that an integrative approach that evaluates consensus results from different methods should be used to achieve high specificity of the classification.

A first step is taken to intersect newly reconstructed transcripts with protein domain database annotations such as PFAM [47]. Transcripts are translated in 6 frames (forward and reverse) in non strand-specific protocols to ensure that the ambiguity of strand attribution of *de novo* assemblers does not compromise the resolution power of the pipeline for the identification of novel lncRNAs. Transcripts presenting a PFAM match in one of the six frames are discarded from further selection.

Another (complementary) strategy that is used for lncRNAs identification is based on the use of machine learning algorithms that evaluate the rate of evolutionary divergence for codon models in transcripts sequences. Protein-coding regions are under strong evolutionary pressure to preserve amino acid content, and they can be identified by analyzing synonymous mutation patterns with programs such as Codon Substitution Frequency (CSF) [27]. Its most recent implementation

(PhyloCSF [27]) generates a probabilistic model that examines the over-representation of evolutionary signatures characteristic of alignments of conserved coding regions, such as the high frequencies of synonymous codon substitutions and conservative amino acid substitutions.

A major drawback of the PhyloCSF [27] approach is the lengthy running time (that can be close to days for thousands of transcripts on a medium-size HPC cluster), so that the classification step may become the bottleneck of the entire pipeline.

Other approaches have been recently developed for lncRNAs identification from RNA-seq data that rely on machine-learning algorithms and linguistic features analysis to provide reliable classification results and cut execution time by orders of magnitude. Among these iSeerna [29], CNCI [30] and CPAT [28] are increasingly used by the scientific community.

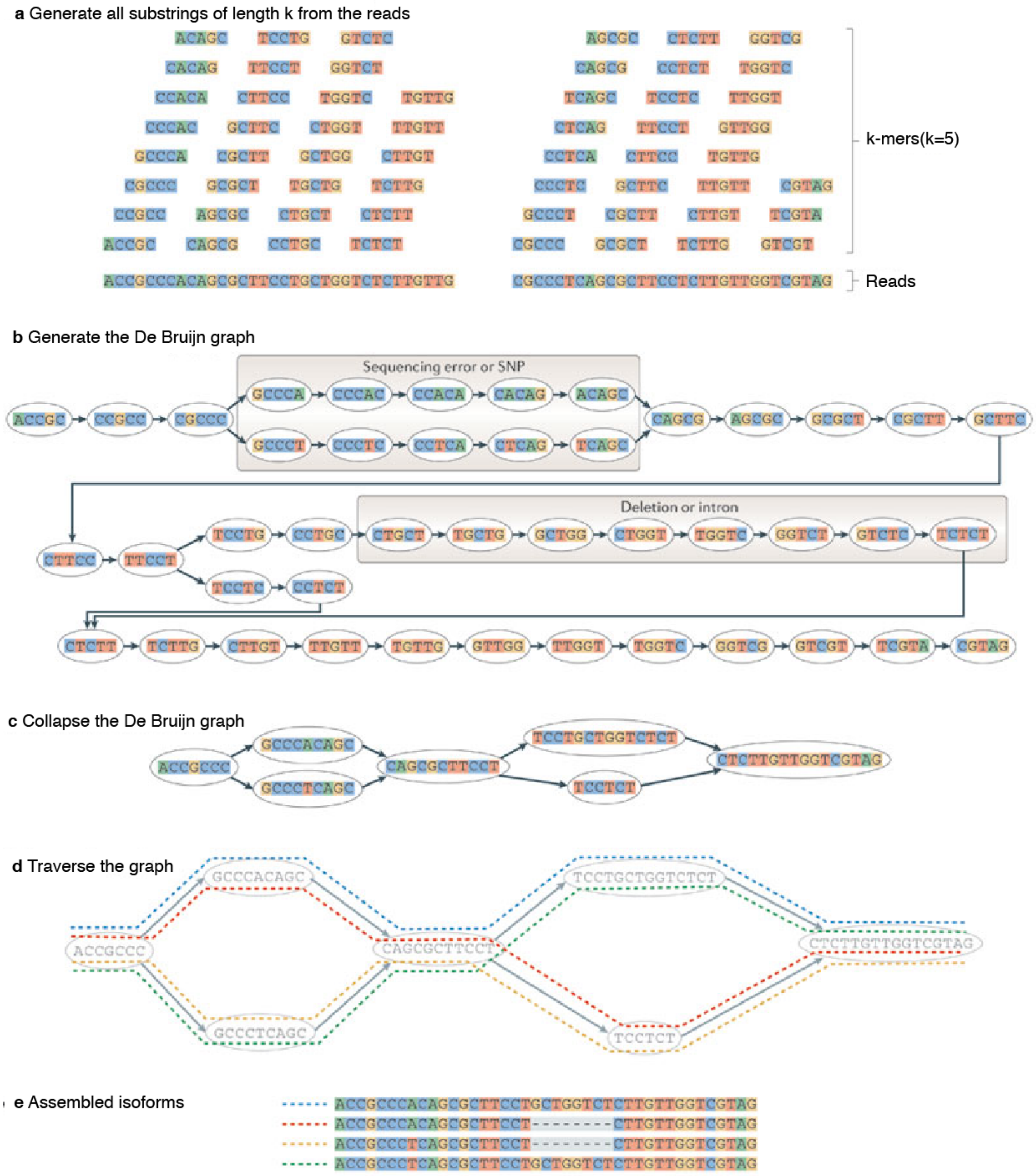
In particular, CPAT [28] uses a logistic regression model built on four sequence features: open reading frame (ORF) size, open reading frame coverage, Fickett TESTCODE statistic and hexamer usage bias.

The Fickett score is a linguistic feature that distinguishes protein-coding transcripts and lncRNAs according to the combinational effect of nucleotide composition and codon usage bias. The most discriminating feature though used by CPAT for classification is the hexamer score, which is related to the position of aminoacids in the three-dimensional structure of proteins. It evaluates the tendency of aminoacids (and hence, codons found in query transcripts) of shaping repetitions of patterns that are arranged to create protein structures.

For a given hexamer sequence S=H_1, H_2, ... , H_m, the hexamer score is calculated as:

$$HexamerScore\ = \frac{1}{m}\sum_{i=1}^{m} log(\frac{F(H_i)}{F'(H_i)})$$

which is the Log-likelihood ratio of the frequency of hexamers (64x64) under the coding (F) and noncoding (F') hypothesis.



a Generate all substrings of length k from the reads

b Generate the De Bruijn graph

c Collapse the De Bruijn graph

d Traverse the graph

e Assembled isoforms

k-mers(k=5)
Reads
ACCGCCCACAGCGCTTCCTGCTGGTCTCTTGTTG
CGCCCTCAGCGCTTCCTCTTGTTGGTCGTAG
Sequencing error or SNP
Deletion or intron
ACCGCCC
GCCCACAGC
GCCCTCAGC
CAGCGCTTCCT
TCCTGCTGGTCTCT
TCCTCT
CTCTTGTTGGTCGTAG
ACCGCCCACAGCGCTTCCTGCTGGTCTCTTGTTGGTCGTAG
ACCGCCCACAGCGCTTCCT--------CTTGTTGGTCGTAG
ACCGCCCTCAGCGCTTCCT--------CTTGTTGGTCGTAG
ACCGCCCTCAGCGCTTCCTGCTGGTCTCTTGTTGGTCGTAG

Fig.6: Overview of the *de novo* transcriptome assembly strategy [23].

RNA-seq data: a statistical introductory overview of differential expression analysis

A typical RNA-seq experiment produces from 10 to 100 million reads that are aligned back to the reference genome (if available), and inferences are performed to estimate the expression of the transcript that generated a particular set of reads, which is of course a measure proportional to the concentration of that transcript in the cell. If we consider K_i to be the number of sequencing reads that have been assigned to a particular genomic region *i*, than the problem can be formulated as a draw of a colored ball from an urn, where there are as many colors as regions *i* and the urn represents a large pool of RNA fragments.

The probabilities *p* for drawing a ball of each color from the urn are given by the frequencies according to which each region *i* is represented inside the pool. Probabilities *p* are not altered by the effect of subsequent draws, as the number of fragments needed for sequencing is much smaller than the total pool of RNA fragments in the pool.

If we consider the reads assigned to a single genomic region *Ki*, these are distributed as a binomial random variable with number of trials N (number of total reads) and probability of success *pi* (the relative frequency of fragments in the total pool arising from *i*). As *pi* shrinks and the number of trials N increases, the binomial distribution converges to a Poisson distribution with mean equal to Npi.

In RNA-seq experiments, the number of 'draws' is on average >10 millions and read counts for genomic region i are typically less than 1/1000. This theoretical demonstration has been empirically supported by Marioni et al. [12], who showed that Poisson distribution fits the expression data for 99.5% of the genes analyzed.

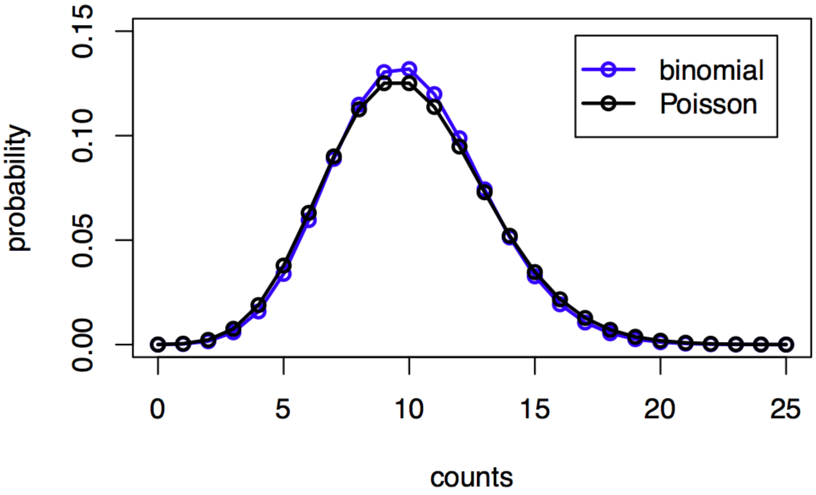


Fig.7 : The binomial distribution converges to the Poisson distribution for a large number of trials (N) and a small probability (p). Data shown are for a B=f(100,1/10) and P=f(10).

Overdispersion of counts data generated from RNA-seq experiments: While the assumptions presented in the last paragraph are well suited for technical replicates, when dealing with biological replicates we need to account for extra source of variability in the data.

This is due to the fact that when producing biological replicates a new pool of DNA fragments is generated, which will not have an identical probability vector $\vec{p}$ of relative frequencies from various regions of the genome.

Hence, these new experimental settings will be modeled after a mixture distribution, since a parameter of the Poisson distribution (the mean) varies according to another distribution. This other distribution was parameterized as a gamma distribution, as 1) it offers support for real non-negative numbers (counts data are of course non-negative integers) and 2) we can specify (at least) two parameters (both the mean and the variance).

When the variance of a Poisson mixture distribution is greater than the mean, counts are said to be 'overdispersed', and modern algorithm for the analysis of RNA-seq data have been implemented to model this variability and produce a reliable identification of differentially expressed genes.

Poisson/Gamma mixture and the negative binomial model: A Poisson/Gamma mixture distribution is most commonly referred to as the negative binomial distribution (a Poisson(λ) distribution, where λ is itself a random variable, distributed as a gamma distribution).

The density for negative binomial distribution K~NB(**μ,α**) with **μ**>0 and **α**>0 is defined as:

$$P(K = k) \ = \frac{\Gamma(k+1/\alpha)}{k!\Gamma(1/\alpha)}(\frac{\mu}{\mu+1/\alpha})^k(1+\mu\alpha)^{-1/\alpha}$$

the mean and the variance of this distribution are given by:

$E(K) = \mu \quad , \quad Var(K) = \mu + \alpha\mu^2$

The use of the negative binomial to study differential expression of genes from RNA-seq data was first introduced by Robinson et al [13], and Anders and Huber[14].

As $\alpha \to 0$, the negative binomial distribution converges to a Poisson distribution, so that it can also be used to model technical replication alone if needed.

Methodologies based on the read count negative binomial distribution assumption for differential expression analysis are EdgeR [13], DESeq (and DESeq2) [15], baySeq [16] and EBSeq [17]. The differential expression analysis performed with these software are all gene-based, while Cufflinks [see next section] also performs isoform-level analysis.

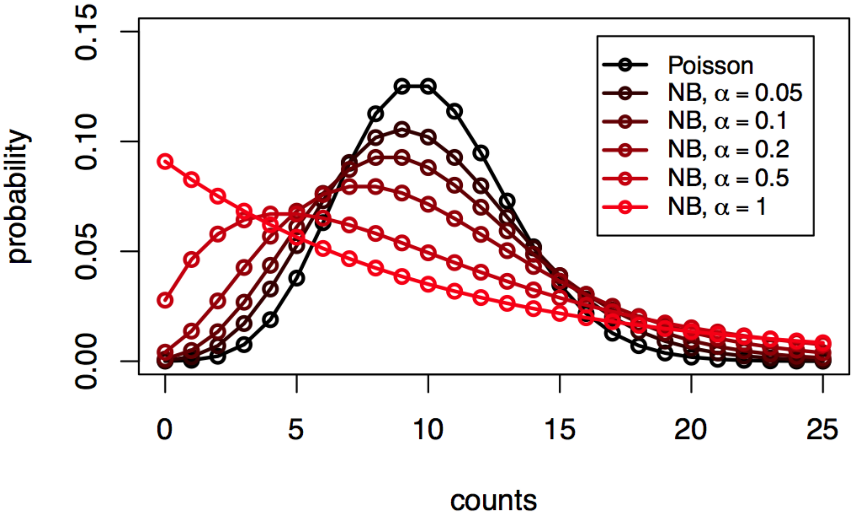


Fig.8 Negative binomial distribution $NB(\lambda, \alpha)$ with different α (variance) values. As α shrinks and tends to zero the distribution approximates a Poisson distribution $\sim P(\lambda)$.

Cuffdiff assumptions for differential expression: Cuffdiff 2 [18] estimates expression at transcript level resolution and controls for variability and read mapping ambiguity by using a beta negative binomial model for fragment counts. The algorithm combines the uncertainty in each transcript's fragment count with the overdispersion predicted to exist for that count by the global model of cross-replicate variability.

The final model (beta negative binomial) arises from the mixture distribution of different beta distributions.

If parameters of the beta distribution are $\alpha$and β, and if

$$X/p \sim NB(r, p)$$

where:

$$p \sim B(\alpha, \beta)$$

then the marginal distribution of X is a beta negative binomial distribution:

$$X \sim BNB(r, \alpha, \beta)$$

A comprehensive list of statistical assumptions used by the most widely used software for differential expression analysis is reported in Table 2 [19]:

Normalization of RNA-seq data: As the total number of reads that are generated in a RNA-seq experiment differs due to technical/biological variability, genes quantification will be affected.

A way to counteract this effect is to normalize gene expression data using a function of the library size from each sample by multiplying each gene read count value for a gene-specific normalization constant (s_{ij}). These normalization constants s_{ij} are considered constant within a sample, $s_{ij} = s_j$, and are estimated with the median-of-ratios method:

$s_j = \underset{i:K\neq 0}{median} \frac{Kij}{K_i}$ with $K_i \ (\prod_{j=1}^{m} \quad K_{ij})^{1/m}$

This normalization methodology was first implemented in DESeq2 [15] and it is currently the default setting for Cuffdiff (Cufflinks suite) [18].

Method	Version	Normalization	Read count distribution assumption	Differential expression test
edgeR	3.0.8	TMM/Upper quartile/RLE (DESeq-like)/None (all scaling factors are set to be one)	Negative binomial distribution	Exact test
DESeq	1.10.1	DESeq sizeFactors	Negative binomial distribution	Exact test
baySeq	1.12.0	Scaling factors (quantile/TMM/total)	Negative binomial distribution	Assesses the posterior probabilities of models for differentially and non-differentially expressed genes via empirical Bayesian methods and then compares these posterior likelihoods
NOIseq	1.1.4	RPKM/TMM/Upper quartile	Nonparametric method	Contrasts fold changes and absolute differences

				within a condition to determine the null distribution and then compares the observed differences to this null
SAMseq (samr)	2.0	SAMseq specialized method based on the mean read count over the null features of the data set	Nonparametric method	Wilcoxon rank statistic and a resampling strategy
Limma	3.14.4	TMM	voom transformation of counts	Empirical Bayes method
Cuffdiff 2 (Cufflinks)	2.0.2-beta	Geometric (DESeq-like)/quartile/classic-fpkm	Beta negative binomial distribution	*t*-test
EBSeq	1.1.7	DESeq median normalization	Negative binomial distribution	Evaluates the posterior probability of differentially and non-differentially expressed entities (genes or isoforms) via empirical Bayesian methods

Table 2: Comparison of software packages to detect differential expression [19]

Modeling gene-wise dispersion estimates: as previously stated, the dispersionαthat characterizes read counts of genes needs to be estimated in the context of a negative binomial distribution model (this is true both for DESeq2 and Cuffdiff 2).

The dispersion can be estimated by taking into account the relationship between mean and variance of gene counts (which is not a linear function in RNA-seq data). This information is incorporated in DESeq2 in a model that produces *a posteriori* (MAP) estimates of gene-wise dispersion. The process is articulated in three sequential steps: 1) count data for each gene is used to produce preliminary gene-wise dispersion estimates by maximum-likelihood estimation, 2) a dispersion trend is fit over all genes using the following parametrization:

$$\alpha_{tr}(\mu) = \frac{\alpha_1}{\mu} + \alpha_0$$

3) the likelihood from (1) is combined with the trended prior (2) to give final *a posteriori* dispersion estimates.

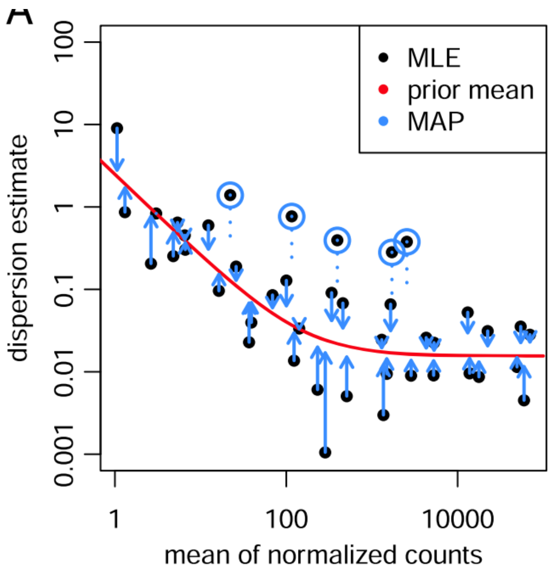


Fig.9: Dispersion estimation performed with DESeq2 [15]. Initial dispersion estimates are 'shifted' closer to the trended estimates by using a (computationally fast) approximation to a full empirical Bayes treatment (MLE= maximum-likelihood estimate, MAP=maximum *a posteriori*)

Testing for differential expression: The quantitative information provided by RNA-seq can be summarized by the ratio of expression values in different groups/conditions, a metric that was (in microarray data analysis the somewhat arbitrary 'twofold' change has been often used to denote 'differential' expression between different conditions).

Cufflinks: Briefly, the log-transformed ratio of expression constitutes a test statistic that follows a standard normal distribution when divided by the variance of the transformed ratio. A two-sided test for significance against a null hypothesis that the ratio is unity (no change) is performed.

DESeq2: In DESeq2, the βestimates (ratio of classes for a single gene or log fold changes) are calculated using a procedure that penalizes genes with low estimated mean values μ_{ij} or high dispersion estimates α_i, so that these are 'shrunken' toward zero. The same happens for datasets having limited degrees of freedom.

A Wald test is then performed that compares the β 'shrunken' estimate divided by its estimated standard error SE to a standard normal distribution.

Scope of the thesis

The work carried out in the context of this thesis represents the first comprehensive transcriptome analysis of human lymphocytes focusing on the expression of long intergenic non coding RNAs (lincRNAs). After deriving a list of 'signature' genes that are specifically expressed in the 13 human lymphocyte subsets we analyzed through RNA-seq, we focused our attention on a TH1-specific lincRNA (linc-MAF-4) that we demonstrated to be involved in the maintenance of TH1 cell identity via an epigenetic-repression of MAF gene.

The follow-up work described in Chapter 5 is aimed to the characterization of lncRNAs' involvement in regulatory processes for tumor-infiltrating lymphocytes (TIL).

References

1) Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research*, *22*(9), 1775-1789.
2) Panzeri, I., Rossetti, G., Abrignani, S., & Pagani, M. (2015). Long Intergenic Non-Coding RNAs: Novel Drivers of Human Lymphocyte Differentiation. *Frontiers in immunology*, *6*.
3) Doerr, A. (2012). Bioinformatics: Predicting PPIs. *Nature methods*, *9*(12), 1139-1139.
4) Ilott, N. E., & Ponting, C. P. (2013). Predicting long non-coding RNAs using RNA sequencing. *Methods*, *63*(1), 50-59.
5) Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357-359.
6) Lunter, G., & Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research*, *21*(6), 936-939.
7) Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21.
8) Marco-Sola, S., Sammeth, M., Guigó, R., & Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature methods*, *9*(12), 1185-1188.

9) Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., et al. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic acids research*, *38*(18), e178-e178.

10) Wu, Thomas D, and Colin K Watanabe. "GMAP: a genomic mapping and alignment program for mRNA and EST sequences." *Bioinformatics* 21.9 (2005): 1859-1875.

11) Trapnell, Cole, Lior Pachter, and Steven L Salzberg. "TopHat: discovering splice junctions with RNA-Seq." *Bioinformatics* 25.9 (2009): 1105-1111.

12) Marioni, John C et al. "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." *Genome research* 18.9 (2008): 1509-1517.

13) Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics* 26.1 (2010): 139-140.

14) Anders, Simon, and Wolfgang Huber. "Differential expression analysis for sequence count data." *Genome biol* 11.10 (2010): R106.

15) Love, Michael I, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biol* 15.12 (2014): 550.

16) Hardcastle, Thomas J, and Krystyna A Kelly. "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data." *BMC bioinformatics* 11.1 (2010): 422.

17) Leng, Ning et al. "EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments." *Bioinformatics* 29.8 (2013): 1035-1043.

18) Trapnell, Cole et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq." *Nature biotechnology* 31.1 (2013): 46-53.

19) Seyednasrollah, Fatemeh, Asta Laiho, and Laura L Elo. "Comparison of software packages for detecting differential expression in RNA-seq studies." *Briefings in bioinformatics* 16.1 (2015): 59-70.

20) Bánfai, Balázs et al. "Long noncoding RNAs are rarely translated in two human cell lines." *Genome research* 22.9 (2012): 1646-1657.

21) Cabili, Moran N et al. "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses." *Genes & development* 25.18 (2011): 1915-1927.

22) Ranzani, Valeria et al. "The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of T cell differentiation by linc-MAF-4." *Nature immunology* 16.3 (2015): 318-325.

23) Martin, Jeffrey A, and Zhong Wang. "Next-generation transcriptome assembly." *Nature Reviews Genetics* 12.10 (2011): 671-682.

24) Guttman, Mitchell et al. "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs." *Nature biotechnology* 28.5 (2010): 503-510.

25) Haas, Brian J et al. "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis." *Nature protocols* 8.8 (2013): 1494-1512.

26) Schulz, Marcel H et al. "Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels." *Bioinformatics* 28.8 (2012): 1086-1092.

27) Lin, Michael F, Irwin Jungreis, and Manolis Kellis. "PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions." *Bioinformatics* 27.13 (2011): i275-i282.

28) Wang, Liguo et al. "CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model." *Nucleic acids research* 41.6 (2013): e74-e74.

29) Sun, Kun et al. "iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data." *BMC genomics* 14.Suppl 2 (2013): S7.

30) Sun, Liang et al. "Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts." *Nucleic acids research* (2013): gkt646.

31) Rinn, J. L., & Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annual review of biochemistry*, *81*.

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

33) Wang, K. C., & Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Molecular cell*, *43*(6), 904-914.

34) Pontier, D. B., & Gribnau, J. (2011). Xist regulation and function explored. *Human genetics*, *130*(2), 223-236.

35) Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, *129*(7), 1311-1323.

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

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

38) Li, Z., Chao, T., Chang, K., Lin, N., Patil, V. S., Shimizu, C., et al. (2014). The long noncoding RNA THRIL regulates TNFα expression through its interaction with hnRNPL. *Proceedings of the National Academy of Sciences*, *111*(3), 1002-1007.

39) Wang, P., Xue, Y., Han, Y., Lin, L., Wu, C., Xu, S., et al. (2014). The STAT3-binding long noncoding RNA lnc-DC controls human dendritic cell differentiation. *Science*, *344*(6181), 310-313.

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

41) Satpathy, Ansuman T, and Howard Y Chang. "Long Noncoding RNA in Hematopoiesis and Immunity." *Immunity* 42.5 (2015): 792-804.

42) Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., et al. (2005). Identification of hundreds of conserved and nonconserved human microRNAs. *Nature genetics*, *37*(7), 766-770

43) Diederichs, S. (2014). The four dimensions of noncoding RNA conservation. *Trends in Genetics*, *30*(4), 121-123.

44) Smith, Martin A et al. "Widespread purifying selection on RNA structure in mammals." *Nucleic acids research* 41.17 (2013): 8220-8236.

45) Volders, P., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., et al. (2013). LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic acids research*, *41*(D1), D246-D251.

46) Qu, Z., & Adelson, D. L. (2012). Evolutionary conservation and functional roles of ncRNA. *Frontiers in genetics*, *3*.

47) Bateman, Alex et al. "The Pfam protein families database." *Nucleic acids research* 32.suppl 1 (2004): D138-D141.

48) Feuerer, Markus et al. "Lean, but not obese, fat is enriched for a unique population of regulatory T cells that affect metabolic parameters." *Nature medicine* 15.8 (2009): 930-939.

49) Rosenblum, Michael D et al. "Response to self antigen imprints regulatory memory in tissues." *Nature* 480.7378 (2011): 538-542.

Chapter 2

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

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

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

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

[3] These authors contributed equally to this work

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

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

Introduction

Lymphocytes enable humans 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 coordinated mostly by distinct CD4$^+$ T cell subsets through signals delivered both by cytokines and by cell-to-cell contacts[1]. The developmental and differentiation programs of CD4$^+$ T lymphocyte subsets with distinct effector functions have been extensively studied in terms of signaling pathways and transcriptional networks, and a certain degree of functional plasticity among different subsets has been established[2]. Indeed, flexibility of the CD4$^+$ T cell subset in the expression of genes encoding 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 of how the phenotype and functions of lymphocytes can be modulated and whether such findings offer new therapeutic opportunities.

In addition to 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 the maintenance of cellular states[4]. In this context, noncoding RNAs are emerging as a new regulatory layer that affects both the development of the immune system and its function[5, 6]. Among the several classes of noncoding RNAs with a specific role in lymphocyte biology, microRNAs are the best characterized[7, 8, 9, 10, 11]. Although thousands of long intergenic noncoding RNAs (lincRNAs) have been identified in the mammalian genome by bioinformatics analyses of transcriptomic data[12, 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 transcript stability[20] directly by base-pairing with specific targets or indirectly as competing endogenous RNAs[21, 22, 23]. A few examples of functional lincRNAs in the mouse immune system have been described. A broad analysis investigating naive and memory CD8$^+$ cells purified from mouse spleen with a custom array of lincRNAs has reported the identification of 96 lymphoid-specific lincRNAs and has suggested a role for lincRNAs in the differentiation and activation of lymphocytes[24]. The lincRNA NeST has been found to be downregulated during lymphocyte activation in a manner reciprocal to the expression of interferon-γ (IFN-γ) and to control susceptibility to infection with Theiler's virus and salmonella in mice through epigenetic regulation of the *Ifng* locus[25, 26]. Subsequently, mouse lincRNA-Cox2 was reported to be induced downstream of signaling via Toll-like receptors and to mediate the activation and repression of distinct sets of genes that are targets of the immune system that encode molecules involved in inflammatory responses[27]. Another study of mouse thymocytes and mature peripheral T cells has allowed the identification of lincRNAs with specific expression patterns during T cell differentiation and of LincR-Ccr2-5′AS, a lincRNA specific to CD4$^+$ T helper type 2 cells (TH2 cells) that is involved in regulating the migration of CD4$^+$ TH2 lymphocytes[28]. Although such studies highlight the relevance of lincRNAs in regulating immune responses, a thorough analysis of their expression profile and function in the human immune system is still lacking.

The present study was based on the analysis of 13 highly purified primary human lymphocytes subsets by high-throughput sequencing technologies for cDNA (RNA-seq analysis). We performed *de novo* transcriptome reconstruction (the creation of a transcriptome without the aid of a reference genome)[29] and discovered over 500 previously unknown long intergenic noncoding RNAs (lincRNAs). We identified several lymphocyte subset–specific lincRNA signatures and found that expression of linc-MAF-4, a chromatin-associated, CD4$^+$ TH1 cell–specific lincRNA, correlated inversely with expression of the transcription factor c-Maf and that its downregulation skewed the differentiation of CD4$^+$ T cells toward the TH2 phenotype. We provide the first comprehensive inventory, to our knowledge, of human lymphocyte lincRNAs and demonstrate that lincRNAs can be key to lymphocyte differentiation. This resource will probably help in providing a better definition of the role of lincRNAs in the differentiation, plasticity and effector functions of lymphocytes.

Discrimination of human lymphocyte subsets by lincRNAs

To assess lincRNA expression in human primary lymphocytes, we extracted RNA from 13 lymphocyte cell subsets (Table 1) purified from peripheral blood mononuclear cells from five healthy donors[11]. We then analyzed the polyadenylated RNA fraction by paired-end RNA sequencing and obtained about 1.7 × 10^9 mapped 'reads'. To enrich for transcripts derived from true active genes, we applied an expression threshold of 0.21 FPKM (fragments per kilobases of exons per million fragments mapped), defined through the integration of RNA-seq data and data on chromatin states from the ENCODE (Encyclopedia of DNA Elements) project[30]. We found a total of 31,902 expressed genes (including both protein-coding genes and noncoding genes) in the 13 subsets (Table 1 and Supplementary Fig. 1a), of which 4,201 were lincRNAs annotated in public resources[12, 31] (Fig. 1). To identify previously unknown lincRNAs expressed in primary human lymphocytes, we used three *de novo*transcriptome–reconstruction strategies based on the combination of two different sequence mappers, TopHat and Star[32, 33], with two different tools for *de novo* transcript assembly, Cufflinks and Trinity[34, 35]. We identified lincRNAs among the newly described transcripts by exploiting the following process. We selected transcripts that were longer than 200 nucleotides and multiexonic that did not overlap with protein-coding genes (and thus excluded unreliable single-exon fragments assembled by RNA-seq). We excluded transcripts with a conserved protein-coding region and those with open reading frames encoding protein domains catalogued in the Pfam database of protein families[36]. We used PhyloCSF, a comparative genomics method that assesses multispecies nucleotide-sequence alignment on the basis of a formal statistical comparison of phylogenetic codon models[37], which efficiently identifies noncoding RNAs, as demonstrated by ribosome-profiling experiments[38]. Finally, we defined a stringent new lincRNA set that included those genes for which at least one lincRNA isoform was reconstructed by two assemblers of three. Through this conservatively multilayered analysis we identified 563 previously unknown lincRNA-encoding genes, which increased by 11.8% the number of lincRNAs known to be expressed in human lymphocytes. The various classes of RNAs were evenly distributed among various lymphocyte subsets (Supplementary Fig. 1b), and the ratio of already annotated and newly identified lincRNAs was similar across different chromosomes (Supplementary Fig. 1c) and across various lymphocyte subsets (Supplementary Fig. 1d). As observed in various cell types[12, 34], lincRNAs were also generally expressed at lower abundance than were protein-coding genes in human lymphocytes (Supplementary Fig. 1e). However, when we categorized transcripts on the basis of their cell-specific expression and non–cell-specific expression (Supplementary Fig. 1f), we found that cell-specific lincRNAs and cell-specific protein-coding genes displayed 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 are protein-coding genes[12, 39], 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 were defined by 15 clusters and protein-coding genes were defined by 24 clusters (Fig. 2a and Supplementary Fig. 2a). Notably, the frequency of genes assigned to the clusters specific for the various lymphocyte subsets was higher for lincRNAs (71%) than for protein-coding genes (34%) (Fig. 2b). This superiority stood out even when we compared lincRNAs with genes encoding membrane receptors (40%) (Fig. 2c), which are generally considered the most accurate markers of various lymphocyte subsets. We obtained similar results with the heuristic expression threshold of FPKM > 1 (Supplementary Fig. 2b). Thus, by RNA-seq analyses of highly purified subsets of primary T lymphocytes and B lymphocytes, we were able to provide a comprehensive landscape of lincRNA expression in human lymphocytes. By exploiting *de novo* transcriptome reconstruction, we discovered 563 previously unknown lincRNAs and found that lincRNAs were effective in marking lymphocyte identity.

Identification of lincRNA signatures in lymphocytes

Next we investigated our data set for the presence of lincRNA signatures in the various lymphocyte subsets. We therefore looked for lincRNAs with a difference in expression of more than 2.5-fold in a given cell subset relative to their expression in all the other subsets (*P*< 0.05 (nonparametric Kruskal-Wallis test)) that were expressed in at least three of five donors and found 172 lincRNAs that met these criteria (Fig. 3a and Supplementary Table 1). We integrated the human transcriptome database with our newly identified transcripts and thus created a new reference with which to assess more thoroughly their expression in other human tissues. Assessing lincRNA signatures in a panel of 16 human tissues (from the Human BodyMap 2.0 project), we found that not only were lymphocyte signature lincRNAs expressed very poorly in nonlymphoid tissues but also most signature lincRNAs were not detectable even in lymphoid tissues (Fig. 3a,b). These findings emphasized the importance of assessing the expression of lincRNAs (as well as of any highly cell-specific transcripts) in purified primary cells rather than in total tissues in which a given cell subset–specific transcript is diluted by the transcripts of all the other cell types of the tissue. We note that the newly identified lincRNAs defined as signatures were more abundant (Fig. 3c) and more cell specific (Supplementary Table 1) than the already annotated lincRNAs defined as signatures. We present here data obtained from the CD4$^+$ TH1 cell subset (Fig. 2b); we obtained similar results for all the other subsets (Supplementary Table 1).

Finally, to confirm and extend our signature data, we assessed expression of the signature lincRNAs of CD4$^+$ TH1 cells (Fig. 3b) by quantitative RT-PCR of a new set of independent samples of primary human CD4$^+$ naive cells, regulatory T cells and TH1 cells, as well as in naive CD4$^+$ T cells that were activated *in vitro* and induced to differentiate toward the TH1 or TH2 phenotype. We confirmed specific-subset expression for 90% of the CD4$^+$ TH1 cell signature lincRNAs (Fig. 3d). Moreover, 90% of the CD4$^+$ TH1 cell signature lincRNAs that were expressed in resting CD4$^+$ TH1 cells purified *ex vivo* also had high expression in naive CD4$^+$ T cells differentiated under TH1-polarizing conditions *in vitro*, whereas they had low expression in naive CD4$^+$ T cells differentiated toward the TH2 phenotype *in vitro* (Fig. 3e). As a corollary to those findings, we observed by RNA-seq that the signature lincRNAs of CD4$^+$ naive cells were mostly downregulated during differentiation toward the TH0 phenotype*in vitro*, whereas the signature lincRNAs of cells of the TH1, TH2 and TH17 subsets of helper T cells were mostly upregulated (Supplementary Fig. 3a). Together our data demonstrated that lincRNAs provided signatures of human lymphocyte subsets and suggested that human CD4$^+$ T lymphocytes acquired most of their memory-specific lincRNA signatures during their activation-driven differentiation from naive cells to memory cells.

Downregulation of linc-MAF-4 skews CD4$^+$ T cells toward TH2 cells

As lincRNAs have been reported to influence the expression of neighboring genes[25, 26, 28,40], we sought to determine whether protein-coding genes proximal to the signature lincRNAs of lymphocytes were involved in key cell functions. For this we used the FatiGO tool from the Babelomics suite for functional enrichment analysis[41] and found that protein-coding genes adjacent to signature lincRNAs showed enrichment for gene ontology terms correlated with the activation of lymphocyte T cells (Fig. 4), which indicated a possible role for signature lincRNAs in lymphocyte function. To obtain proof of concept of this hypothesis, we chose to characterize in depth linc-MAF-4 (lnc-MAF-2 in the LNCipedia database[42]), a signature lincRNA of TH1 cells located 139.5 kilobases upstream of *MAF*. This gene encodes transcription factor c-Maf, which is involved in TH2 differentiation[43] but is also required for the efficient development of TH17 cells[44] and controls transcription of the gene encoding interleukin 4 in CD4$^+$ follicular helper T cells[45]. Our sequencing data showed that high expression of linc-MAF-4 correlated with a low abundance of *MAF* transcripts in CD4$^+$ TH1 cells; conversely, TH2 cells had low expression of linc-MAF-4 and abundant *MAF* transcripts (data not shown). The anti-correlation of expression between lincRNAs and their neighboring genes is not a common feature of all lincRNAs[12, 16] and is probably restricted to a limited number of *cis*-acting lincRNAs. We also confirmed this observation in our data set (data not shown). Moreover, we observed no correlation between the expression of linc-MAF-4 and its proximal upstream protein-coding genes *CDYL2* and *DYNLRB2* (Supplementary Fig. 4a).

We observed a similar inverse relation between linc-MAF-4 and MAF when we differentiated naive CD4$^+$ T cells *in vitro* toward the TH1 or TH2 phenotype. In T lymphocytes differentiating toward the TH1 phenotype, *MAF* transcripts increased up to day 3 and then decreased thereafter (Fig. 5a). Conversely, *linc-MAF-4* was poorly expressed for the first 3 d but then increased progressively (Fig. 5a). In CD4$^+$ T lymphocytes differentiating toward the TH2 phenotype, the abundance of both *MAF* transcripts and c-Maf protein increased constantly up to day 8, while linc-MAF-4 remained constantly low (Fig. 5a and Supplementary Fig. 4c), similar to what we observed for CD4$^+$ T lymphocytes differentiating toward the TH17 phenotype (Supplementary Fig. 4d).

We further characterized the transcriptional regulation of *MAF* by assessing the abundance of histone H3 trimethylated at Lys4 (H3K4me3) and occupancy by RNA polymerase II at the*MAF* promoter region in TH1 and TH2 cells. Consistent with the higher active transcription of*MAF* in CD4$^+$ TH2 cells, we found enrichment for H3K4me3 in TH2 cells relative to its abundance in TH1 cells and that binding of RNA polymerase II at *MAF* promoter was higher in TH2 than in TH1 cells (Fig. 5b). Notably, knockdown of linc-MAF-4 in activated CD4$^+$ naive T cells led to increased *MAF* expression (Fig. 5c and Supplementary Fig. 4e). All the results presented above indicated that modulation of *MAF* transcription in T cells depended on tuning of its promoter setting, and suggested direct involvement of linc-MAF-4 in the regulation of *MAF* transcription.

We then assessed the overall effect of the knockdown of linc-MAF-4 on the differentiation of CD4$^+$ T cells by transcriptome profiling and gene set–enrichment analysis. We defined as reference gene sets the groups of genes upregulated in CD4$^+$ naive T cells differentiated *in vitro* toward the TH1 or TH2 phenotype (Supplementary Table 2). We found that the CD4$^+$TH2 cell gene set showed enrichment for genes overexpressed in cells in which linc-MAF-4 was knocked down, whereas the CD4$^+$ TH1 cell gene set showed depletion of those same genes (Fig. 5d). Concordant with those findings, the expression of *GATA3* and *IL4,* two genes characteristic of TH2 cells, was increased after knockdown of linc-MAF-4 (Fig. 5e andSupplementary Fig. 4f). Together these results demonstrated that downregulation of linc-MAF-4 contributed to skewing of the differentiation of CD4$^+$ T cells toward the TH2 phenotype.

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

Since the gene encoding linc-MAF-4 maps in relative proximity to *MAF* (within 139.5 kilobases), we sought to determine whether linc-MAF-4 was able to downregulate *MAF*transcription, and we investigated whether their genomic regions could physically interact. We exploited chromosome-conformation capture analysis to determine the relative crosslinking frequencies among regions of interest. We assessed the conformation of the genomic regions of the gene encoding linc-MAF-4 (called '*linc-MAF-4*' here) and *MAF* in differentiated CD4$^+$ TH1 cells. We used common reverse-primer mapping of the *MAF*promoter region in combination with a set of primers spanning the locus and analyzed interactions by PCR. We detected specific interactions between the *MAF* promoter and the 5′ and 3′ end regions of *linc-MAF-4* (Fig. 6a and Supplementary Fig. 5a,b), which indicated the existence of an *in cis* chromatin-looping conformation that brought *linc-MAF-4* in close proximity to the *MAF* promoter. Notably, subcellular fractionation of CD4$^+$ TH1 lymphocytes differentiated *in vitro* revealed considerable enrichment for linc-MAF-4 in the chromatin fraction (Fig. 6b). Because other chromatin-associated lincRNAs regulate neighboring genes by recruiting specific chromatin remodelers, we assessed by RNA-immunoprecipitation assay the interaction of linc-MAF-4 with various chromatin modifiers, including activators and repressors (data not shown), and found specific enrichment for linc-MAF-4 in the immunoprecipitates of two chromatin modifiers, EZH2 and LSD1 (Fig. 6c and Supplementary Fig. 5c). In agreement with those findings, we found that knockdown of linc-MAF-4 in activated CD4$^+$ naive T cells reduced the abundance of both EZH2 and LSD1 and correlated with lower enzymatic activity of EZH2 at the *MAF* promoter, as demonstrated by a lower abundance of H3K27me3 at this locus (Fig. 6d). Notably, the content of H3K27me3 was not diminished at either the *MYOD1* promoter region (a known target of EZH2) or at a region within the chromatin loop between *linc-MAF-4* and *MAF* marked by H3K27me3 (Supplementary Fig. 5d). Together these results demonstrated a long-distance interaction between the genomic regions of *linc-MAF-4* and *MAF*, through which linc-MAF-4 might act as a scaffold to recruit both EZH2 and LSD1 and modulate the enzymatic activity of EZH2 on the *MAF* promoter and thus regulate its transcription (Fig. 6e).

Mammalian genomes encode more long noncoding RNAs than initially thought[16, 46], and the identification of lincRNAs with a role in cellular processes is growing steadily. As there are relatively few examples of functional long noncoding RNAs in the immune system[24, 25, 26,27, 28], with the present study we have presented a comprehensive landscape of the expression of lincRNAs in 13 subsets of human primary lymphocytes. Moreover, we have identified a lincRNA (linc-MAF-4) that seemed to have a key role in the differentiation of CD4$^+$ helper T cells.

LincRNAs have been reported to have high tissue specificity[12], and our study of lincRNA expression in highly pure primary human lymphocyte has provided added value because it allowed the identification of lincRNAs whose expression was restricted to a given lymphocyte cell subset. Notably, we found that lincRNAs defined cellular identity better than protein-coding genes did, including those that encode surface receptors that are generally considered the most precise markers of lymphocyte subsets. Due to their specificity of expression, human lymphocyte lincRNAs that are not yet annotated in public resources would have not been identified without *de novo* transcriptome reconstruction. Indeed, by exploiting three different *de novo* strategies, we identified 563 previously unknown lincRNAs and increased by 11.8% the number of lincRNAs known to be expressed in human lymphocytes. As our conservative analysis was limited to 13 cellular subsets, it remains unclear how many novel lincRNAs could be identified by transcriptome analysis of all of the several hundreds of human cell types.

We compared our data with published analyses of lincRNA expression in the mouse immune system[28], exploiting the LNCipedia database[42]. We found that 51% of the human lincRNA signature was conserved in mice, which is similar to the overall conservation between human lincRNAs and mouse lincRNAs (60%). However, further studies will be needed to assess whether their function is also conserved.

Given our findings, signature lincRNAs might be exploited to discriminate and differentiate at the molecular level those cell subsets that cannot be distinguished easily on the basis of cell surface markers because of their cellular heterogeneity, such as CD4$^+$ regulatory T cells. However, as lincRNA expression in a tissue is averaged across all the cell types that compose that tissue, transcriptome analysis of unfractionated tissue-derived cells may underestimate the expression of cell-specific lincRNAs. In fact, the great majority of our lymphocyte lincRNA signatures could not be detected among RNA extracted from total lymphoid tissues (peripheral blood and lymph nodes), although these same tissues contained cells from all of the lymphocytes subsets we assessed.

The role of lincRNAs in differentiation has been described for various cell types[17, 20, 23, 47,48]. In the mouse immune system, it has been found that lincRNA expression changes during the differentiation of naive CD8$^+$ T cells into memory CD8$^+$ T cells[24] and during the differentiation of naive CD4$^+$ T cells into distinct lineages of helper T cells[28]. We have shown for human primary lymphocytes that activation-induced differentiation of CD4$^+$ naive T cells was associated with increased expression of lincRNAs belonging to the CD4$^+$ TH1 cell signature, which suggests that upregulation of TH1 cell lincRNAs is part of the cell-differentiation transcriptional program. Indeed, linc-MAF-4, one of the TH1 cell signature lincRNAs, had low expression in TH2 cells, and its experimental downregulation skewed differentiating helper T cells toward a TH2 transcription profile. We found that linc-MAF-4 regulated transcription by exploiting a chromosome loop that brought its genomic region close to the promoter of *MAF*. We propose that the chromosome organization of this region allows a linc-MAF-4 transcript to recruit both EZH2 and LSD1 and to modulate the enzymatic activity of EZH2 that negatively regulates *MAF* transcription via a mechanism of action similar to that shown for the lincRNAs HOTAIR[49] and MEG3 (ref. 50). We therefore have provided mechanistic proof of the concept that lincRNAs can be important regulators of CD4$^+$ T cell differentiation. Given the number of specific lincRNAs expressed in various lymphocyte subsets, it can be postulated that many other lincRNAs might contribute to cell differentiation and to the definition of identity in human lymphocytes. These findings and the high cell specificity of lincRNAs suggest that lincRNAs might be highly specific molecular targets for the development of new therapies for diseases (such as autoimmunity, allergy and cancer) in which altered CD4$^+$ T cell functions have a pathogenic role.

Discussion

Purification of primary immunological cell subsets.

Blood buffy coat cells of healthy donors were obtained from Fondazione Istituto di Ricovero e Cura a Carattere Scientifico Ca'Granda Ospedale Maggiore Policlinico in Milan, and peripheral blood mononuclear cells were isolated by ficoll-hypaque density-gradient centrifugation. The ethical committee of Fondazione Istituto di Ricovero e Cura a Carattere Scientifico Ca'Granda Ospedale Maggiore Policlinico approved the use of peripheral blood mononuclear cells from healthy donors for research purposes, and informed consent was obtained from subjects. Human blood primary lymphocyte subsets were purified to a purity of >95% by cell sorting through the use of various combinations of surface markers (Table 1). For *in vitro* differentiation experiments, resting naive CD4$^+$ T cells were purified to a purity of >95% by negative selection with magnetic beads with an isolation kit for human CD4$^+$ Naive T cells (Miltenyi) and were stimulated with Dynabeads Human T-Activator CD3/CD28 (Life Technologies). Interleukin 2 (IL-2) was added at 20

IU/ml (202-IL; R&D Systems). TH1 polarization was initiated with 10 ng/ml IL-12 (219-IL; R&D Systems) and TH2-neutralizing antibody anti-IL-4 (2 µg/ml; MAB3007; R&D Systems). TH2 polarization was induced by activation with phytohemagglutinin (4 µg/ml; L2769; Sigma) in the presence of IL-4 (10 ng/ml; 204-IL; R&D Systems), and neutralizing anti-IFN-γ (2 µg/ml; MAB 285; R&D Systems) and anti-IL-12 (2 µg/ml; MAB219; R&D Systems). For intracellular staining of GATA-3 and c-Maf, cells were harvested and then were fixed for 30 min at 4 °C in Fixation/Permeabilization Buffer (eBioscience). Cells were stained for 30 min at 4 °C with anti-GATA-3 (TWAJ; eBioscience) and anti-c-Maf (sym0F1; eBioscience) in washing buffer. Cells were then washed two times, resuspended in autoMACS buffer (Miltenyi) and analyzed by flow cytometry.

RNA isolation and RNA sequencing.

Total RNA was isolated with an 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 libraries generated were loaded on to the cBot automated clonal amplification system (Illumina) for clustering on a HiSeq Flow Cell v3. The libriaries clustered on a HiSeq Flow Cell v3 were then sequenced with a HiScanSQ optical imaging system (Illumina). A paired-end run (with a read length of 101 bases) was performed with an SBS Kit v3 DNA sequencing kit (Illumina). Real-time analysis and base calling was performed with HiSeq Control Software Version 1.5 (Illumina).

RNA-seq.

RNA-seq data representative of 13 lymphocyte populations were collected for transcriptome reconstruction. Five biological replicates were analyzed for all populations except for CD8$^+$TCM cells and CD5$^+$ B cells (four samples). The whole data set was aligned to human genome assembly GRCh37 (Genome Reference Consortium Human Build 37) with TopHat software (version 1.4.1)[33] for a total of over 1.7 × 10^9 mapped paired-end reads (30 million reads per sample on average). These data were also mapped with the aligner STAR (version 2.2.0)[32]. RNA-seq data sets of 16 human tissues belonging to the Illumina Human BodyMap 2.0 project (ArrayExpress accession code E-MTAB-513) were mapped according to the same criteria.

Reference annotation.

An initial custom reference annotation of unique, non-redundant transcripts was built by integration of the Ensembl database (version 67 from May 2012) with the lincRNAs identified by another group[13] through the use of the Cuffcompare tool (version 2.1.1) of the Cufflinks suite[34]. The annotated human lincRNAs were extracted from Ensembl through the use of the BioMart software suite (version 67) and were categorized 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 collection of receptor-encoding genes defined in BioMart under GO term GO:000487 (2,043 genes encoding molecules with receptor activity function) and the class of genes encoding molecules 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 were nonredundant lincRNA-encoding genes.

De novo **genome-based transcripts reconstruction.**

A comprehensive catalog of lincRNAs specifically expressed in human lymphocyte subsets was generated with a *de novo* genome-based transcripts reconstruction procedure by three different approaches. Two aligners were used: TopHat (version 1.4.1) and STAR (version 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 (version 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 with which transcript abundance was estimated. By this method, about 3 × 10^4 to 5 × 10^4 previously unknown transcripts were identified in each lymphocyte population. The third approach used genome-guided Trinity software (additional information available athttp://pasa.sourceforge.net/#A_ComprehensiveTranscriptome), which generates novel transcripts by local assembly on previously mapped reads from specific location. STAR was used instead of the Trinity default aligner[29]. Each candidate transcript was then processed via the PASA 'pipeline' (Program to Assemble Spliced Alignments; a genome annotation tool), which reconstructs the complete transcript and gene structures, resolving incongruences derived from transcript misalignments and alternatively splices events, refining the reference annotation when there was enough evidence and proposing new transcripts and genes in case no previous annotation was able to explain the new data (Supplementary Note).

Identification of previously unknown lincRNA-encoding genes.
Annotated transcripts and previously unknown isoforms of known genes were discarded, and only previously unknown genes and their isoforms located in intergenic positions were retained. 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[36] was run for each transcript to identify occurrences of any protein family domain documented in the Pfam database (release 26; both PfamA and PfamB were used). All six possible frames were considered for the analysis, and the matching transcripts were excluded from the final catalog.

The coding potential for all the remaining transcripts was then evaluated by the PhyloCSF comparative genomics method (phylogenetic codon substitution frequency)[37], which was run on a multiple sequence alignment of 29 mammalian genomes (in multi-alignment file (MAF) format) (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/) to obtain the open reading frames that encoded proteins of over 29 amino acids in length across all three reading frames and had the best scores. For efficient accessing of the MAFs, the biogem plugin of the bio-maf Ruby (MAF parser for the BioRuby open-source bioinformatics library for Ruby programming code; https://github.com/csw/bioruby-maf)[51] was used. This library provides indexed and sequential access to MAF data, and also performs fast manipulations on it and writes modified MAFs. Transcripts with at least one open reading frame with a PhyloCSF score of over 100 were excluded from the final catalog. The threshold of 100 for the PhyloCSF score was determined as described[13] to optimize specificity and sensitivity for the classification of coding and noncoding transcripts annotated in the RefSeq reference sequence database of the National Center for Biotechnology Information (RefSeq coding and RefSeq lincRNAs). A PhyloCSF score of 100 corresponds to a false-negative rate of 6% for coding genes (i.e., 6% of coding genes are classified as noncoding) and a false-positive rate of ~10% (i.e., 9.5% of noncoding transcripts are classified as coding).

De novo transcriptome data integration.

Duplicates among the transcripts identified with the same *de novo* method were resolved through the use of Cuffcompare (version 2.1.1). In the same way, the resulting three data sets were further merged to generate a nonredundant atlas of lincRNAs in human lymphocytes and only those genes identified by at least two of the three software programs used 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 relative 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 nonredundant catalog of newly identified lincRNAs includes 4,666 previously unknown transcripts referencing 3,005 previously unknown genes.

Definition of lincRNA signatures.

Analysis of differences in expression among the 13 cell subsets profiled was performed with the Cuffdiff program of Cufflinks (version 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 'calls' for differences in expression for less-abundant genes and transcripts. Only genes expressed at an FPKM value over 0.21 (ref. 29) were considered in the downstream analysis to filter out genes that are merely byproducts of 'leaky' gene expression, sequencing errors, and/or off-target read mapping. After a pseudo-count of 1 was added to the raw FPKM value for each gene, with the application of log2 transformation and *z*-score normalization, K-means clustering with Euclidean metric was performed on lincRNA expression values with the MultiExperiment Viewer tool (version 4.6) (Supplementary Note). The same procedure was then applied to the expression values of genes encoding proteins, products involved in metabolic processes and receptors. The Silhouette function[52] was used to select an appropriate K value (number of clusters). K values ranging from 13 to 60 were tested, and the value associated with the highest Silhouette score for each class of genes was selected. The number of clusters that maximized the Silhouette score was 15 for lincRNA (Supplementary Fig. 2a), 24 for protein-coding genes, 23 genes encoding receptors and 36 for genes encoding products involved in metabolic processes. The centroid expression profile of each cluster was then evaluated to associate each cluster to a single cellular population (Fig. 2).

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 (Supplementary Note). The estimation procedure for the JS score was adapted by the building of a reference model composed of 13 cell subsets. For the lincRNAs selected, the intrapopulation consistency among different samples was subsequently evaluated to minimize the biological variability: only genes expressed in at least three of five of the samples profiled (or three of four replicates for CD8$^+$ TCM cells and CD5$^+$B cells) whose maximal expression value was >2.5-fold that in all other lymphocyte subsets were considered. Finally, a nonparametric Kruskal-Wallis test was applied to select only lincRNA genes with a significant difference in expression 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.

GO enrichment analysis.

A GO enrichment analysis was performed for biological process terms associated with protein-coding genes that were proximal to lincRNA signatures at the genomic level. For each lincRNA signature, the proximal protein-coding gene was selected regardless of the sense of transcription. The FatiGO tool of the Babelomics suite (version 4.3.0) was used to identify the GO terms that showed enrichment, among the 158 protein-coding genes (input list). All protein-coding genes that were 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 GO semantic similarity analysis on the 51 GO terms with adjusted*P* value lower than 0.1, which resulted from previous analysis with the G-SESAME (gene semantic similarity analysis and measurement) tool. This analysis provides as a result a symmetric matrix in which each value represents a score for similarity 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'.

Transfection of siRNA into naive CD4$^+$ T cells.

300 nM fluorescein isothiocyanate (FITC)-labeled-siRNA targeting linc-MAF-4 or FITC-labeled-AllStars negative control (Qiagen) was transfected into activated CD4$^+$ naive T cells through the use of Lipofectamine 2000 according to the manufacturer's protocol (Life Technologies). FITC$^+$ cells were sorted and lysed 72 h after transfection. siRNAs sequences are provided in Supplementary Table 3.

Gene-expression analysis.

Gene expression in transfected activated CD4$^+$ naive cells was analyzed by Illumina Direct Hybridization Assays according to the standard protocol (Illumina). Total RNA was isolated, underwent quality control and was quantified as described above; for each sample, 500 ng total RNA was reverse transcribed according to the Illumina TotalPrep RNA Amplification kit (AMIL1791; LifeTechnologies) and cRNA was generated by 14 h of *in vitro* transcription. Samples were hybridized 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 underwent quantile normalization, with no subtraction of background values, and average signals were calculated on the gene level data for genes whose detection *P* value was lower than 0.001 in at least one of the cohorts considered.

Gene set–enrichment analysis (GSEA).

GSEA is a statistical methodology used to evaluate whether a given gene set shows significant enrichment for a list of gene markers (ranked by their correlation with a phenotype of interest). To evaluate this degree of 'enrichment', the software calculates an enrichment score (ES) by moving down the ranked list; i.e., it increases the value of the sum if the marker is included in the gene set and decreases 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 by comparison of gene-expression data obtained from activated CD4$^+$ naive T cells transfected with siRNA specific for linc-MAF-4 or control siRNA. The experimentally generated data set from cells differentiated *in vitro* (in TH1- or TH2-polarizing conditions) from CD4$^+$ naive T cells of the same donors in which linc-MAF-4 was downregulated were used to construct reference gene sets for TH1 and TH2 cells. RNA for analysis of gene expression in differentiating TH1 and TH2 cells was collected 72 h 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 the production of IFN-γ and IL-13 by TH1 and TH2 cells. The TH1 and TH2 data sets were ranked as log2 ratios of the expression values for each gene in the two conditions (TH1/TH2), and the genes with the greatest upregulation or downregulation (with log2 ratios ranging from |3| to |0.6|) were assigned to the TH1 or TH2 reference sets, respectively.

Genes from the TH1 gene list that were downregulated in a comparison of TH1 cells versus cells transfected with control siRNA and genes from the TH2 gene list that were downregulated in a comparison of TH2 cells versus cells transfected with control siRNA were filtered out, which resulted in a TH1 cell–specific gene set (74 genes) and a TH2 cell–specific gene set (141 genes) (Supplementary Table 2). GSEA was then performed on the data set for the comparison of cells transfected linc-MAF-4-specific siRNA versus cells transfected with control siRNA. The metric used for the analysis is the log2 ratio of classes, with 1,000 gene set permutations for testing of significance.

Quantitative RT-PCR analysis.

For reverse transcription, equal amounts of DNA-free RNA (500 ng) were reverse-transcribed with SuperScript III in the conditions suggested by the manufacturer (LifeTechnologies). Diluted cDNA was then used as input for quantitative RT-PCR to assess the expression of*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) with Inventoried TaqMan Gene Expression assays (LifeTechnologies). For assessment of linc-MAF-4 and confirmation of $CD4^+$ TH1 cell signature lincRNAs, specific primers were designed, and 2.5 µg RNA from $CD4^+$ TH1 cells, regulatory T cells or naive cells was used for reverse transcription with SuperScript III (LifeTechnologies). Quantitative RT-PCR was performed on diluted cDNA with PowerSyberGreen (LifeTechnologies), and the specificity of each amplified product was monitored through the use of melting curves at the end of each amplification reaction. The primers used in quantitative PCR are listed inSupplementary Table 3.

Cell fractionation.

TH1 cells differentiated *in vitro* were resuspended for 10 min on ice in RLN1 buffer (50 mM Tris-HCl pH 8, 140 mM NaCl, 1.5 mM MgCl2, 0.5% NP-40) supplemented with SUPERaseIn (Ambion). After a centrifugation at 300*g* for 2 min, the supernatant was collected as the cytoplasmic fraction. The pellet was resuspended for 10 min on ice in RLN2 buffer (50 mM Tris-HCl, pH 8, 500 mM NaCl, 1.5 mM MgCl2 and 0.5% NP-40) supplemented with RNase inhibitors. Chromatin was pelletted at maximum speed for 3 min. The supernatant represented the nuclear fraction. All fractions were resuspended in TRIzol (Ambion) to a volume of 1 ml, and RNA was extracted following a standard protocol.

RNA immunoprecipitation.

TH1 cells differentiated *in vitro* underwent crosslinking by ultraviolet irradiation at 400 mJ/cm^2in ice-cold Dulbecco's-PBS and then were pelleted at 1,350*g* for 5 min. The pellet was resuspended in ice-cold lysis buffer (25 mM Tris-HCl pH 7.5, 150 mM NaCl and 0.5% NP-40) supplemented with 0.5 mM β-mercaptoethanol, Protease Inhibitor Cocktail Tablets cOmplete, EDTA-free (Roche) and SUPERaseIn (Ambion) and was incubated with rocking at 4 °C until lysis was complete. The debris were centrifuged at 13,000*g* for 10 min. The lysate was precleared for 30 min at 4 °C with Dynabeads Protein G (Novex) and then was incubated for 2 h with 7 µg anti-EZH2 (39875; Active Motif) or anti-LSD1 (ab17721; Abcam), or with anti-HA (sc7392; Santa Cruz) as mock control. The lysate was coupled for 1 h at 4 °C to Dynabeads Protein G (Novex). Immunoprecipitates were washed for five times with lysis buffer. RNA was then extracted according to the protocol of the mirVana miRNA Isolation Kit (Ambion). The abundance of RNA transcripts encoding linc-MAF-4 or the negative controls β-actin, RNU2.1 and a region upstream the TSS of linc-MAF-4 (linc-MAF-4 control) was assessed by quantitative RT-PCR.

Chromatin immunoprecipitation.

TH1 and TH2 cells differentiated *in vitro* were crosslinked for 12 min in their medium with 1:10 dilution of fresh formaldehyde solution (50 mM HEPES-KOH, pH 7.5, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA and 11% formaldehyde). Subsequently, they were treated for 5 min with 1:10 dilution of 1.25 M glycine and were centrifuged at 1,350*g* for 5 min at 4 °C. Cells were lysed at 4 °C 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). Nuclei were pelleted at 1,350*g* for 5 min at 4 °C and were washed in LB2 (10 mM Tris-HCl, pH 8.0, 200 mM NaCl, 1 mM EDTA and 0.5 mM EGTA) supplemented protease inhibitors. Nuclei were again pelleted at 1,350*g* for 5 min at 4 °C and then were 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 and 0.5% N-lauroylscarcosine) supplemented with protease inhibitors. Cell debris were pelleted at 20,000*g* for 10 min at 4 °C, followed by chromatin immunoprecipitation overnight at 4 °C in LB3 supplemented with 1% Triton X-100 and protease inhibitors, with anti-H3K4me3 (07-473; Millipore), anti-H3K27me3 (07-449; Millipore), antibody to the RNA polymerase II STD repeat YSPTSPS (ab5408; Abcam), LSD1 (ab17721; Abcam), anti-EZH2 (39875; Active Motif) or no antibody (as negative control) . The next day, Dynabeads Protein G (Novex) were added, followed by incubation for 2 h at 4 °C with rocking. Then, the beads were washed twice with low-salt wash buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.1% SDS, 2 mM EDTA and 1% Triton X-100) and with a high-salt wash buffer (20 mM Tris-HCl, pH 8.0, 500 mM NaCl, 0.1% SDS, 2 mM EDTA and 1% Triton X-100). Samples obtained by immunoprecipitation with antibodies to histones (identified above) were also washed with a LiCl solution (10 mM Tris-HCl, pH 8.0, 250 mM LiCl, 1% NP-40 and 1 mM EDTA). All samples were finally washed with 50 mM NaCl in 1× Tris-EDTA buffer. Elution was performed overnight at 65 °C in 50 mM Tris-HCl pH 8.0, 10 mM EDTA and 1% SDS. Samples were treated for 2 h at

37 °C with 0.02 µg/µl RNase A (Sigma) and for 2 h at 55 °C with 0.04 µg/µl proteinase K (Sigma). DNA was purified with phenol-chloroform extraction.

Chromosome-conformation capture.

For chromosome-conformation capture analysis[53] cells were crosslinked and digested as describe above for chromatin immunoprecipitation. Nuclei resuspended in 500 µl of 1.2× NEB3 buffer (New England BioLabs) with 0.3% SDS were incubated at 37 °C for 1 h and then were incubated for another 1 h. with 2% Triton X-100. Samples underwent digestion overnight at 37 °C (with shaking) with 800 U of BglII (New England BioLabs). Digestion was checked by the separation of digested samples and undigested control samples by electrophoresis through a 0.6% agarose gel. Then, the samples were incubated for 25 min at 65 °C with 1.6% SDS and were incubated for 1 h at 37 °C with 1.15× ligation buffer (New England BioLabs) and 1% Triton X-100. Ligation with 1,000 U T4 DNA ligase (New England BioLabs) was performed for 8 h at 16 °C and at 22 °C for 30 min. DNA was purified by phenol-chloroform extraction after digestion with RNase A (Sigma) and proteinase K (Sigma). As controls, bacterial artificial chromosomes corresponding to the region of interested were digested overnight at 37 °C with 100 U BglII in NEB3 buffer in a volume of 50 µl. Then, fragments underwent ligation overnight at 22 °C with 400 U T4 DNA ligase in a volume of 40 µl. PCR products amplified with GoTaq Flexi (Promega) for bacterial artificial chromosomes and samples were separated by electrophoresis through 2.5% agarose gels and quantified with ImageJ software. Primers are listed in Supplementary Table 3.

Statistical analysis.

Unless indicated otherwise in the figure legend(s), a one-tailed, paired *t*-test was performed on experimental data with Prism (GraphPad Software). For multiple comparisons of human lymphocytes subsets, a nonparametric Kruskal-Wallis test was used. Analysis of variance and Dunnet post-hoc test was applied for statistical analysis of RNA-immunoprecipitation experiments in Figure 6c.

Accession codes.

ArrayExpress: E-MTAB-2319.

These authors contributed equally to this work.

Valeria Ranzani, Grazisa Rossetti, Ilaria Panzeri, Alberto Arrigoni & Raoul J P Bonnal

Contributions

V.R., A.A. and R.J.P.B. set up 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, analyzed the data and contributed to the preparation of the manuscript; S.C., P.G., E.P., E.S. and B.B. performed experiments and analyzed 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; and all authors discussed and interpreted the results.

Competing financial interests

The authors declare no competing financial interests.

Corresponding authors

- Massimiliano Pagani or
- Sergio Abrignani

Figure and Table Legends

Figure 1: For the identification of lincRNAs, RNA-seq data generated from 63 lymphocyte samples were processed by the quantification of lincRNAs already annotated in public resources (left) and by *de novo* genome-based transcripts reconstruction for the quantification of previously unknown lincRNAs expressed in human lymphocytes (right) through the use of reference annotation–based assembly by Cufflinks software with the aligners TopHat and STAR and by an approach that integrates Trinity and PASA software (bottom right). Only transcripts reconstructed by at least two assemblers were considered. Newly

identified transcripts were filtered with a computational analysis pipeline to select for lincRNAs.

Figure 2: (a) Expression profiles of lincRNA and protein-coding genes across 13 human lymphocyte subsets (key at right; numbers above profile peaks correspond to key) according to K-means cluster definition. Black lines indicate mean expression of genes belonging to the same cluster. (b) Specificity of lincRNA-encoding genes (left) and protein-coding genes (right) across 13 human lymphocyte populations (above columns); order of rows and columns based on K-means clustering; color intensity (key) indicates *z*-score log2-normalized raw FPKM counts estimated by Cufflinks software; numbers at top left indicate percent assigned to specific clusters (additional information,Supplementary Fig. 2a). (c) Analysis as in b of genes encoding receptors (top) and molecules involved in metabolic processes (bottom).

Figure 3: (**a**) Heat map of normalized expression values of lymphocyte signature lincRNAs selected on the basis of a difference in expression of >2.5-fold (relative to expression in all other subsets), intrapopulation consistency (expressed in at least three of five samples) and a *P* value of <0.05 (nonparametric Kruskal-Wallis test); normalized expression values were calculated as log2 ratio between expression in the lymphocyte subsets and in a panel of human lymphoid and nonlymphoid tissues of the Human BodyMap 2.0 project (additional information, Supplementary Table 1). (**b**) Expression of CD4$^+$ TH1 cell signature

lincRNAs (presented as in **a**). S, sense; AS, antisense. (**c**) Expression of newly identified and previously annotated lincRNAs (key) in human lymphocyte subsets and lymphoid or nonlymphoid human tissues (presented as the 2.5–97.5 percentile). (**d**) Quantitative RT-PCR analysis of the expression of TH1 cell signature lincRNAs by primary CD4$^+$naive cells (Naive), regulatory T cells (Treg) and TH1 cells (TH1) sorted from the peripheral blood mononuclear cells of healthy donors. (**e**) Quantitative RT-PCR analysis of the expression of TH1 cell signature lincRNAs over time in CD4$^+$ naive T cells differentiated in TH1- or TH2-polarizing conditions; results (average values) are presented as relative quantity (RQ) relative to expression at time zero. Data are from at least four experiments (**a**,**b**), one experiment with 63 independent samples (**c**), three independent experiments (**d**; average ± s.e.m.) or two independent experiments (**e**).

Figure 4: Semantic similarity scores for all gene-ontology (GO) term pairs clustered by a hierarchical clustering method (left), with adjusted *P* values for each GO term (middle), as well as common ancestors (right); red bars indicate gene-ontology terms with significant enrichment.

Figure 5: (**a**) Quantitative RT-PCR analysis of the expression of *linc-MAF-4* and *MAF* in activated CD4$^+$naive T cells differentiated in TH1- or TH2-polarizing conditions (additional information,Supplementary Fig. 4b,c). AU, arbitrary units. (**b**) Occupancy of H3K4me3 and RNA polymerase II at the *MAF* locus (top) or the control *IFNG* locus (bottom)

in CD4$^+$ naive T cells differentiated in TH1- or TH2-polarizing conditions at day 8 after activation, analyzed by chromatin immunoprecipitation followed by quantitative PCR and presented relative to input DNA. (**c**) Quantitative RT-PCR analysis of the expression of *linc-MAF-4* and *MAF* in activated CD4$^+$ naive T cells (in the absence of polarizing cytokines) 72 h after transfection of small interfering RNA (siRNA) targeting linc-MAF-4 or control (ctrl) siRNA. (**d**) Gene set–enrichment analysis, presented as enrichment score profiles for genes in activated CD4$^+$ naive T cells after transfection of siRNA targeting linc-MAF-4 or control siRNA compared with that of the CD4$^+$ TH1 cell reference gene set or the TH2 cell reference gene set, respectively. Nominal $P < 0.05$. (**e**) Quantitative RT-PCR analysis of the expression of *GATA3* and *IL4* transcripts in activated CD4$^+$ naive T cells transfected with siRNA as in **d**. *$P < 0.05$ and **$P < 0.01$ (one-tailed *t*-test). Data are representative of four independent experiments (**a**; average ± s.e.m.) or are from at least five (**b**, top) or ten (**b**, bottom) independent experiments (average and s.e.m.), six independent experiments (**c**,**e**; average and s.e.m.) or four independent experiments (**d**; average).

Figure 6: (**a**) Chromosome-conformation capture analysis of the interactions between a 'bait' region M1 (red line) at the 5′ end of *MAF* and 24 'prey' regions spanning the *linc-MAF-4–MAF* genomic locus (L1–L24; horizontal axis) in CD4$^+$ naive T cells differentiated in TH1-polarizing conditions at day 8 after activation. Top, organization of the genomic locus. (**b**) Abundance of linc-MAF-4 transcripts, as well as of

linc-00339, MALAT1 and RNU2.1 (cytoplasmic, nuclear and chromatin-associated control transcripts, respectively), in the cytoplasm, nucleus and chromatin of $CD4^+$ naive T cells differentiated in TH1-polarizing conditions at day 8 after activation. (**c**) RNA-immunoprecipitation assay of the interaction of LSD1 and EZH2 with *linc-MAF-4*, and with the controls *ACTB*, *RNU2.1*and a region upstream of the transcriptional start site of *linc-MAF-4* (*linc-MAF-4* ctrl), in $CD4^+$ naive T cells differentiated in TH1-polarizing conditions at day 8 after activation; results are presented relative to control immunoprecipitation. $*P < 0.05$ and $**P < 0.01$ (analysis of variance and Dunnet post-hoc test). (**d**) Occupancy of EZH2, H3K27me3 and LSD1 at the *MAF* locus in activated $CD4^+$naive T cells transfected with siRNA targeting linc-MAF-4 or control siRNA, analyzed by chromatin immunoprecipitation followed by quantitative PCR. $*P < 0.05$ (one-tailed *t*-test). (**e**) Model for linc-MAF-4-mediated repression of *MAF* in TH1 lymphocytes: when linc-MAF-4 is expressed, it recruits chromatin remodelers (i.e., LSD1 and EZH2) at the 5′ end of *MAF*, taking advantage of a DNA loop that brings the 5′ and 3′ ends of *linc-MAF-4* in close proximity to the 5′ end of *MAF*. Data are from three independent experiments (**a**,**b**; average and s.e.m.), six independent experiments (**c**; average and s.e.m.) or at least three independent experiments (**d**; average and s.e.m.).

Table 1

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

Figure 1

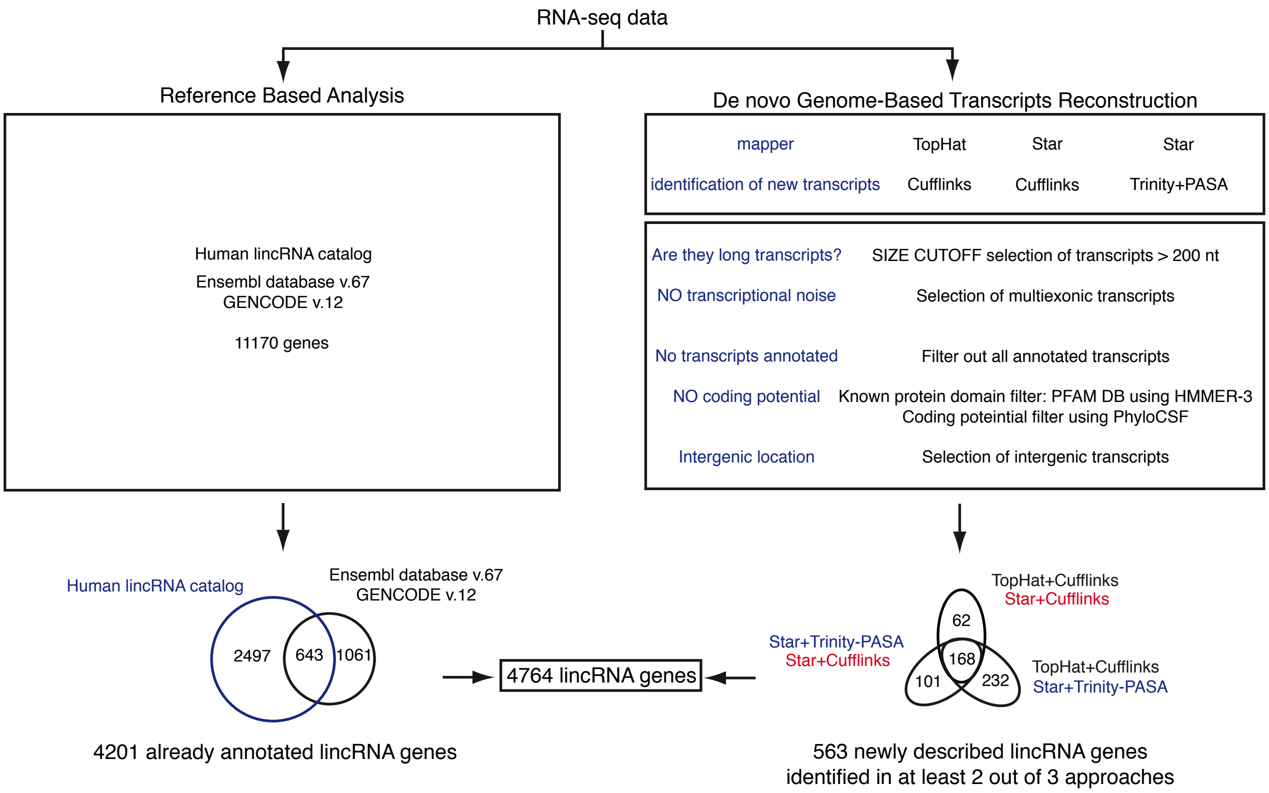

RNA-seq data
Reference Based Analysis
Human lincRNA catalog
Ensembl database v.67
GENCODE v.12
11170 genes
De novo Genome-Based Transcripts Reconstruction
mapper
TopHat
Star
Star
identification of new transcripts
Cufflinks
Cufflinks
Trinity+PASA
Are they long transcripts?
SIZE CUTOFF selection of transcripts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

Figure 2

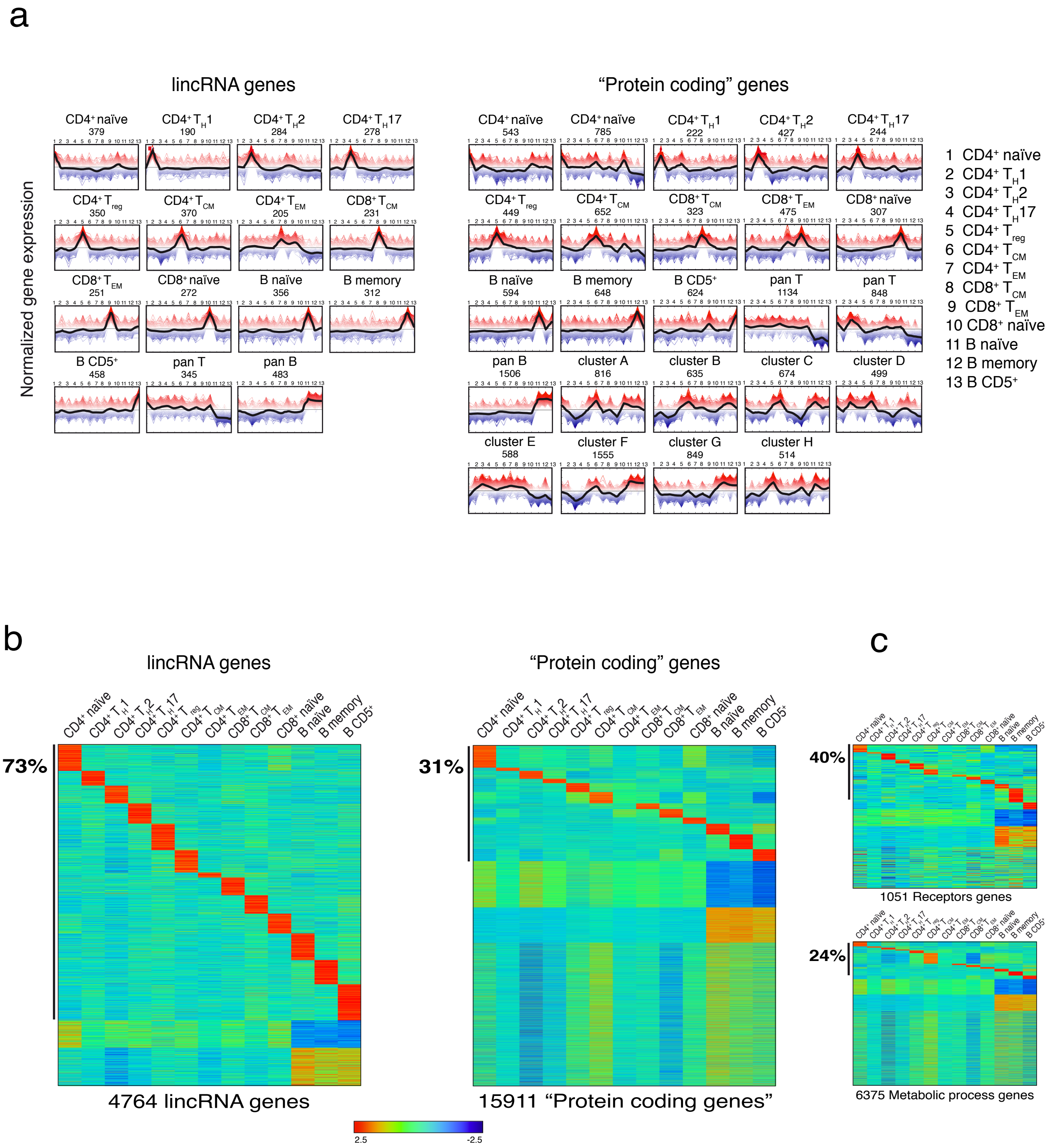

a
lincRNA genes
"Protein coding" genes
Normalized gene expression
CD4$^+$ naïve 379
CD4$^+$ T$_H$1 190
CD4$^+$ T$_H$2 284
CD4$^+$ T$_H$17 278
CD4$^+$ T$_{reg}$ 350
CD4$^+$ T$_{CM}$ 370
CD4$^+$ T$_{EM}$ 205
CD8$^+$ T$_{CM}$ 231
CD8$^+$ T$_{EM}$ 251
CD8$^+$ naïve 272
B naïve 356
B memory 312
B CD5$^+$ 458
pan T 345
pan B 483
CD4$^+$ naïve 543
CD4$^+$ naïve 785
CD4$^+$ T$_H$1 222
CD4$^+$ T$_H$2 427
CD4$^+$ T$_H$17 244
CD4$^+$ T$_{reg}$ 449
CD4$^+$ T$_{CM}$ 652
CD8$^+$ T$_{CM}$ 323
CD8$^+$ T$_{EM}$ 475
CD8$^+$ naïve 307
B naïve 594
B memory 648
B CD5$^+$ 624
pan T 1134
pan T 848
pan B 1506
cluster A 816
cluster B 635
cluster C 674
cluster D 499
cluster E 588
cluster F 1555
cluster G 849
cluster H 514
1 2 3 4 5 6 7 8 9 10 11 12 13
1 CD4$^+$ naïve
2 CD4$^+$ T$_H$1
3 CD4$^+$ T$_H$2
4 CD4$^+$ T$_H$17
5 CD4$^+$ T$_{reg}$
6 CD4$^+$ T$_{CM}$
7 CD4$^+$ T$_{EM}$
8 CD8$^+$ T$_{CM}$
9 CD8$^+$ T$_{EM}$
10 CD8$^+$ naïve
11 B naïve
12 B memory
13 B CD5$^+$
b
lincRNA genes
"Protein coding" genes
CD4$^+$ naïve
CD4$^+$ T$_H$1
CD4$^+$ T$_H$2
CD4$^+$ T$_H$17
CD4$^+$ T$_{reg}$
CD4$^+$ T$_{CM}$
CD4$^+$ T$_{EM}$
CD8$^+$ T$_{CM}$
CD8$^+$ T$_{EM}$
CD8$^+$ naïve
B naïve
B memory
B CD5$^+$
73%
31%
4764 lincRNA genes
15911 "Protein coding genes"
2.5
-2.5
c
40%
1051 Receptors genes
24%
6375 Metabolic process genes

Figure 3

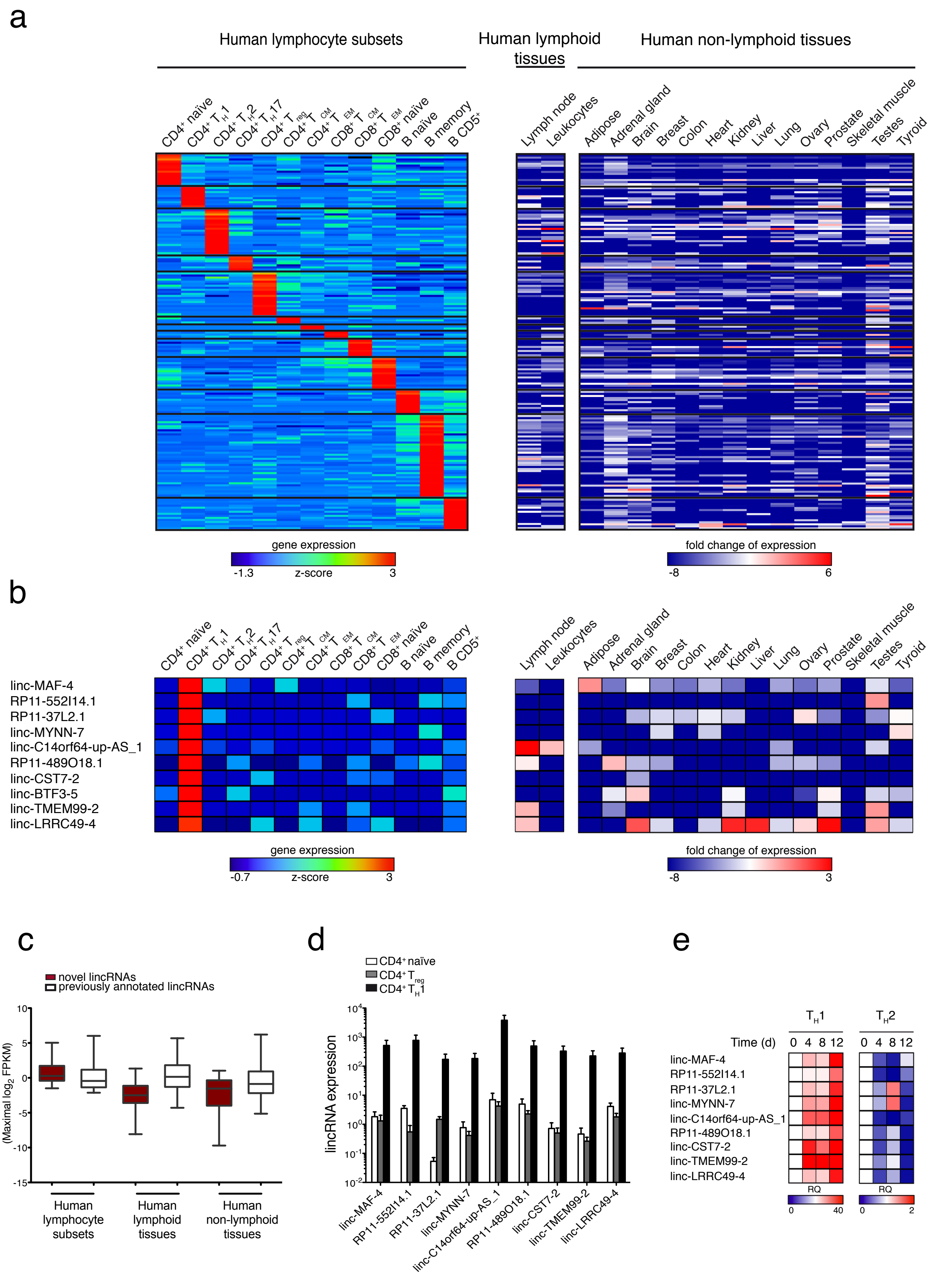

a
Human lymphocyte subsets
Human lymphoid tissues
Human non-lymphoid tissues
CD4+ naïve
CD4+ TH1
CD4+ TH2
CD4+ TH17
CD4+ Treg
CD4+ TCM
CD4+ TEM
CD8+ TCM
CD8+ TEM
CD8+ naïve
B naïve
B memory
B CD5+
Lymph node
Leukocytes
Adipose
Adrenal gland
Brain
Breast
Colon
Heart
Kidney
Liver
Lung
Ovary
Prostate
Skeletal muscle
Testes
Tyroid
gene expression
-1.3
z-score
3
fold change of expression
-8
6
b
linc-MAF-4
RP11-552I14.1
RP11-37L2.1
linc-MYNN-7
linc-C14orf64-up-AS_1
RP11-489O18.1
linc-CST7-2
linc-BTF3-5
linc-TMEM99-2
linc-LRRC49-4
gene expression
-0.7
z-score
3
fold change of expression
-8
3
c
novel lincRNAs
previously annotated lincRNAs
Gene expression (Maximal log2 FPKM)
10
5
0
-5
-10
-15
Human lymphocyte subsets
Human lymphoid tissues
Human non-lymphoid tissues
d
CD4+ naïve
CD4+ Treg
CD4+ TH1
lincRNA expression
10^4
10^3
10^2
10^1
10^0
10^-1
10^-2
linc-MAF-4
RP11-552I14.1
RP11-37L2.1
linc-MYNN-7
linc-C14orf64-up-AS_1
RP11-489O18.1
linc-CST7-2
linc-TMEM99-2
linc-LRRC49-4
e
TH1
TH2
Time (d)
0 4 8 12
0 4 8 12
linc-MAF-4
RP11-552I14.1
RP11-37L2.1
linc-MYNN-7
linc-C14orf64-up-AS_1
RP11-489O18.1
linc-CST7-2
linc-TMEM99-2
linc-LRRC49-4
RQ
0
40
RQ
0
2

Figure 4

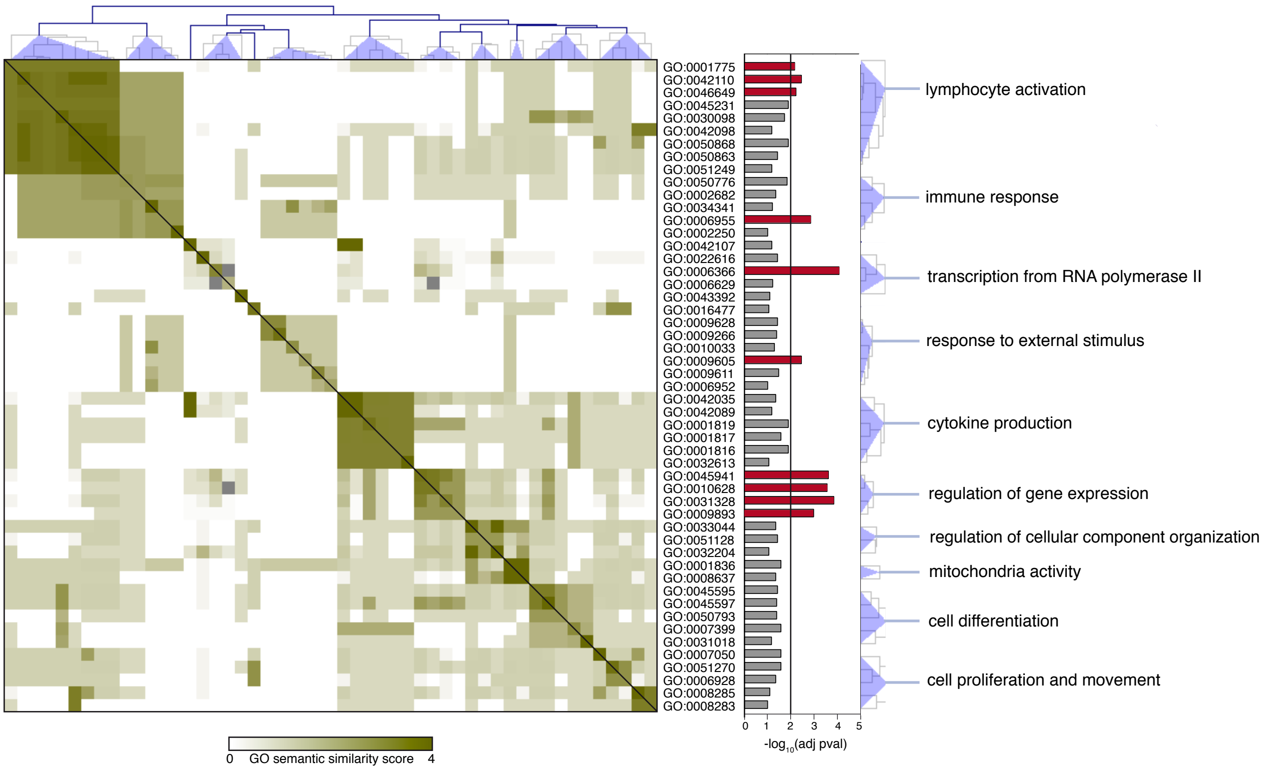
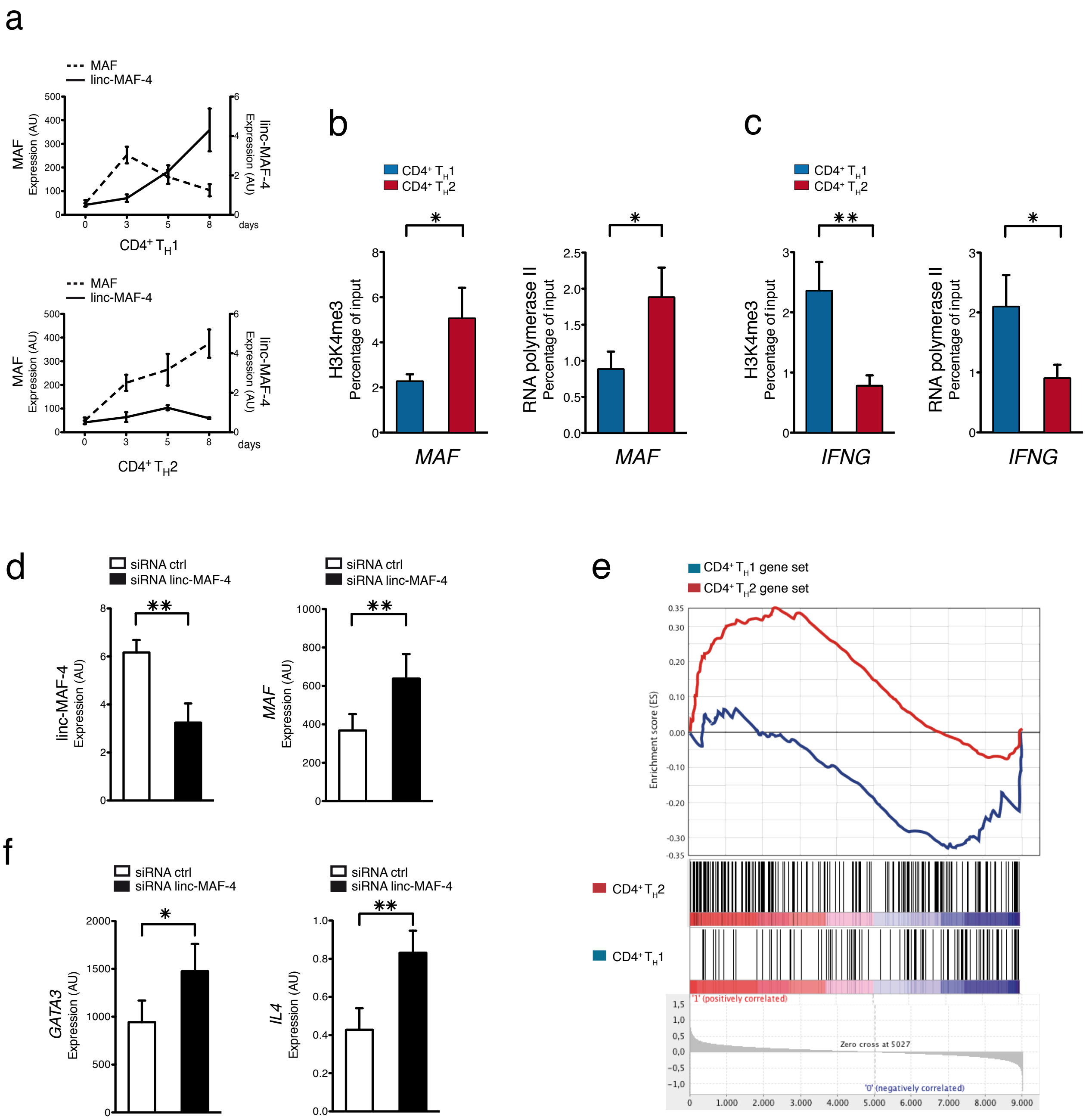
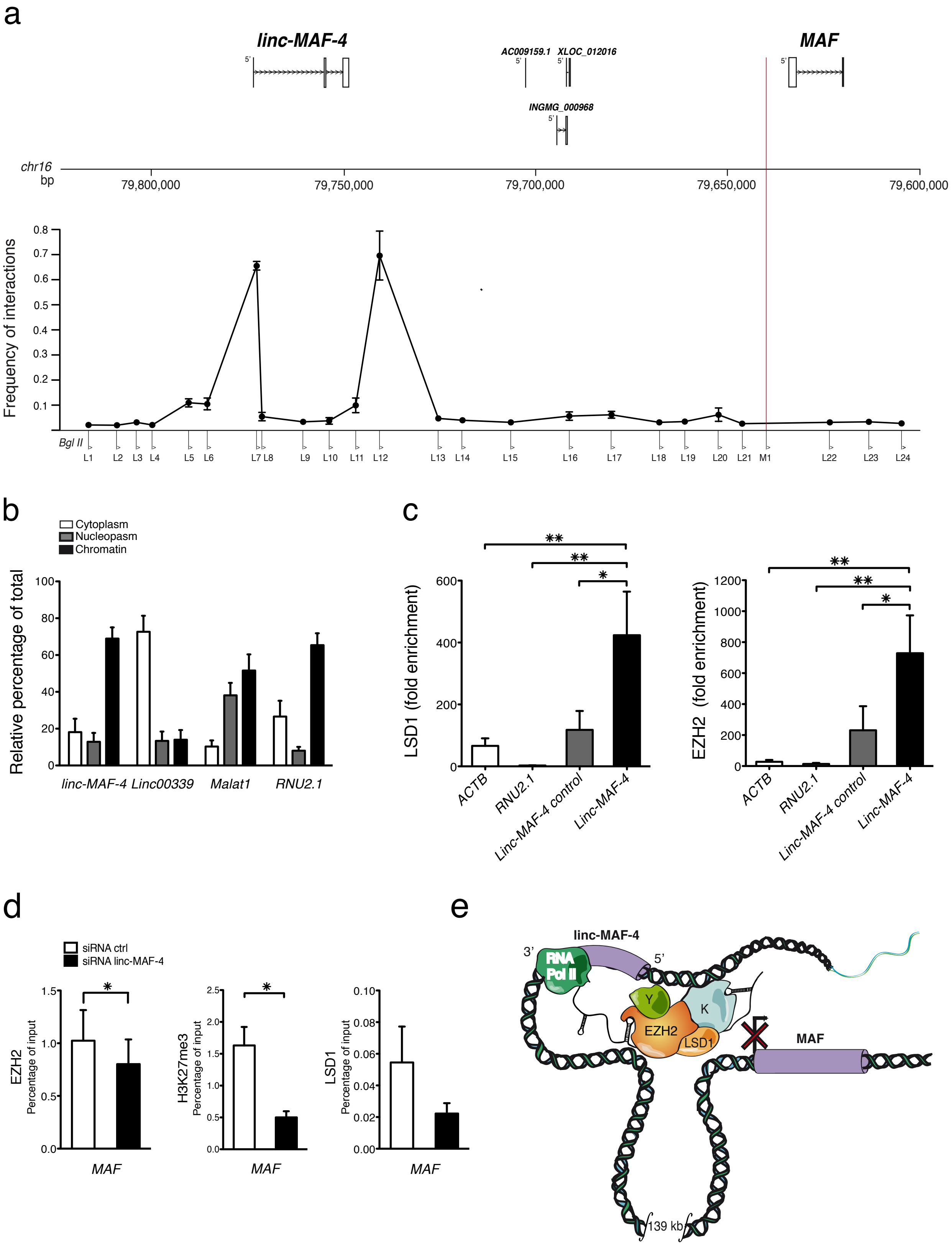

GO:0001775
GO:0042110
GO:0046649
GO:0045231
GO:0030098
GO:0042098
GO:0050868
GO:0050863
GO:0051249
GO:0050776
GO:0002682
GO:0034341
GO:0006955
GO:0002250
GO:0042107
GO:0022616
GO:0006366
GO:0006629
GO:0043392
GO:0016477
GO:0009628
GO:0009266
GO:0010033
GO:0009605
GO:0009611
GO:0006952
GO:0042035
GO:0042089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

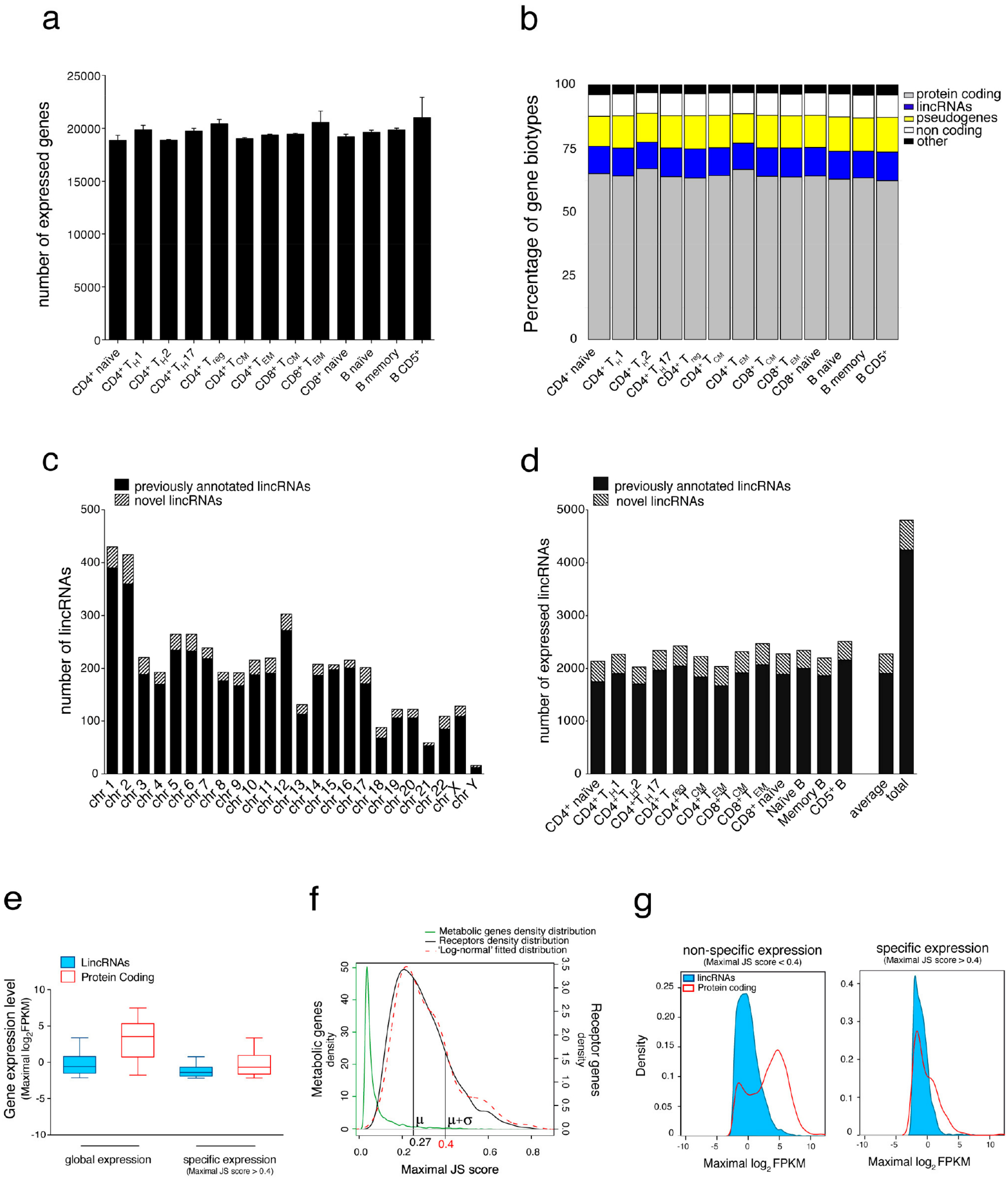
Figure 5



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

Figure 6


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



Supplementary Figure 1

Distribution and expression of lincRNAs in primary human lymphocytes subsets.

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

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

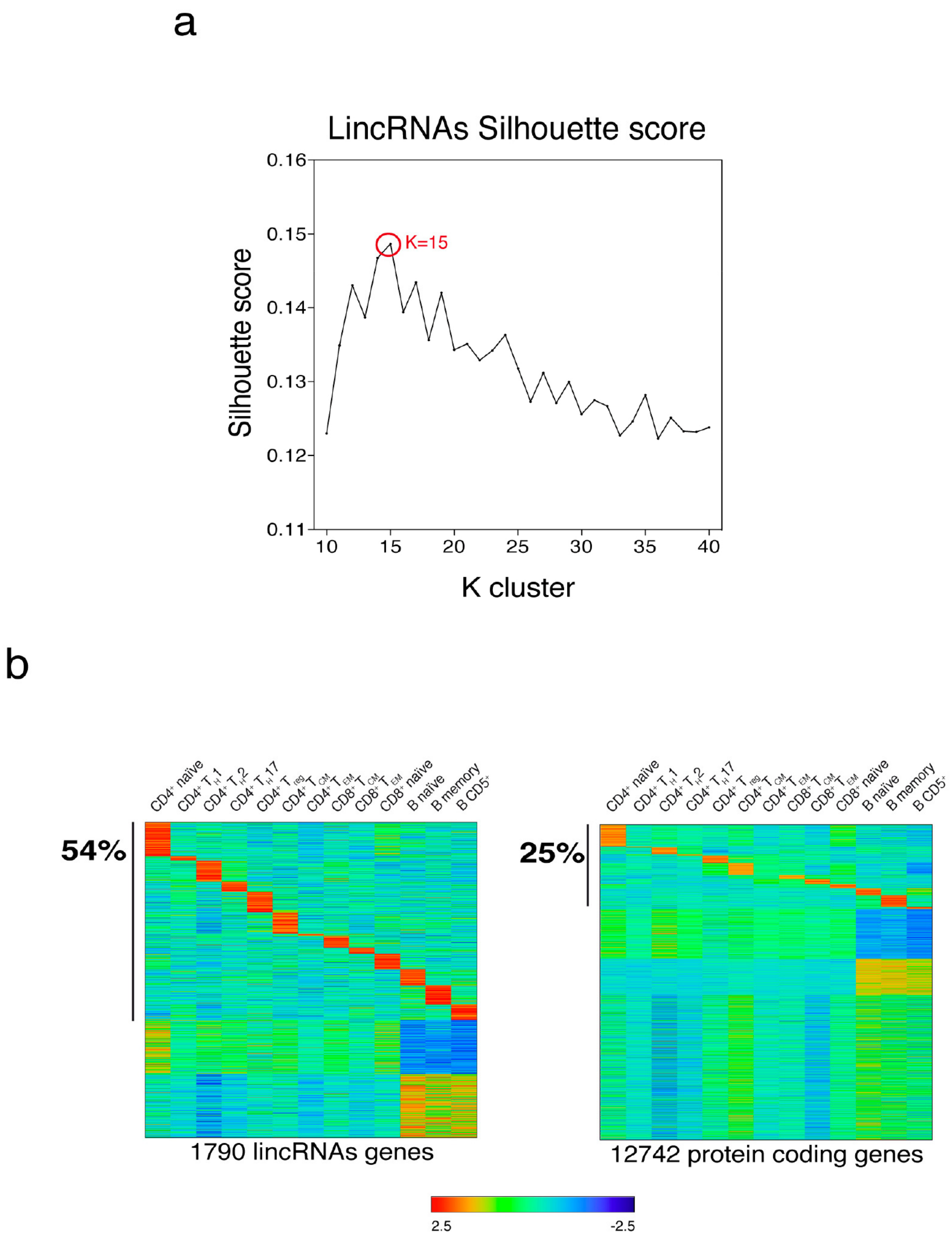
(c) Distribution of novel (striped) and previously annotated (black) lincRNAs in all human chromosomes

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

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

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

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

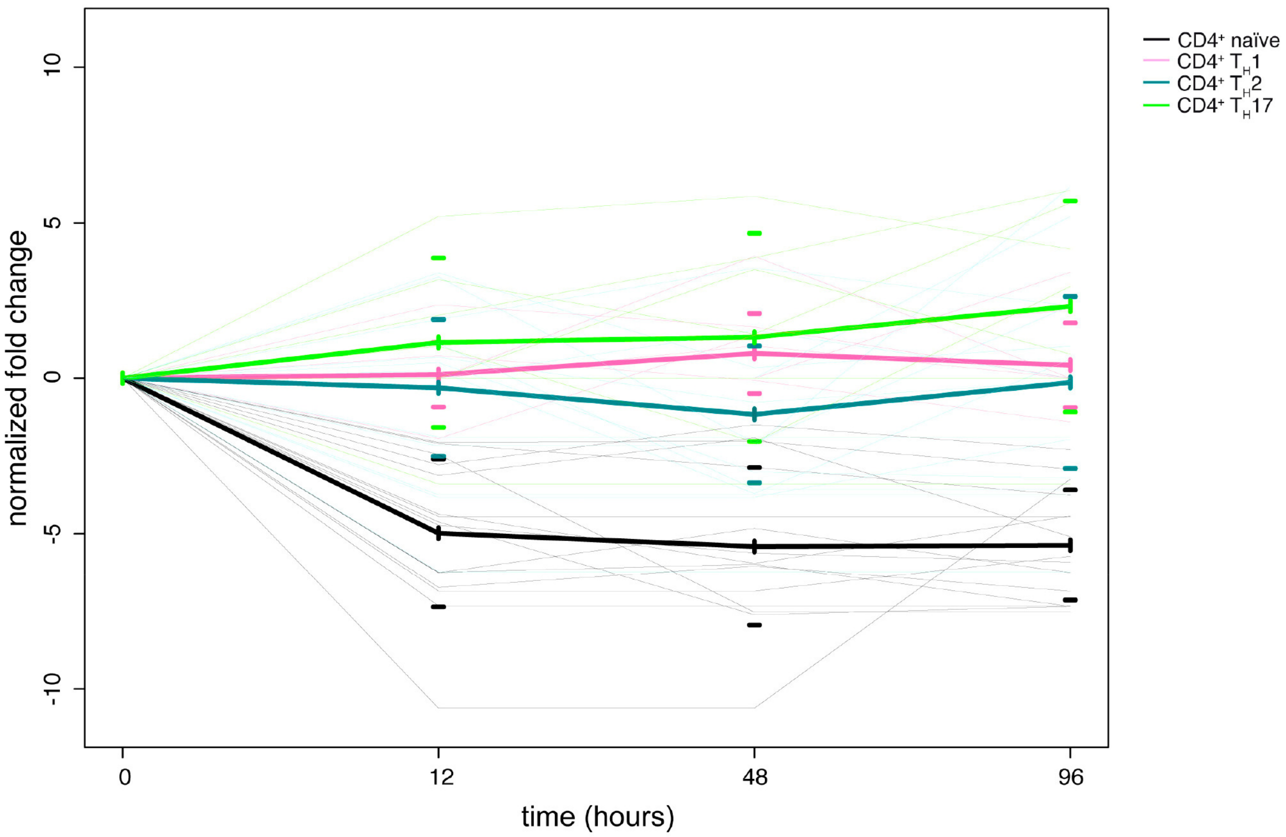


Supplementary Figure 2

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

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

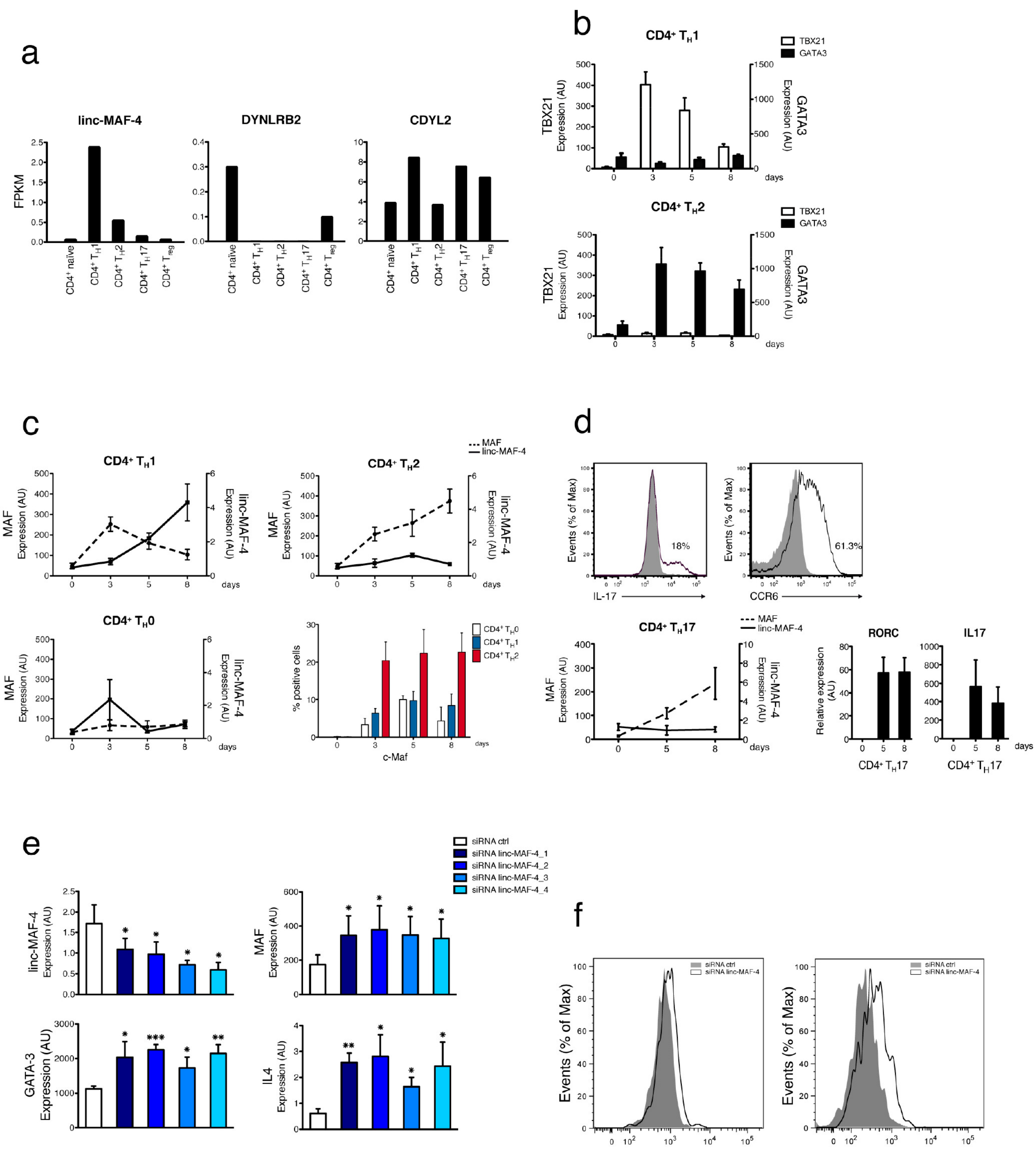
(b) Specificity of lincRNAs and protein coding genes (FPKM >1) by K-Means clustering across 13 human lymphocyte populations. Colour intensity represents the Z-score log2-normalized raw FPKM counts estimated by Cufflinks



Supplementary Figure 3

LincRNA signatures in a differentiation time course.

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



Supplementary Figure 4

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

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

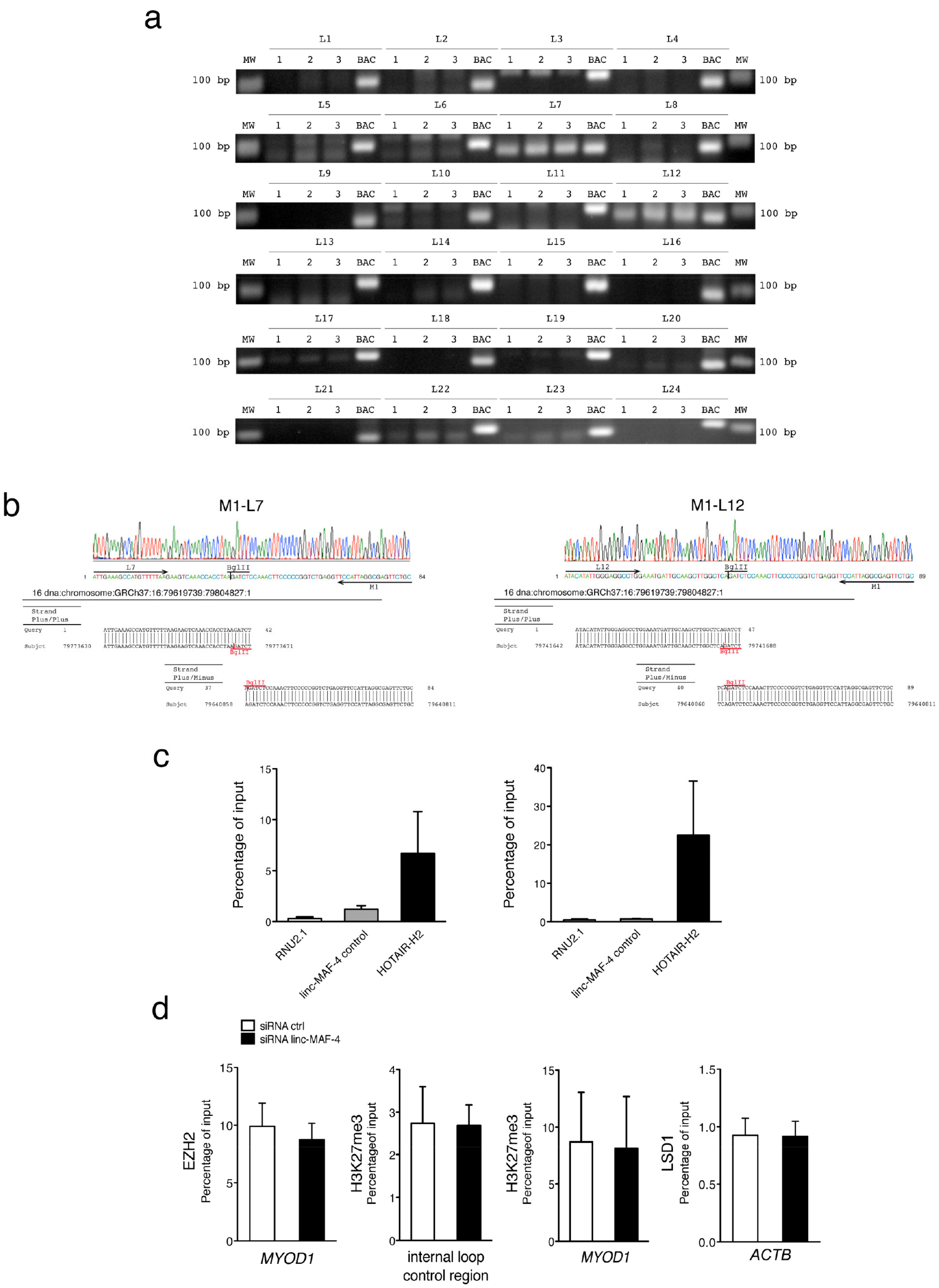
(b) Expression of TBX21 an GATA3 in activated CD4+ naïve T cells differentiated in T_H1 or T_H2 polarizing conditions assessed at different time points by RT-qPCR (average of four independent experiments ± SEM)

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

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

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

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


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

Supplementary Figure 5

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where:

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

and:

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

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

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

Specificity score of gene expression patterns: Jensen-Shannon divergence

The clustering results were integrated with an entropy-based methodology that assigns a cell-specificity score to each gene based on Jensen–Shannon divergence (Trapnell et al., 2010).
The JS divergence of two discrete probability distributions $p1$, $p2$, is defined to be:

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

where H is the entropy of a discrete probability distribution:
$p = (p_1, p_2 .. p_n), 0 \leq p_i \leq 1$ and $\sum_{i=1}^{n} p_i = 1$

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

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

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

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

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

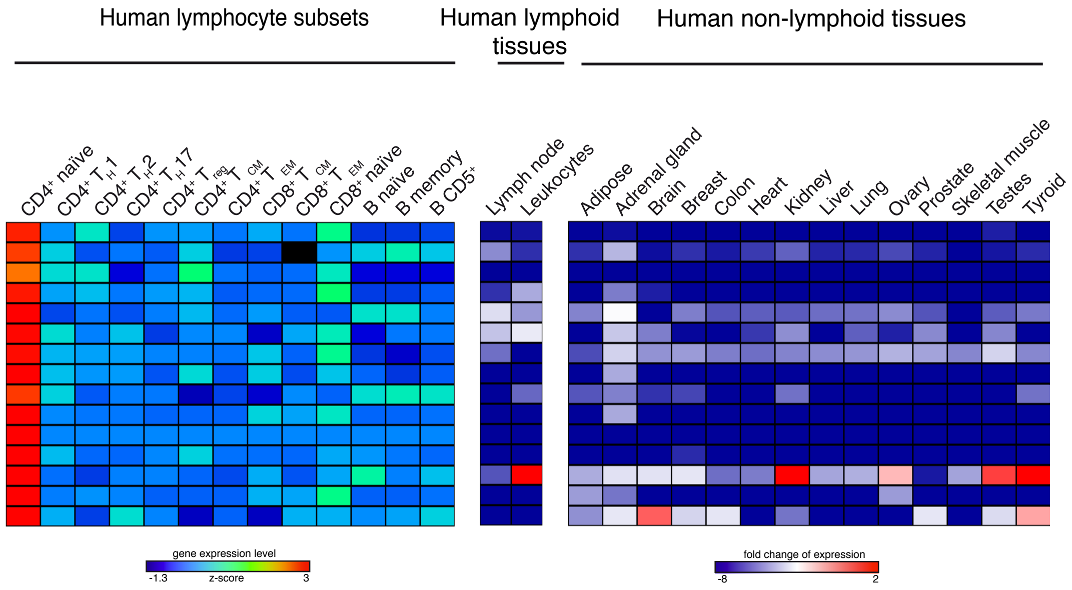
The threshold value for the JS score was calculated by considering one standard deviation away from the mean of the fitted distribution (0.4).

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

Table 1

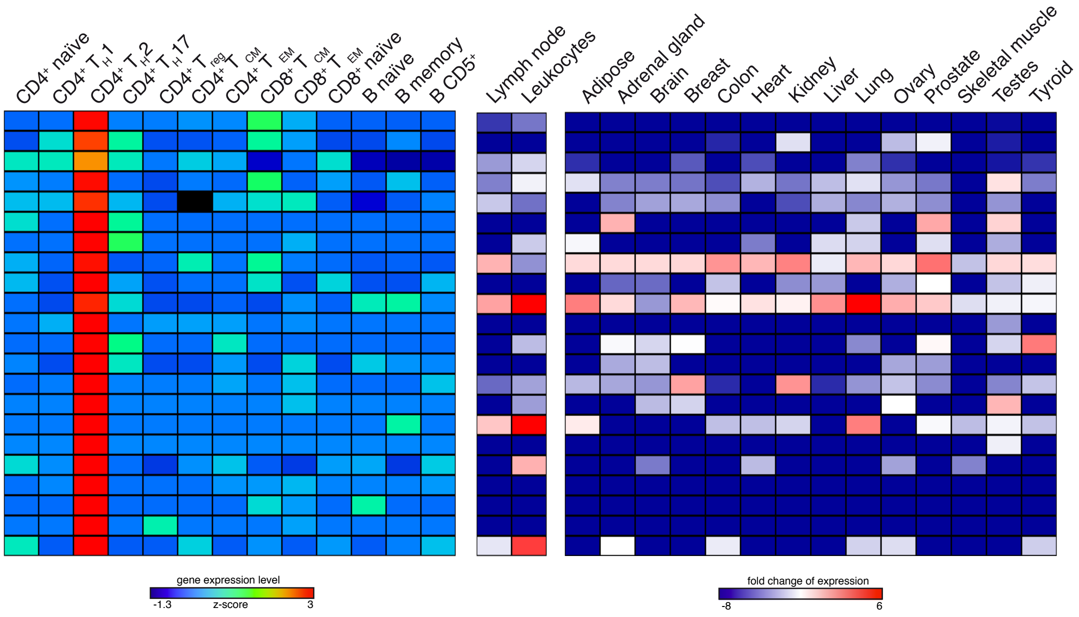
CD4⁺ naïve signature

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



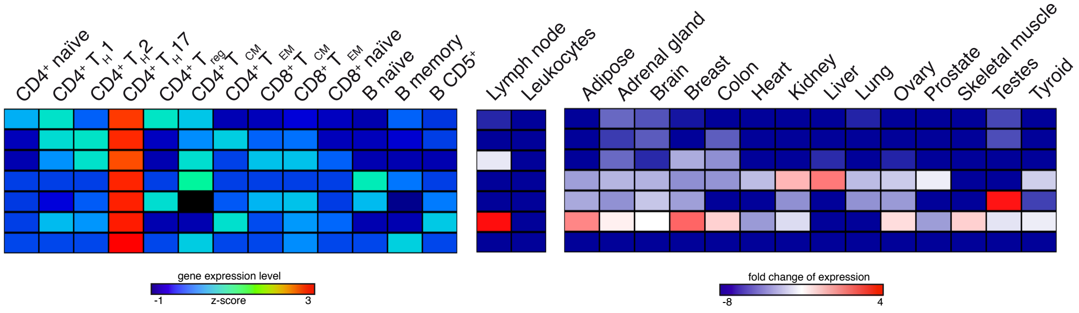
CD4⁺ T_H2 signature

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



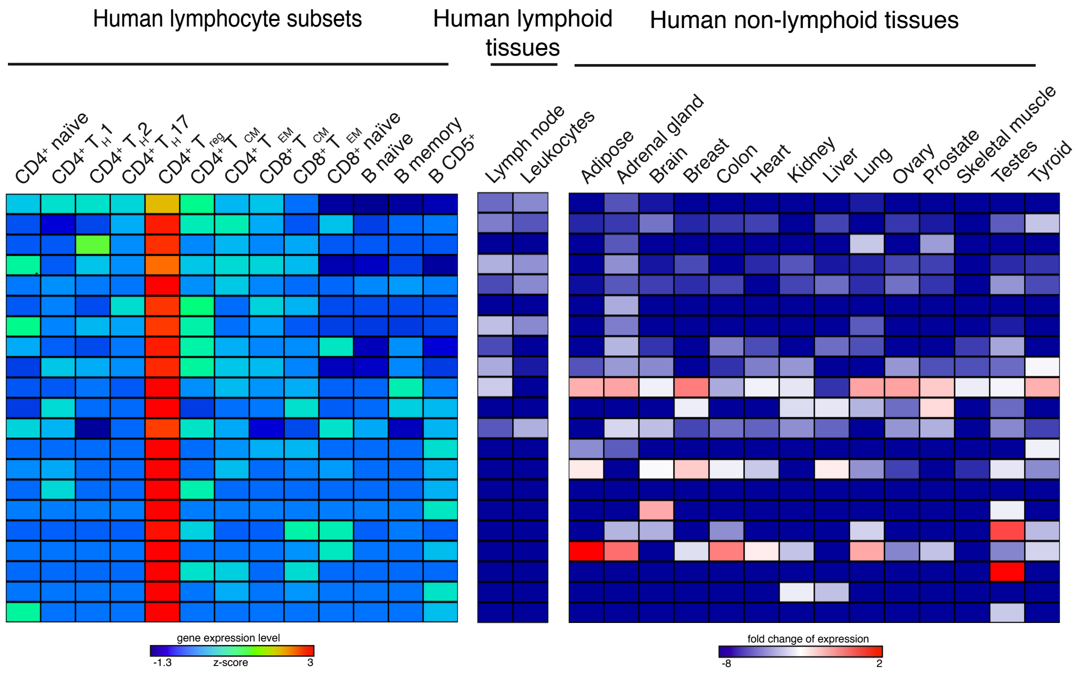
CD4⁺ T_H17 signature

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



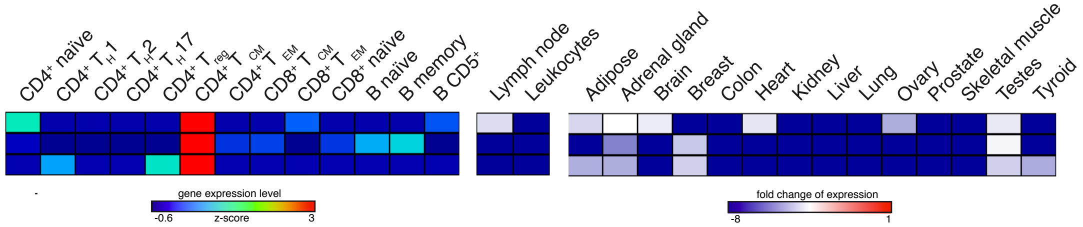
CD4+ T_{reg} signature

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



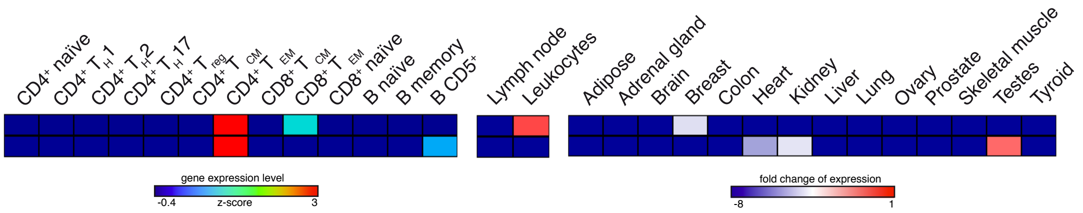
CD4+ T_{CM} signature

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



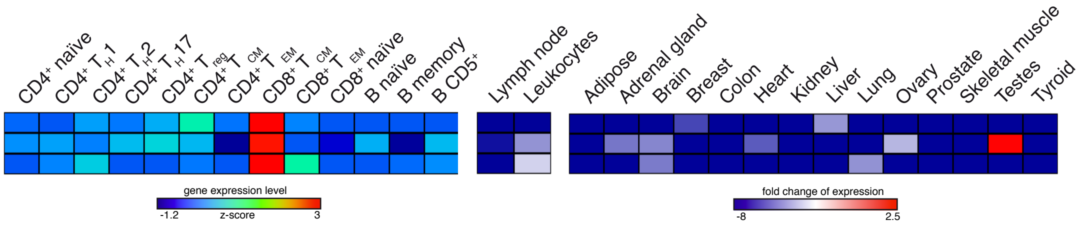
CD4+ T_{EM} signature

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



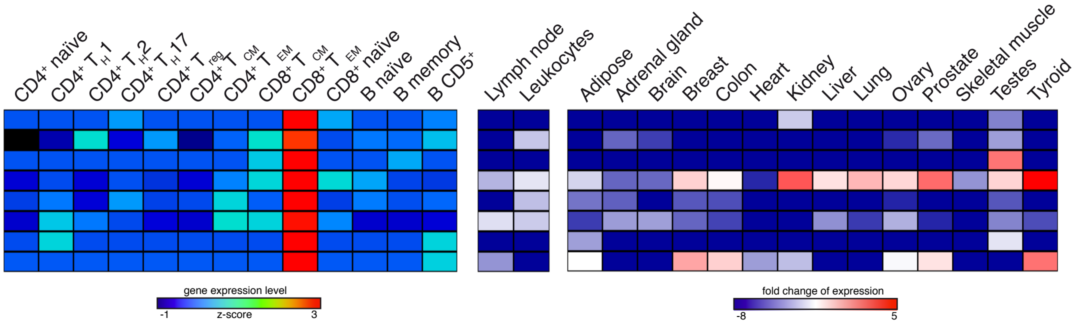
CD8+ T_{CM} signature

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



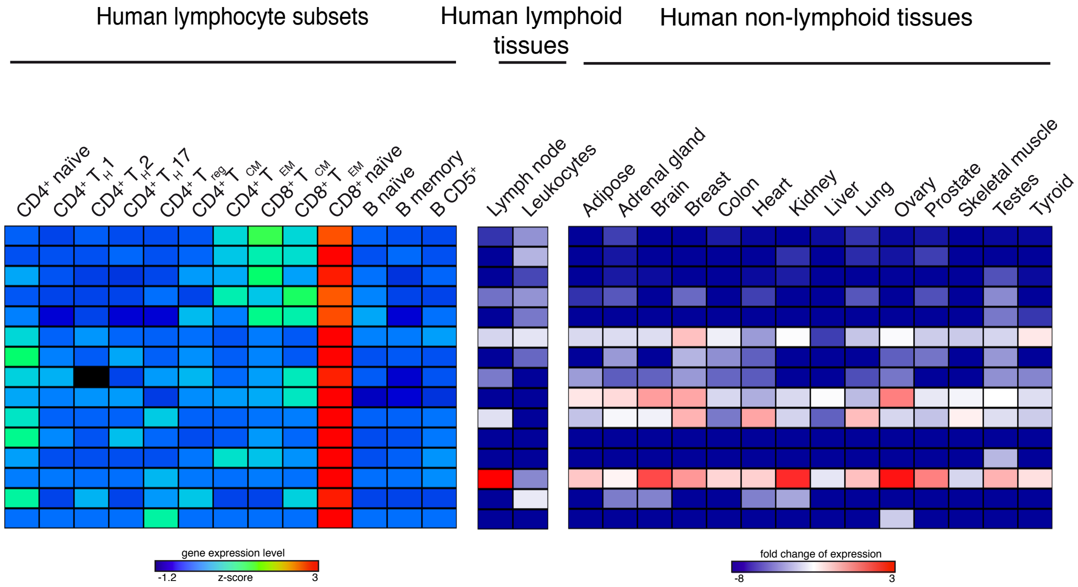
CD8+ T_{EM} signature

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



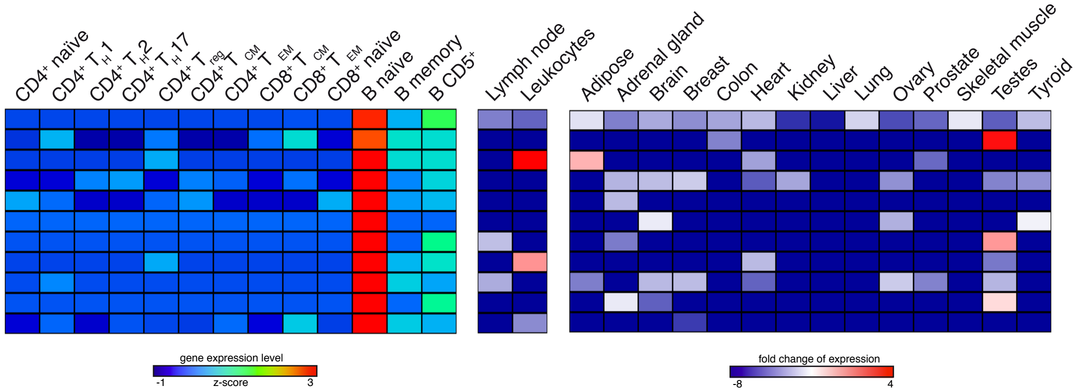
CD8⁺ naïve signature

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



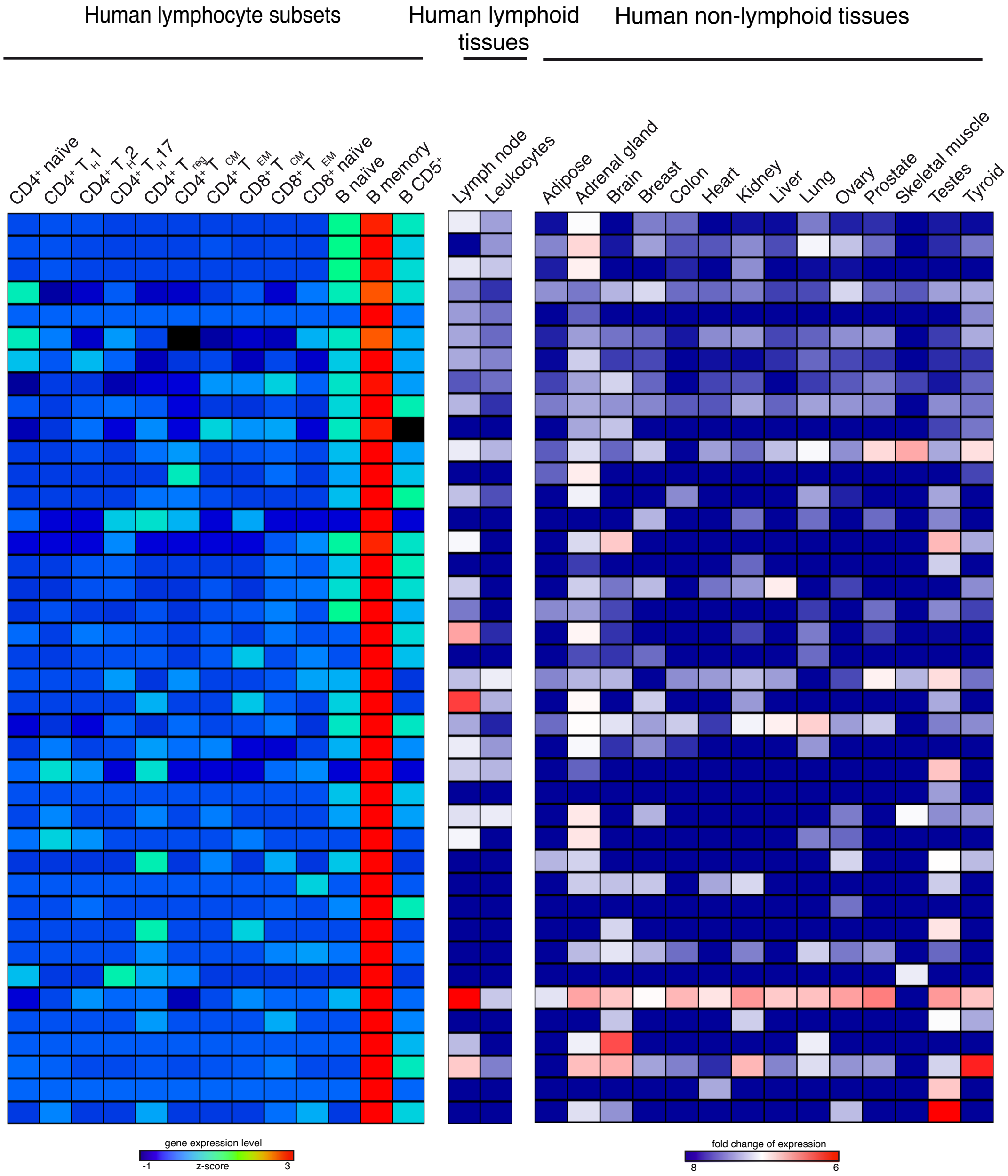
B Naïve signature

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



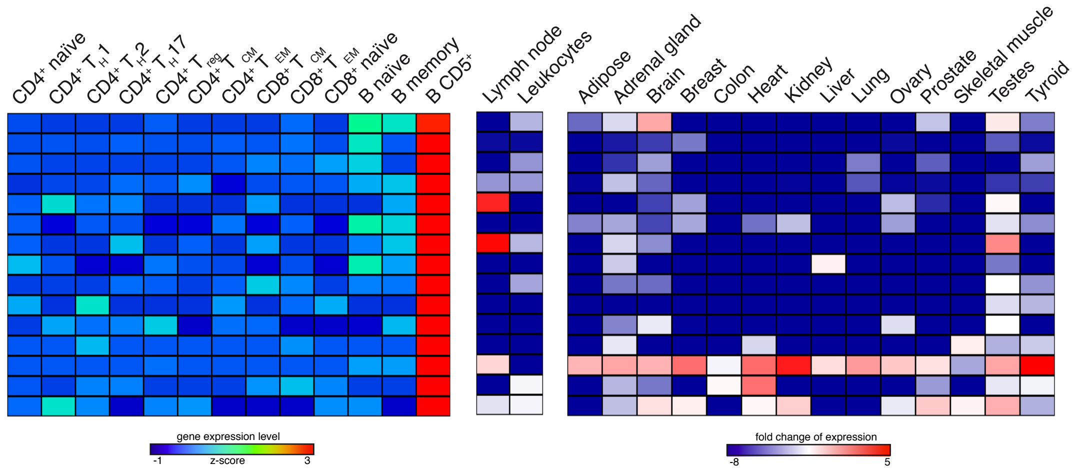
B memory signature

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



B CD5^{+} signature

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



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

Table 3

VALIDATION PRIMERS

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

siRNA

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

ChIP PRIMERS

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

3C PRIMERS

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

RIP PRIMERS

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

References

1. Zhu, J., Yamane, H. & Paul, W.E. Differentiation of effector CD4 T cell populations. Annu. Rev. Immunol. 28, 445–489 (2010).

2. Zhou, L., Chong, M.M. & Littman, D.R. Plasticity of CD4+ T cell lineage differentiation. Immunity 30, 646–655 (2009).

3. O'Shea, J.J. & Paul, W.E. Mechanisms underlying lineage commitment and plasticity of helper CD4+ T cells. Science 327, 1098–1102 (2010).

4. Kanno, Y., Vahedi, G., Hirahara, K., Singleton, K. & O'Shea, J.J. Transcriptional and epigenetic control of T helper cell specification: molecular mechanisms underlying commitment and plasticity. Annu. Rev. Immunol. 30, 707–731 (2012).

5. O'Connell, R.M., Rao, D.S., Chaudhuri, A.A. & Baltimore, D. Physiological and pathological roles for microRNAs in the immune system. Nat. Rev. Immunol. 10, 111–122 (2010).

6. Pagani, M. et al. Role of microRNAs and long-non-coding RNAs in CD4+ T-cell differentiation. Immunol. Rev. 253, 82–96 (2013).

7. Cobb, B.S. et al. T cell lineage choice and differentiation in the absence of the RNase III enzyme Dicer. J. Exp. Med. 201, 1367–1373 (2005).

8. Koralov, S.B. et al. Dicer ablation affects antibody diversity and cell survival in the B lymphocyte lineage. Cell 132, 860–874 (2008).

9. O’Connell, R.M. et al. MicroRNA-155 promotes autoimmune inflammation by enhancing inflammatory T cell development. Immunity 33, 607–619 (2010).

10. Rodriguez, A. et al. Requirement of bic/microRNA-155 for normal immune function. Science 316, 608–611 (2007).

11. Rossi, R.L. et al. Distinct microRNA signatures in human lymphocyte subsets and enforcement of the naive state in CD4+ T cells by the microRNA miR-125b. Nat. Immunol. 12, 796–803 (2011).

12. Cabili, M.N. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 25, 1915–1927 (2011).

13. Derrien, T. et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 22, 1775–1789 (2012).

14. Hrdlickova, B. et al. Expression profiles of long non-coding RNAs located in autoimmune disease-associated regions reveal immune cell-type specificity. Genome Med 6, 88 (2014).

15. Fatica, A. & Bozzoni, I. Long non-coding RNAs: new players in cell differentiation and development. Nat. Rev. Genet. 15, 7–21 (2014).

16. Guttman, M. et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458, 223–227 (2009).

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

18. Khalil, A.M. et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc. Natl. Acad. Sci. USA 106, 11667–11672 (2009).

19. Yoon, J.H. et al. LincRNA-p21 suppresses target mRNA translation. Mol. Cell 47, 648–655 (2012).

20. Kretz, M. et al. Control of somatic tissue differentiation by the long non-coding RNA TINCR. Nature 493, 231–235 (2013).

21. Poliseno, L. et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465, 1033–1038 (2010).

22. Sumazin, P. et al. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. Cell 147, 370–381 (2011).

23. Cesana, M. et al. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell 147, 358–369 (2011). 24. Pang, K.C. et al. Genome-wide identification of long noncoding RNAs in CD8+ T cells. J. Immunol. 182, 7738–7748 (2009).

25. Collier, S.P., Collins, P.L., Williams, C.L., Boothby, M.R. & Aune, T.M. Cutting edge: influence of Tmevpg1, a long intergenic noncoding RNA, on the expression of Ifng by Th1 cells. J. Immunol. 189, 2084–2088 (2012).

26. Gomez, J.A. et al. The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-γ locus. Cell 152, 743–754 (2013).

27. Carpenter, S. et al. A long noncoding RNA mediates both activation and repression of immune response genes. Science 341, 789–792 (2013).

28. Hu, G. et al. Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. Nat. Immunol. 14, 1190–1198 (2013).

29. Haas, B.J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8, 1494–1512 (2013).

30. Hart, T., Komori, H.K., LaMere, S., Podshivalova, K. & Salomon, D.R. Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics 14, 778 (2013).

31. Flicek, P. et al. Ensembl 2013. Nucleic Acids Res. 41, D48–D55 (2013). 32. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).

33. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111 (2009).

34. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515 (2010).

35. Rhind, N. et al. Comparative functional genomics of the fission yeasts. Science 332, 930–936 (2011).

36. Finn, R.D. et al. The Pfam protein families database. Nucleic Acids Res. 38, D211–D222 (2010).

37. Lin, M.F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics 27, i275–i282 (2011).

38. Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S. & Lander, E.S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell 154, 240–251 (2013).

39. Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F. & Mattick, J.S. Specific expression of long noncoding RNAs in the mouse brain. Proc. Natl. Acad. Sci. USA 105, 716–721 (2008).

40. Ørom, U.A. et al. Long noncoding RNAs with enhancer-like function in human cells. Cell 143, 46–58 (2010).

41. Al-Shahrour, F., Minguez, P., Vaquerizas, J.M., Conde, L. & Dopazo, J. BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. Nucleic Acids Res. 33, W472–W476 (2005).

42. Volders, P.J. et al. LNCipedia: a database for annotated human lncRNA transcript sequences and structures. Nucleic Acids Res. 41, D246–D251 (2013).

43. Ho, I.C., Lo, D. & Glimcher, L.H. c-maf promotes T helper cell type 2 (Th2) and attenuates Th1 differentiation by both interleukin 4-dependent and -independent mechanisms. J. Exp. Med. 188, 1859–1866 (1998).

44. Liu, X., Nurieva, R.I. & Dong, C. Transcriptional regulation of follicular T-helper (Tfh) cells. Immunol. Rev. 252, 139–145 (2013).

45. Sato, K. et al. Marked induction of c-Maf protein during Th17 cell differentiation and its implication in memory Th cell development. J. Biol. Chem. 286, 14963–14971 (2011).

46. Mattick, J.S. The genetic signatures of noncoding RNAs. PLoS Genet. 5, e1000459 (2009).

47. Klattenhoff, C.A. et al. Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. Cell 152, 570–583 (2013).

48. Cabianca, D.S. et al. A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. Cell 149, 819–831 (2012).

49. Tsai, M.C. et al. Long noncoding RNA as modular scaffold of histone modification complexes. Science 329, 689–693 (2010).

50. Kaneko, S. et al. Interactions between JARID2 and noncoding RNAs regulate PRC2 recruitment to chromatin. Mol. Cell 53, 290–300 (2014).

51. Bonnal, R.J. et al. Biogem: an effective tool-based approach for scaling up open source software development in bioinformatics. Bioinformatics 28, 1035–1037 (2012).

52. Rousseeuw, P.J. & Leroy, A.M. John Wiley & Sons. in Wiley Series in Probability and Mathematical Statistics Applied Probability and Statistics (Wiley, New York, 1987).

53. Bodega, B. et al. Remodeling of the chromatin structure of the facioscapulohumeral muscular dystrophy (FSHD) locus and upregulation of FSHD-related gene 1 (FRG1) expression during human myogenic differentiation. BMC Biol. 7, 41 (2009).

Chapter 3

Title

De novo Transcriptome profiling of highly purified human lymphocytes primary cells.

Authors

Raoul J.P. Bonnal[1], Valeria Ranzani[1], Alberto Arrigoni[1], Serena Curti[1], Ilaria Panzeri[1], Paola Gruarin[1], Sergio Abrignani[1,2], Grazisa Rossetti[1], Massimiliano Pagani[1,3]

Affiliations

1. Istituto Nazionale Genetica Molecolare ‘Romeo ed Enrica Invernizzi’, Via F. Sforza 35, 20122 Milan, Italy
2. Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Via Festa del Perdono 7, 20122 Milano, Italy
3. Department of Medical Biotechnology and Translational Medicine, Università degli Studi di Milano, Via Festa del Perdono 7, 20122 Milano, Italy

corresponding authors: Raoul J.P. Bonnal (bonnal@ingm.org), Grazisa Rossetti (rossetti@ingm.org) , Massimiliano Pagani (pagani@ingm.org)

Published on:

Scientific Data **2, Article number: 150051 (2015)**
doi:10.1038/sdata.2015.51

Abstract

To help better understand the role of long noncoding RNAs in the human immune system, we recently generated a comprehensive RNA-seq data set using 63 RNA samples from 13 subsets of T (CD4$^+$ naive, CD4$^+$ T$_H$1, CD4$^+$ T$_H$2, CD4$^+$ T$_H$17, CD4$^+$ T$_{reg}$, CD4$^+$ T$_{CM}$, CD4$^+$ T$_{EM}$, CD8$^+$ naive, CD8$^+$ T$_{CM}$, CD8$^+$ T$_{EM}$) and B (Naive B, Memory B, CD5+ B) lymphocytes. The number of biological replicates for each subset was 5 except for CD8+ CM and B_CD5 populations that included 4 replicates. RNA-Seq data were generated by an Illumina HiScanSQ sequencer using the TruSeq v3 Cluster kit. 2.193 billion of paired-ends reads, 2 x 100 bp, were sequenced and after filtering a total of 1.7 billion reads were mapped. Using different de novo transcriptome reconstruction techniques over 500 previously unkonwn lincRNAs were identified. The current data set could be exploited to drive the functional characterization of lincRNAs, identify novel genes and regulatory networks associated with specific cells subsets of the human immune system.

Background & Summary

With emerging technologies, it is becoming evident that the vast majority of the genome is transcribed (the so-called “dark matter of the genome”) and produces a diverse population of non-protein-coding RNAs (ncRNAs), including long non-coding RNAs (lncRNAs). LncRNAs are transcripts of more than 200 base pairs in length that are often expressed with higher cell-specificity compared to protein-coding genes[1] despite having lower expression levels. LncRNAs fold in functional domains that allow them to interact with other RNA molecules, DNA and proteins exerting a plethora of different functions in the cells, as chromatin remodeling (XIST, HOTAIR), transcriptional activation or repression, competition with microRNAs (linc-MD1, PTEN ceRNAs), splicing, RNA trafficking, mRNA stability, imprinting, gene-dosage compensation and translation (lncRNA-21p), among others[2]. LncRNAs are also frequently expressed only in specific developmental stages, hinting to their involvement in cell fate determination[1]. Moreover, lncRNAs have been implicated in the maintenance of stem cell pluripotency and differentiation[3], in the establishment of the cardiovascular lineage (12) and in the control of somatic tissue differentiation[4]. Altogether these findings clearly point out the fundamental role of lncRNAs in the control of cell differentiation and in the maintenance of cell identity. Indeed in the mouse immune system lncRNAs expression changes during naive to memory $CD8^+$ T cell differentiation[5] and during naive $CD4^+$ T cells differentiation into distinct helper T cell lineages[6] and our results on human $CD4^+$ T lymphocytes specific lncRNAs[7] are in agreement with the findings in mice. In this work 63 RNA samples from 13 subsets of T ($CD4^+$ naive, $CD4^+$ T_H1, $CD4^+$ T_H2, $CD4^+$ T_H17, $CD4^+$ T_{reg}, $CD4^+$ T_{CM}, $CD4^+$ T_{EM}, $CD8^+$ naive, $CD8^+$ T_{CM}, $CD8^+$ T_{EM}) and B (Naive B, Memory B, CD5+ B) lymphocytes were collected. The hierarchy of T and B cells during differentiation of the analyzed subsets is depicted in Fig. 1a as well as the number of biological replicates for each cell population. After RNA-seq sequencing the exploitation of different *de novo* transcriptome reconstruction led to the identification of over 500 previously unknown lincRNAs[7]. The general experimental design is shown in Fig. 1b. As recent findings suggest that lncRNAs might contribute to the definition of lymphocytes identity and to the modulation of their functional plasticity, our data set could be used as a resource to guide the validation and functional characterization of lincRNAs and to identify genes and regulatory networks associated with specific cells subsets of the human immune system.

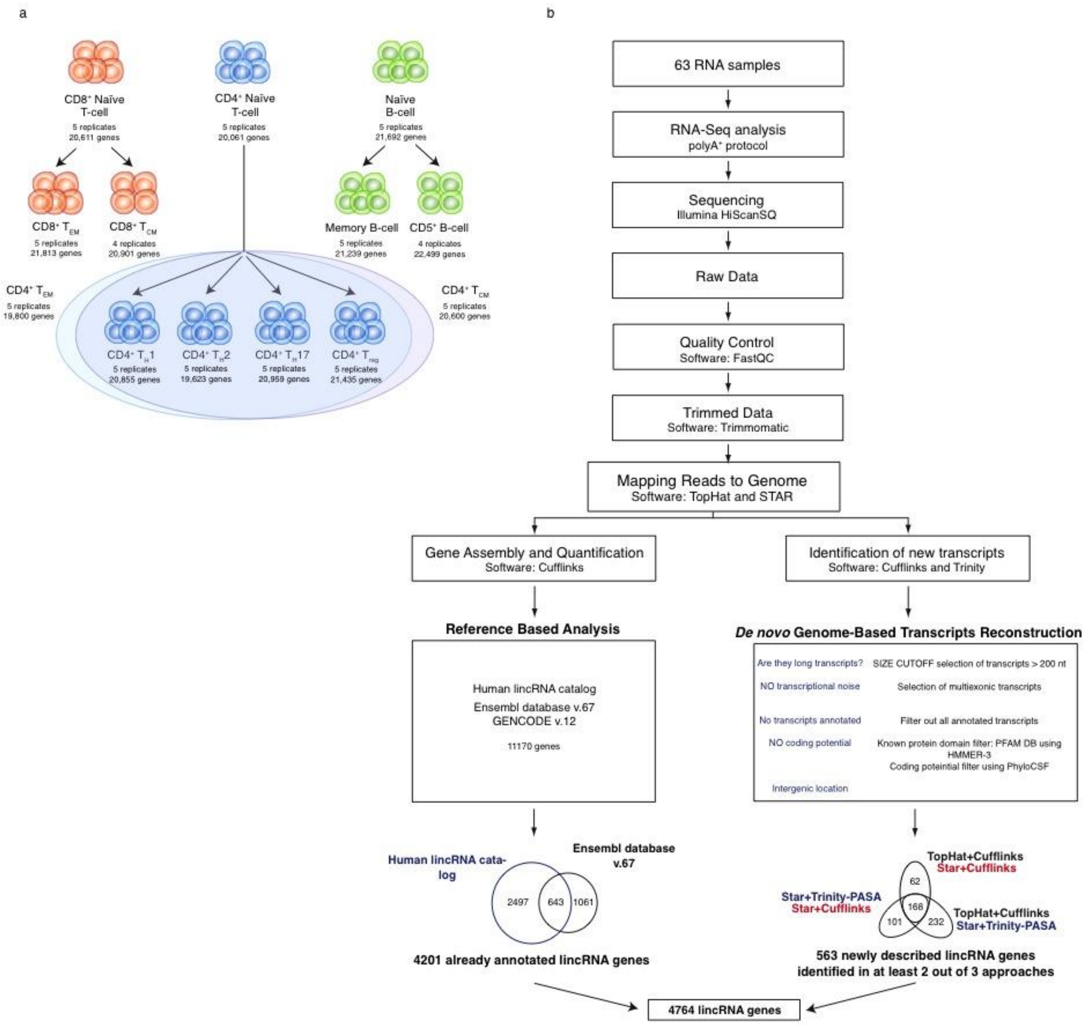


Figure 1 Description of the study with (a) Hierarchical representation of the different cell subset originating from hematopoietic stem cells. In this study 13 human primary lymphocyte subsets were profiled: CD4 Naïve; CD4 Th1; CD4 Th2; CD4 Th17; CD4 Treg; CD4 TEM; CD4 TCM; CD8 Naïve; CD8 TEM; CD8 TCM; B Naïve; B Memory; CD5 B. The number of biological replicates and the expressed genes (FPKM > 0.21) for each

population is indicated. The total number of samples profiled in this study is 63. (b) General overview of the bioinformatic steps and approaches used for the identification of novel lincRNAs.

Methods

Purification of primary immunological cell subsets

These methods are expanded from the descriptions in our previous article[7]. Blood buffy coat cells of healthy donors were obtained from Fondazione Istituto di Ricovero e Cura a Carattere Scientifico Ca'Granda Ospedale Maggiore Policlinico in Milan, and peripheral blood mononuclear cells were isolated by ficoll-hypaque density- gradient centrifugation. The ethical committee of Fondazione Istituto di Ricovero e Cura a Carattere Scientifico Ca'Granda Ospedale Maggiore Policlinico approved the use of peripheral blood mononuclear cells from healthy donors for research purposes, and informed consent was obtained from subjects. Human blood primary lymphocyte subsets were purified to a purity of >95% by cell sorting through the use of various combinations of surface markers (see Table 1).

Table 1 Purification and RNA-Seq of human primary lymphocyte subsets

Subset	Purity (%)	Sorting phenotype	Donors
$CD4^+$ naïve	99,8 ± 0,1	$CD4^+$ $CCR7^+$ $CD45RA^+$ $CD45RO^-$	5

CD4$^+$ T$_H$1	99,9 ± 0,05	CD4$^+$ CXCR3$^+$	5
CD4$^+$ T$_H$2	99,7 ± 0,3	CD4$^+$ CRTH2$^+$ CXCR3$^-$	5
CD4$^+$ T$_H$17	99,1 ± 1	CD4$^+$ CCR6$^+$ CD161$^+$ CXCR3$^-$	5
CD4$^+$ T$_{reg}$	99,0 ± 0,8	CD4$^+$ CD127$^-$ CD25$^+$	5
CD4$^+$ T$_{CM}$	98,4 ± 2,8	CD4$^+$ CCR7$^+$ CD45RA$^-$ CD45RO$^+$	5
CD4$^+$ T$_{EM}$	95,4 ± 5,5	CD4$^+$ CCR7$^-$ CD45RA$^-$ CD45RO$^+$	5
CD8$^+$ T$_{CM}$	98,3 ± 0,8	CD8$^+$ CCR7$^+$ CD45RA$^-$ CD45RO$^+$	4
CD8$^+$ T$_{EM}$	96,8 ± 0,9	CD8$^+$ CCR7$^-$ CD45RA$^-$ CD45RO$^+$	5
CD8$^+$ naïve	99,3 ± 0,2	CD8$^+$ CCR7$^+$ CD45RA$^+$ CD45RO$^-$	5
B naïve	99,9 ± 0,1	CD19$^+$ CD5$^-$ CD27$^-$	5
B memory	99,1 ± 0,8	CD19$^+$ CD5$^-$ CD27$^+$	5
B CD5$^+$	99,1 ± 0,8	CD19$^+$ CD5$^+$	4

Purity achieved (middle left) by the sorting of 13 human lympocyte subsets (isolated from peripheral blood lymphocytes of four to five different donors per subset) by various surface marker combinations (Sorting phenotype). T$_{reg}$,regulatory T cells; T$_{CM}$, central

memory T cells; T_{EM}, effector memory T cells; B, B cells. Data are representative of at least four experiments (mean ± s.d. for purity).

RNA isolation and RNA sequencing.

Total RNA was isolated with a mirVana Isolation Kit (Ambion). Libraries for Illumina sequencing were constructed from 100 ng of total RNA with the Illumina TruSeq RNA Sample Preparation Kit v2 (Set A). The libraries generated were loaded on to the cBot automated clonal amplification system (Illumina) for clustering on a HiSeq Flow Cell v3. The libriaries clustered on a HiSeq Flow Cell v3 were then sequenced with a HiScanSQ optical imaging system (Illumina). A paired-end run (with a read length of 100 bases) was performed with an SBS Kit v3 DNA sequencing kit (Illumina). Real-time analysis and base calling was performed with HiSeq Control Software Version 1.5 (Illumina). CASAVA 1.8.2 (Illumina) software was used to demultiplex reads into specific sample and groups, the software was configured to operate with "--mismatches='1'" allowing one mismatch during the identification of the indexes (Data Citation 1).

RNA-seq trimming and mapping.

To improve sequence quality, samples data were cleaned by Trimmomatic[8] using the following parameters (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15

MINLEN:50) giving in input the forward and reverse “fastq” sequences for each sample. Only the reads that passed the quality or length threshold on both strands were considered for mapping. The whole data set was aligned to human genome assembly GRCh37 (Genome Reference Consortium Human Build 37) using both TopHat[9] (version 1.4.1) and STAR[10] (version 2.2.0). The reference genome was indexed using Bowtie[11] (version 0.12.9) for TopHat alignment. Both TopHat and STAR were used with default parameters; only for TopHat we specified the mate-inner-dist parameter for each sample of our data set please see the associated Metadata Record. RNA-seq data from the Illumina Human BodyMap 2.0 project (Data Citation 2) consisting of 16 human tissues were downloaded, processed and mapped using to the same criteria.

Reference annotation.

Ensembl database (version 67 from May 2012) annotation was integrated with a previously published catalogue of lincRNAs[1] (Data Citation 3) using Cuffcompare which is provided by the Cufflinks[12] (version 2.1.1) suite. BioMart was used to categorize Ensembl annotation in different classes by their biotype: ‘lincRNA’ (5,804 genes), protein-coding genes (21,976 genes), receptor-encoding using GO term GO:000487 (2,043 genes encoding molecules with receptor activity function) and the class of genes encoding molecules involved in metabolic processes corresponding to GO term GO:0008152 (7,756 genes). The final reference annotation consisted of 195,392 transcripts that referred to 62,641 genes, 11,170 of which were non-redundant lincRNA-encoding genes.

De novo genome-based transcripts reconstruction.

To identify putative novel genes, specifically expressed in our datasets and not yet annotated, we combined multiple tools and their outputs following a *de novo* genome-based transcripts reconstruction procedure. Samples were aggregated in meta datasets corresponding to the 13 lymphocyte populations. These meta datasets were aligned to the reference genome using two mappers: TopHat and STAR. The resulting 26 alignments were used as independent inputs for Cufflinks configured to use the RABT[13] assembler for the identification of novel transcripts. The following parameters were used in combination with Cufflinks: "-g" to guide the assembly by the reference annotation, "-u" multi-read and "-b" fragment-bias correction to improve the accuracy of transcripts abundance estimation. With these approaches we identified about 3 × 10^4 to 5 × 10^4 previously unknown transcripts for each lymphocyte population. A third approach was the Genome-guided Trinity[14] pipeline (Supplementary File 1: example of command lines) (release 2012-10-05, http://trinityrnaseq.github.io/#genome_guided) that generates de novo transcripts by local assembly on previously mapped reads from specific locations. We used STAR instead of the Trinity's default aligner GSNAP[15], as it performed better in terms of both accuracy and computing time. For the first alignment phase STAR was used with the default parameters. The "Genome-guided Trinity" suite was used with the parameters suggested in the main documentation (default). Each candidate transcript was then processed via the Program to Assemble Spliced Alignments[16] (PASA, http://pasapipeline.github.io/). PASA is a genome annotation tool that reconstructs the complete transcript and gene structures, resolves incongruences derived from transcript misalignments and alternatively splices events, refines the reference annotation and proposes new transcripts and genes in case no previous annotation can explain the new data. PASA was configured to use STAR as aligner. We recompiled STAR to enable it for handling long reads (putative transcripts); the file "IncludeDefine.h", from the source code, was modified setting the variable "MAX_READ_LENGTH" to a value of "100000". Recompiling the source tree using the GNU "make" utility with the command "make STARlong" generated the desired modified binary version of STAR.

Identification of previously unknown lincRNA-encoding genes.

Data generated by the three different approaches, Trinity/Cufflinks; STAR/Cufflinks; STAR/Trinity, were separately processed to identify unknown lincRNA-encoding genes.

The three de novo approaches applied to each lymphocyte population generate transcripts and genes without prior knowledge on they ability to encode for proteins or not. In order to identify only the putative novel lincRNAs, known transcripts and previously unknown isoforms of already annotated genes were filtered out. To perform this filtering we compared the reference annotation with the datasets produced by each approach using a custom script; this can be done also using consolidated tools as the UCSC bedtools[17] or Cuffcompare. Transcriptional noise and low polymerase fidelity can create artifactual transcripts therefore only multi-exonic transcripts longer than 200 bases were retained in our analysis. Protein family domains available from Pfam[18] (PfamA and PfamB) database (release 26) were searched in all transcripts using the HMMER3[19] algorithm and those transcripts that matched at least one of all six possible frames were discarded. Another criteria commonly accepted to define lincRNA is the evaluation of their coding potential; absence of coding potential is distinctive of putative lincRNA. PhyloCSF[20] is a comparative genomics method (phylogenetic codon substitution frequency) built upon a multiple sequence alignment of 29 mammalian genomes in multi-alignment file format (MAF, http://genome.ucsc.edu/FAQ/FAQformat.html#format5.

(Data Citation 4). The entire set of novel transcripts that passed the previous filters was used as input for PhyloCSF. Transcripts scoring more than 100 decibans (PhyloCSF scores were obtained using option --*frames=6*) were excluded from the final catalog. This threshold was calculated by Cabili et al., as it corresponds to a false-negative rate of

6% for coding genes (i.e., 6% of coding genes are classified as noncoding) and a false-positive rate of ~10% (i.e., 9.5% of noncoding transcripts are classified as coding). They optimized PhyloCSF specificity and sensitivity threshold for the classification of coding and noncoding transcripts on the RefSeq reference sequence database of the National Center for Biotechnology Information (RefSeq coding and RefSeq lincRNAs).

De novo **transcriptome data integration.**

In order to create a comprehensive and unique annotation of novel lincRNAs identified in lymphocytes, duplicates generated by the three approaches used must be resolved. To accomplish this task Cuffcompare was used. For each *de novo* reconstruction approach Cuffcompare merged the transcripts generated by all the populations. The result is a set of three distinct annotations corresponding to TopHat/Cufflinks, Star/Cufflinks, Star/Trinity/PASA. These three lincRNA sets were further merged to generate a non redundant atlas of lincRNAs in human lymphocytes and only those genes identified by at least two out of the three software programs were considered. After data integration through Cuffocmpare, a custom script was used to remove and substitute the XLOCs and TCONs, assigned by the software, with their original and public names.

New lincRNAs were then uniquely identified with a name that contains the prefix ‘linc-’; the Ensembl gene name of the nearest protein-coding gene (irrespective of the strand); the location of the lincRNA relative to the sense of transcription of the nearest protein-coding gene: ‘up’ or ‘down’; the description of the concordance of the transcription between the lincRNA and its nearest coding gene: ‘sense’ or ‘antisense’; a counter to distinguish between lincRNA that share the same nearest protein-coding gene. An example of template name is ‘linc-geneX-(up|down)-(sense|antisense)_#n’. The resulting annotation has been integrated concatenating it to the GRCh37 version 67 provided by Ensembl.

The resulting integrated annotation comprises 563 novel lincRNAs genes and 1,797 novel transcripts, publish in a previous work[7] is avaliable in Data Citation 1.

Data Records

In this study we deposited 1 dataset, which contains the RNA-Seq raw reads in fastq format (Data Citation 1 and Supplementary Table 1 (which is a simplified version of the ISA-TAB please see the associated Metadata Record). This dataset contains 63 samples in total, grouped by 13 lymphocyte subsets with 4 or 5 biological replicates each. Supplementary Table 1 is an XLSX with the following header: Source, the original source name used by the lab; Name, assigned by the provider; SubSet, the lymthocyte subset; Antibody, the antibodies used for sorting; InnerSize, the estimated inner size; R1 URI, the forward reads uri for download; R1 MD5SUM, checksum for the forward reads; R2 URI, the reverse reads uri for download; R2 MD5SUM, checksum for the reverse reads. The annotation of the 563 newly described lincRNA (Data Citation 1: the new 563 annotated lincRNAs[7]) is a General Transfer Format (GTF).

Technical Validation

RNA-seq raw data quality

Assessing the quality of the data performing the Quality Control (QC) is crucial to the whole study. RNA-seq data generated were initially analyzed with FASTQC and a summary plot with the data from all samples is depicted in Fig. 2a. The quality of the reads during the sequencing tends to decrease but it can be further improved using specific software that removes low quality bases reducing the length of the read or directly discard the whole read when its quality is too low. To perform the trimming and filtering Trimmomatic was run on each sample and the data were later on reanalysed with FASTQC to confirm the quality improvements. The summary of the resulting data is shown in

Fig. 2a. Another criteria to measure the QC for NGS reads is the % of GC content, which is improved by the filtering (Fig. 2c)

During the study two mapping software were used, TopHat and STAR. To exclude the possibility of discordance between the two aligners, the mapping results were compared to assess their mapping performance. The alignments with the two software showed a good concordance (99%) with a slight advantage of STAR in terms of mapped reads **(**Fig. 2d**)**.

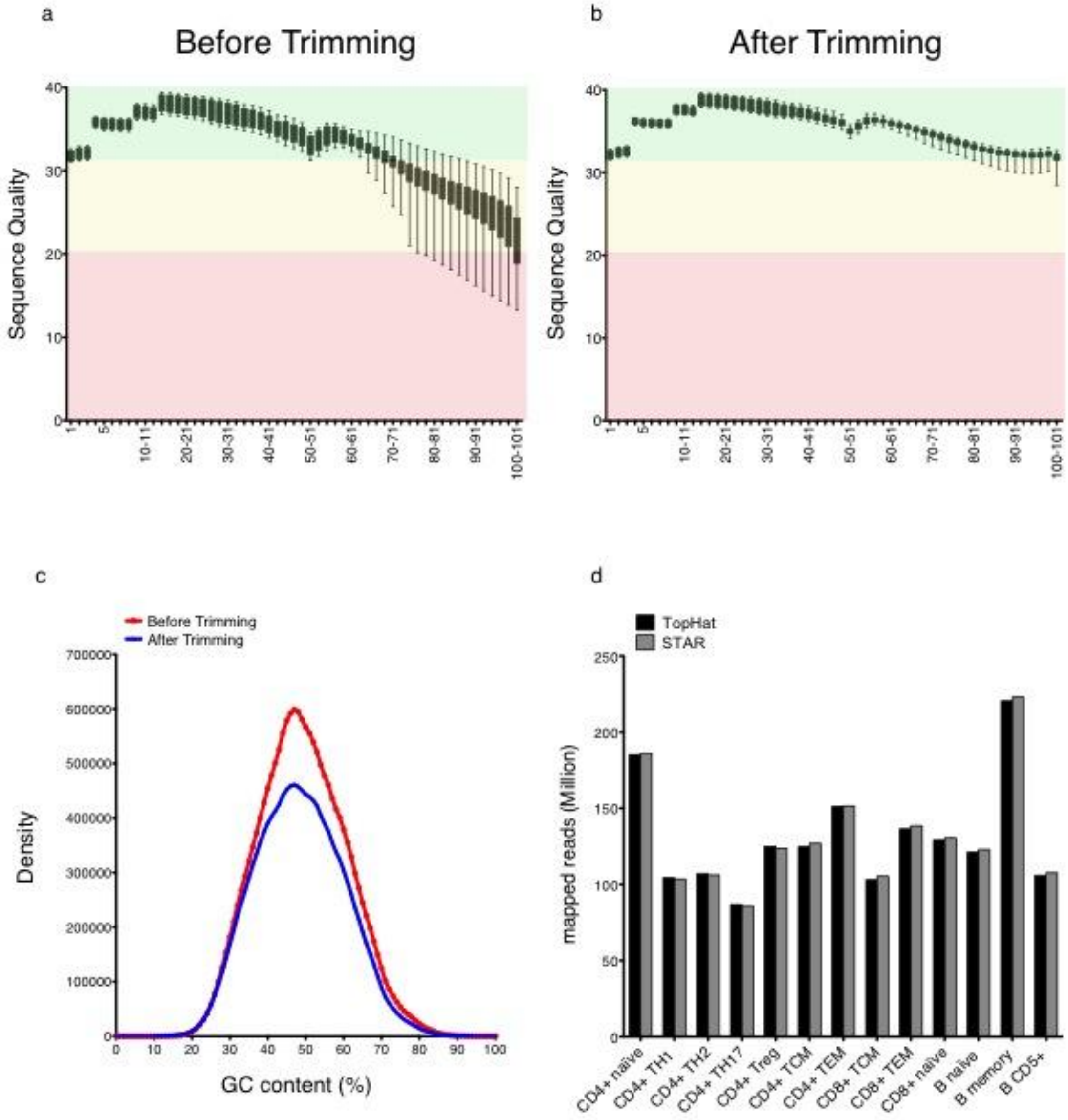


Figure 2 Quality control assessments. (a) Phred quality score of the avarage distribution over all reads across all samples in each base before and (b) after trimming. (c) %GC content before and after trimming. (d) Comparison of the mapped reads for all the human lymphocyte subsets profiled using TopHat and STAR.

RNA-seq biological replicates

Biological replicates are fundamental to guarantee data consistency, in this study the lymphocyte populations profiled have 4 biological replicates for CD5 B and CD8 TCM and 5 biological replicates for all the other populations. In order to establish the congruency among biological replicates Principal Component Analysis (PCA) Fig. 3a and hierarchical clustering Fig. 3b were performed. A good separation between B and T cells samples is achieved by PCA on normalized read counts using DESeq2[21]. Comparable results are obtained using hierarchical clustering on the same data. Moreover, similarity between biological replicates of the same population is obtained, showing a good consistency and correlation among them.

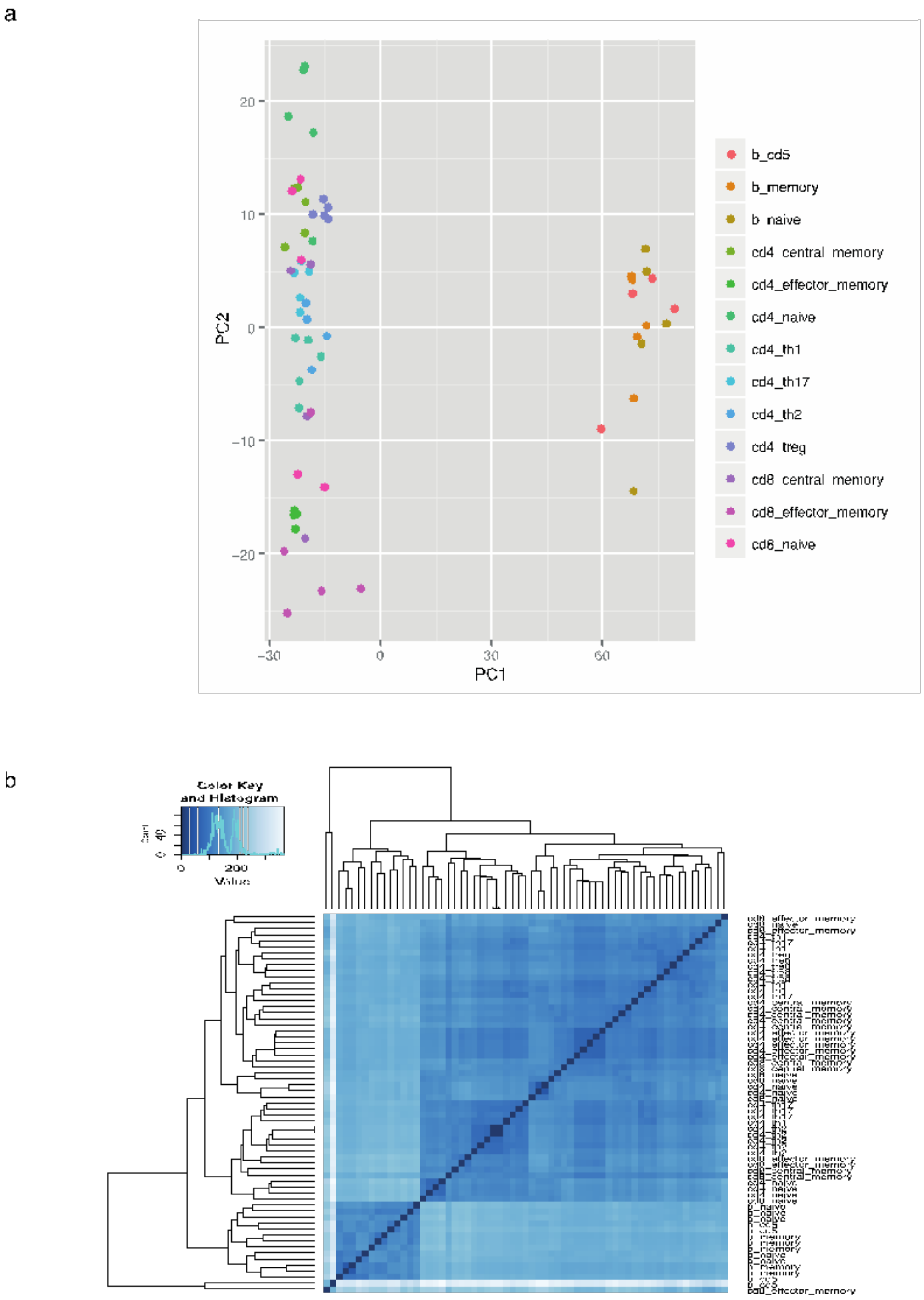

a
PC2
PC1
20
10
0
−10
−20
−30
0
30
60
b_cd5
b_memory
b_naive
cd4_central_memory
cd4_effector_memory
cd4_naive
cd4_th1
cd4_th17
cd4_th2
cd4_treg
cd8_central_memory
cd8_effector_memory
cd8_naive
b
Color Key
and Histogram
Count
0
40
0
200
Value
cd8_effector_memory

Figure 3 Analysis of intra-population consistency: Principal Component Analysis and hierarchical clustering. (a) Principal Component Analysis (PCA) performed using DESeq2 rlog-normalized RNA-seq data. Loadings for principal components 1 (PC1) and PC2 are reported in graph (on x and y-axes). (b) Hierarchical clustering analyses performed using DESeq2 rlog-normalized RNA-seq data. Color code (from white to dark blue) refers to the distance metric used for clustering (dark blue corresponds to the maximum of correlation values).

Low abundance transcript consistency

As reported in literature, lincRNAs are expressed at very low levels (Hart et al., BMC Genomics 2013), so it is necessary to define a FPKM threshold to discriminate low abundant functional transcripts from technical or biological noise. To define a sensible FPKM threshold, Hart et al.[22] integrated RNAseq data and CHIPseq data of 17 human cell lines from ENCODE project. They established a relationship between the gene expression levels and promoters activities and set the FPKM value when the fraction of active promoters is equal to the fraction of repressed promoters as an expression cutoff. We calculated a threshold of 0.21 FPKM as the mean of the data reported in the paper. Then, we considered for the downstream analysis only genes whose expression values were at least 0.21 FPKM in one population.

De novo transcript identification

Multiple combinations of software and filters were used for the identification of lincRNAs in the 13 lymphocytes populations. Moreover, we considered only newly described lincRNAs detected in at least 2 out of 3 de novo approaches to improve the reliability of the data.

LincRNAs discrimination between coding and non-coding RNA depends on the algorithm used to asses the coding potential, in this study was used PhyloCSF. The final dataset of putative lincRNAs was further processed using iSeeRNA[24] in order to verify our results using a different approach based on Support Vector Machines (SVM). The classification we obtained is highly concordant, as ~99% of the putative lincRNAs contained in the final catalogue are classified as 'noncoding' by iSeeRNA.

Usage Notes (optional)

This study was performed on the version 67 from May 2012 of Ensembl GRCh37 but the current version of the Ensembl human genome is GRCh38 version 80. In order to consider the catalogue of newly described lincRNAs generated in this study, researchers must update the annotation associated with the (Data Citation 1 Array Express E-MTAB-2319) using the liftover software (https://genome.ucsc.edu/cgi-bin/hgLiftOver) from UCSC or the assembly converter (http://www.ensembl.org/Homo_sapiens/Tools/AssemblyConverter) from Ensembl.

Software used during this study went through minor and major code base updates. The more notable software suite that has been updated during the time is Trinity and is strongly suggested to use the latest release downloadable from https://github.com/trinityrnaseq/trinityrnaseq.

For the evaluation of the coding potential of de novo transcripts we suggest to use other recently developed software that perform the classification more efficiently than PhyloCSF, such as iSeeRNA, CNCI[23] and CPAT[24]. It has been demonstrated that these algorithms have a higher level of accuracy, and execution times are considerably faster[23–25].

Acknowledgements

M. Moro and M.C. Crosti for technical assistance with cell sorting; B.J. Haas and A. Dobin for help with the integration of genome-guided Trinity with STAR aligner; S. Donizetti for the IT support; the Istituto Nazionale Genetica Molecolare Bioinformatics Facility for support; Supported by Il Consiglio Nazionale delle Ricerche–Il Ministero dell'Istuzione dell'Universita e della Ricerca (Progetto Bandiera EPIGEN), Fondazione Cariplo (grant n. 2013-0955), the Associazione Italiana per la Ricerca sul Cancro (grant n. IG2013-ID14596), the European Research Council (grant n. 269022 to S.A.; grant n. 617978 to M.P.) and Fondazione Romeo ed Enrica Invernizzi.

Author Contributions

R.J.P.B set up the bioinformatics pipelines and wrote the manuscript;
V.R. set up the bioinformatics pipelines and contributed to the preparation of the manuscript;
A.A. set up the bioinformatics pipelines and contributed to the preparation of the manuscript;
G.R. designed and performed the main experiments, analyzed the data and contributed to the preparation of the manuscript;
S.C., I.P. and P.G. prepared the library and performed the experiments;
S.A. and M.P. designed the study, supervised research and wrote the manuscript;
and all authors discussed and interpreted the results.

Competing Financial Interests

The authors declare no competing financial interests.

References

Bibliographic information for any works cited in the above sections.

1. Cabili, M. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25,** 1915–1927 (2011).

2. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Insights Into Functions. *Nat. Rev. Genet.* **10,** 155–159 (2009).

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

4. Kretz, M. *et al.* Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* **493,** 231–5 (2013).

5. Carpenter, S. *et al.* A long noncoding RNA mediates both activation and repression of immune response genes. *Science* **341,** 789–92 (2013).

6. Hu, G. *et al.* Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. *Nat. Immunol.* **14,** 1190–8 (2013).

7. Ranzani, V. *et al.* The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of T cell differentiation by linc-MAF-4. *Nat. Immunol.* (2015). doi:10.1038/ni.3093

8. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

9. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).

10. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

11. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

12. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–5 (2010).

13. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-seq. *Bioinformatics* **27,** 2325–2329 (2011).

14. Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* **8,** 469–477 (2011).

15. Wu, T. D. & Watanabe, C. K. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21,** 1859–1875 (2005).

16. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9,** R7 (2008).

17. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

18. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40,** D290–D301 (2011).

19. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11,** 431 (2010).

20. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27,** 275–282 (2011).

21. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2 Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. (2014).

22. Hart, T., Komori, H. K., LaMere, S., Podshivalova, K. & Salomon, D. R. Finding the active genes in deep RNA-seq gene expression studies. *BMC Genomics* **14,** 778 (2013).

23. Sun, L. *et al.* Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* **41,** (2013).

24. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41,** e74 (2013).

25. Sun, K. *et al.* iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC Genomics* **14 Suppl 2,** S7 (2013).

Data Citations

Bibliographic information for the data records described in the manuscript.

1. Pagani M., Bonnal R. J.P., Array Express E-MTAB-2319 (2015).
2. Schroth G., Array Express E-MTAB-513, (2011)
3. Cabili M., HTTP Download http://www.broadinstitute.org/genome_bio/human_lincrnas/?q=lincRNA_catalog
4. UCSC, FTP Download, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/, (2009)

Chapter 4

Title: Analysis RNA-seq and non coding RNA

Authors

Alberto Arrigoni1, Valeria Ranzani1, Grazisa Rossetti1, Ilaria Panzeri1, Sergio Abrignani1,2, Raoul J.P. Bonnal1, Massimiliano Pagani1,3

Affiliations

1. Istituto Nazionale Genetica Molecolare ‘Romeo ed Enrica Invernizzi’, Via F. Sforza 35, 20122 Milan, Italy
2. Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Via Festa del Perdono 7, 20122 Milano, Italy
3. Department of Medical Biotechnology and Translational Medicine, Università degli Studi di Milano, Via Festa del Perdono 7, 20122 Milano, Italy

Corresponding authors: Raoul J.P. Bonnal (bonnal@ingm.org), Massimiliano Pagani (pagani@ingm.org)

Accepted for publication on 'Methods in molecular biology'

Summary

RNA-Seq is an approach to transcriptome profiling that uses deep-sequencing technologies to detect and accurately quantify RNA molecules originating from a genome at a given moment in time. In recent years, the advent of RNA-Seq has facilitated genome-wide expression profiling, including the identification of novel and rare transcripts like non coding RNAs and novel alternative splicing isoforms.

Here, we describe the analytical steps required for the identification and characterization of non coding RNAs starting from RNA-Seq raw samples, with a particular emphasis on long non coding RNAs (lncRNAs).

Keywords: RNA-seq, lncRNAs, bioinformatics

Introduction

In recent years, advances in transcriptome reconstruction technologies have made possible the identification and the characterization of thousands of novel long non coding RNAs (lncRNAs) from short reads RNA-seq data [2,3,4].

The rapid increase of sequencing depth and read length has considerably improved the accuracy of transcripts reconstruction and offers the unprecedented possibility to characterize lncRNAs on a global scale.

LncRNAs are defined as transcripts of length >200 nucleotides that are characterized by a low coding potential [1]. The choice of this length threshold is somewhat arbitrary, but it is instrumental in order to separate lncRNAs from other non-coding RNA classes, such as microRNAs (miRNAs), short interfering RNAs (siRNAs), Piwi-interacting RNAs (piRNAs), small nucleolar RNAs (snoRNAs), and other short RNAs [1].

LncRNAs are broadly classified according to the genomic context in which they are located (Fig.1): - antisense lncRNAs are transcripts that span at least one exon of a nearby protein coding, and are transcribed in the opposite direction - intronic lncRNAs originate from intronic regions, and they do not overlap any annotated exon - bidirectional lncRNAs are transcripts that initiate in a divergent fashion from the promoter of a protein-coding gene - intergenic lncRNAs are lncRNAs with separate transcriptional units from protein coding genes [5].

In contrast to what has been reported for other non-coding RNA classes [6], long ncRNAs lack strong inter-species conservation [7]. Moreover, the evolutionary linkage between lncRNAs in different species is difficult to infer since the majority of approaches for conservation studies are based on primary sequence analysis. More comprehensive studies that integrate primary sequence analysis, structure and functional role of lncRNAs are therefore needed to assess their conservation on a global scale [7].

This aspect is also relevant in light of the correlation that exists between lncRNAs structure and function, as it has been demonstrated that they physically associate with chromatin modifiers to alter the epigenetic state of target regions, whether *in cis* or *in trans* [5].

LncRNAs are specifically expressed in different tissues, as demonstrated by integrative studies that used RNA-seq to accurately detect and quantify them [2]. This contributed to develop the notion that lncRNAs have been used by evolution as molecular switches whose activity influences the onset and the maintenance of differentiative states of several different tissues and cellular populations [1].

The *de novo* identification of transcripts from RNA-seq data is performed using algorithms that follow two slightly different approaches: *mapping-first* algorithms like Cufflinks [8] and Scripture [9] and *assembly-first* methods like Trinity [10], SOAPdenovo [11], and Oases [12] [**Note 3**].

Then, some peculiar characteristics of non-coding transcripts are leveraged in order to isolate lncRNAs from protein coding transcripts in datasets that are tipically constituted by thousands of previously unidentified transcripts. The analysis of evolutionary patterns across different species (PhyloCSF [13]) and (more recently) classifiers trained on linguistic features (iSeeRNA [15], CPAT [14], PLEK [16]) are used for this purpose.

Moreover, the use of ChIP (Chromatin ImmunoPrecipitation) coupled with NGS (ChIP-seq) provides important and complementary information that fosters the identification process of novel lncRNAs [17]. These methodologies rely on the generation of chromatin maps based on the presence of epigenetics marks (H3K4me3, H3K36me3) and Pol II occupancy to define novel transcriptional units.

Here, we present an analytical protocol (Fig.2) for the characterization of lncRNAs starting from raw RNA-seq samples obtained from sequencing Poly-A$^+$ fractions using paired-end Illumina reads.

Materials

Description of a standard bioinformatics architecture for NGS data analysis and software requirements

- Any UNIX-based operating system (Linux, BSD, Solaris, Mac OSX) could be used to perform the analyses described in the 'Methods' section, although the majority of tools for NGS analyses have been developed with Linux as first choice. The methodologies described in the following sections have been tested on Ubuntu 12.04.5 LTS.
- Since the pipeline involves the use of the STAR mapper, we recommend the use of a workstation with at least 30 GB of RAM.
- Experimental settings: the pipeline described in the 'Methods' section was designed and tested for paired-end, poli-A$^+$ Illumina libraries. Reads length is >=100 pb.

Annotation sources and integration

- Reference annotation .gtf files and .fasta genomic sequence files (hg19/hg38) can be obtained at http://www.ensembl.org/info/data/ftp/index.html (Ensembl FTP data repository) or ftp://hgdownload.cse.ucsc.edu/goldenPath/ (UCSC FTP data repository).

- Integration of available transcript annotations not included in UCSC/Ensembl databases can be merged to the official annotation release by using Cufflinks' utility Cuffcompare (8).

Software versions used for this protocol

- cufflinks/2.2.1, fastqc/0.11.3, samtools/1.2, STAR/2.4.1c, trimmomatic/0.33, cutadapt/1.8, R/3.2.0 (DESeq2), HTSeq/0.6.1, CPAT v1.2

Methods

1) Preprocessing:

1. **Quality assessment** (pre-trimming): [**Note 1**] Raw .fastq files resulting from a paired-end sequencing experiment are analyzed in order to assess overall quality. The initial inspection is carried out using FastQC v0.11.3. This analysis step is important, as it may raise attention regarding library preparation problems that may have occurred (the analysis is carried out both on 'forward' and 'reverse' strand reads). [<fastqc> --outdir . <sample_path>/R1(R2).fastq.gz]
2. **Adapter removal**: Adapters sequences are removed using cutadapt [24]. This is usually necessary when the read length of the sequencing machine is longer than the molecule that is sequenced (for example when sequencing microRNAs).

'Cutadapt' is run both for R1 and R2 indicating adapters' sequences [--anywhere <adapter1> --anywhere <adapter2> --overlap 10 --times 2 --mask-adapter]

3. **Trimming**: Trimmomatic [18] is used with a sliding window approach to remove lower quality bases at the end. Standard parameters used for phred33 encoding: ILLUMINACLIP (LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15), MINLEN parameter is set to 50.
4. **Quality assessment (**post-trimming**)**: The output of Trimmomatic is used as input to FastQC in order to inspect the results of the pre-processing steps (adapter removal and trimming) (results for high quality Illumina data are reported in Panel 1)

2) Alignment to the reference genome:

1. **Reference indexing:** In order to map the short reads to a reference genome, a .fasta reference index should be built for the STAR aligner [19].
2. **Mapping**: [**Note 2**] Paired-end reads are mapped to the reference genome using the STAR aligner [STAR --genomeDir <index_star> --runThreadN <cpu_number> --readFilesIn <trimmed>_R1.fastq.gz <trimmed>_R2_P.fastq.gz --readFilesCommand zcat]. The .sam output of STAR is then converted to its compressed format .bam [samtools view -bS aln.sam > aln.bam], and it is indexed for further analyses (e.g. for use with genome browser and further quality inspection of

mapping) using command 'samtools index' (https://github.com/samtools).

3. **Alignment statistics**: Statistics are calculated on output bam files [samtools flagstat bam_name>]

3) Data exploration:

1. **BAM sorting**: In order to perform data exploration using DESeq2 .bam files are sorted by gene identifier, using command [samtools sort -n].
2. **Read counts**: The overlap of reads with annotation features found in the reference .gtf is calculated using HT-seq [20]. The output computed for each sample (raw read counts) is used as input for Bioconductor's DESeq2.
3. **Principal Component Analyses of RNA-seq samples**: (Panel 2A) raw counts produced with HT-seq are parsed using DESeq2 (v 1.8.1)[21] for each sample and the biological labels (e.g. cellular population identifiers) are provided in a tabular 'samplesheet' along with the counts files. Raw counts are normalized using DESeq2's function 'rlog', which outputs 'variance stabilized' values and transforms the original count data to the log scale. Normalized counts are used to calculate and plot Principal Component Analysis (PCA) (using DESeq2's 'plotPCA' function). (http://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf) (example PCA performed using 12

samples belonging to 2 different biological classes is depicted in Panel.2A)

4. **Heatmap of sample-to-sample distances**: (Panel 2B) normalized counts used to produce the PCA plot of the previous step are also used to calculate sample-to-sample distances with DESeq2's functions 'hclust' and 'heatmap.2'. (example hierarchical cluster analysis performed using 12 samples belonging to 2 different biological classes is reported in Panel 2B).

4) De novo identification of transcripts:

1. **BAM sorting**: In order to perform *de novo* discovery using Cufflinks [8], the .bam files of every sample should be sorted by coordinate using samtools.
2. **Cufflinks *de novo***: [**Note 3**] [**Note 4**] Transcripts identification is performed on a per-sample basis using a tool from the Cufflinks suite ('RABT') [8]. For the software to discover novel transcripts, the reference .gtf file used in the initial mapping step must not be supplied. The resulting file "transcripts.gtf" found in the output folder contains the assembled transcripts identified by Cufflinks [cufflinks -o OutputDirectory/ mappedReads.bam]
3. **Cuffmerge:** [**Note 5**] Predicted transcripts (contained in one .gtf file for each sample) are merged using Cufflinks utility 'cuffmerge'. The .gtf reference file is supplied to ‘cuffmerge’ so that newly discovered genes and transcripts are integrated in the original annotation. The resulting .gtf file is used as input for the

differential expression analysis [cuffmerge <gtfs_list> -g <ref.gtf> -s <ref_fasta> -p <cpu_number>] [**Note 6**] [**Note 13**]

4. **Length filter**: the transcripts identified using the *de novo* approach are filtered by length, as lncRNAs are by definition longer than 200 pb. Sequences' lengths can be evaluated after the results of 'gtf_to_fasta' (included in the Cufflinks suite).

5) De novo identification of lncRNAs: coding potential evaluation

1. The first step of the classification pipeline filters out the transcripts for which a PFAM match [22] is reported by HMMER's utility 'hmmscan'. The input for this analysis is a multi .fasta file containing the translated nucleotide sequences of each transcript (all six possible transcription frames are considered) [**Note 10**]
2. The coding potential of the remaining transcripts is calculated using CPAT [14]. The threshold for the combined classifier's score is 0.364, where transcripts scoring <0.364 are considered putative 'noncoding' (threshold calculated for the human transcriptome by the authors of CPAT) [**Note 11**]

6) Differential expression:

1. **Cuffdiff**: [**Note 7**] Genes/isoforms differential expression is performed using Cuffdiff (included in the Cufflinks suite), having as input the .bam files produced by the quantification step. Input .bam files should be grouped in the command line (1.bam,2.bam 3.bam,4.bam) in order to describe different biological classes/cellular populations (class separator is the white space). Genes/transcripts annotation used for the differential expression

analysis must be the .gtf file produced in the previous step by cuffmerge. [cuffdiff [options]* <transcripts.gtf> <sample1.bam>,<sample2.bam> <sample3.bam>,<sample4.bam>].

2. The selection of differentially expressed genes/isoforms from Cuffdiff results is carried out by extracting P-value statistics and log-fold change ratios for the classes identified in input. Among the several files produced by Cuffdiff, *gene_exp.diff* can be loaded into Excel/R and used for further analyses and inspection. [**Note 8**]
3. **Consistency filter**: Cuffdiff's output is used to assess the consistency of novel genes across different samples of the same class/cellular population [**Note 9**]

7) Downstream analyses and visualization

1. Downstream analyses [**Note 12**] are performed on normalized gene expression values obtained from Cuffdiff output. These can be loaded into external visualization software such as Mev (http://www.tm4.org/mev.html), Excel, or R. Cluster analysis is then performed using K-means algorithm [**Note 14**] in order to identify the lncRNAs that are specifically expressed in a single cellular population/class [2].
2. To further investigate the specificity of expression values for lincRNAs the JS (Jensen Shannon) score is calculated on the vector of values returned by Cuffdiff using an appropriate model distribution [2] [23].

Notes

1) We recommend using compressed .fastq files in order to limit disk occupancy. Compressed files are directly parsed by software used in subsequent steps of the pipeline, i.e. mapping step (STAR).
2) If several samples are to be analyzed consequently using STAR*,* the option "--genomeLoad LoadAndKeep" should be added to the command line. STAR will then load the genome index into shared memory so that it can use it the next time a calculation is run.
3) Cufflinks uses a *mapping-first* approach to attempt transcripts reconstruction by leveraging genome sequence information to reduce the computational complexity of calculations. Other algorithms are based on an *assembly-first* approach, and though being more computationally expensive can reach higher sensitivity (more transcripts are identified). It is often a good idea to combine these two different approaches in order to obtain higher yields. For human data, it is advisable to start from more than 50 Million read pairs to balance specificity and sensitivity of detection (This estimate is based on the Trinity publication [10], where 52.6 million 76bp read pairs were used for the reconstruction).
4) If multiple de novo strategies (e.g. Cufflinks, Scripture [9], Trinity [10]) are used annotation outputs (.gtf files) must be merged into

a single coherent annotation (using 'cuffmerge' utility from the Cufflinks suite).

5) The Cufflinks *de novo* procedure followed by Cuffmerge generates a new custom nomenclature for novel genes and isoforms, which are respectively named with a standard code followed a progressive integer ‘id’ (XLOC_id and TCONS_id). These new codes can be conveniently renamed when releasing a new catalogue not to overlap any existing external annotation [see Note 13 for further reference].
6) *De novo* transcripts reconstruction is strictly biased by the number and the quality of raw reads that are used for the discovery. For this reason, it may be necessary to pull all available information in a single Cufflinks run in order to increase detection sensitivity. Otherwise, when attempting to characterize transcripts isoforms, data should be kept separate.
7) Different normalization options are available for Cuffdiff: 'classic-fpkm', 'geometric' and 'quartile'. The default for this parameter is 'geometric', so that FPKMs and fragment counts are scaled via the median of the geometric means of fragment counts across all libraries, as described by Anders and Huber in [25]. We recommend using the default option to obtain expression values that are comparable to those obtained with DESeq2.
8) Performing differential analyses using Cuffdiff can be time-consuming, and this may impact significantly on the execution average time of the pipeline. For this reason, DESeq2 can be

used as an alternative. It should be noted that the differential analysis performed by DESeq2 is at gene-level, while for isoforms expression characterization (based on exons occupancy evaluation) DEXSeq is available on Bioconductor. Alternative approaches also present in Bioconductor are 'edgeR' (http://bioconductor.org/packages/release/bioc/html/edgeR.html) and 'bayseq' (http://www.bioconductor.org/packages/release/bioc/html/baySeq.html).

9) For newly identified genes/isoforms, it is important to consider expression consistency across different biological replicates. When analyzing data from different individuals, a discrepancy may indicate biological differences that are peculiar to the set of donors chosen for the experiment. A high consistency ensures that the observed results are in fact real and not technical/biological artifacts.

10) It has been demonstrated [3] that the filtering approach based on PFAM and the subsequent CPAT analysis are consistent with each other, though the combined use of these different methodologies increases specificity of the final results [2].

11) PhyloCSF has been widely used for the prediction of novel lncRNAs in many works. It has been recently demonstrated though [14,15,16] that classifiers based on linguistic features only (or the integration of different variables like conservation and ORF length prediction) are more accurate and considerably faster

(even by orders of magnitude). For these reasons, we suggest to rely on CPAT (or IseeRNA) for the classification step of the pipeline.

12) Downstream analyses can alternatively be performed on normalized counts data (rlogs) produced during the previous 'data exploration' step. Love et al. demonstrate in a recent paper [21] that variance-stabilized 'rlogs' produced using DESeq2 can be used to perform robust hierarchical cluster analyses.

13) At the time of writing, lncRNA transcripts nomenclature is still source for debate, as only some general guidelines have been proposed by the Human Genome Nomenclature Committee (HGNC). Thus, we advise the readers to follow the suggestions proposed in a recent review by Mattick & Rinn (26). Antisense lncRNAs are annotated according to the genomic context and take their names after the overlapping genes, while intergenic lncRNAs should be named as LINC-X, where X is a numerical unique identifier.

14) In downstream analyses, K-means clustering is used to identify lncRNAs that are specifically expressed in a single cellular subset or condition. In order to calculate the ideal number of clusters (K) a 'silhouette' measure can be used (available in the R package 'cluster'). Alternatively, a plot of the within groups sum of squares by number of clusters extracted can be used (http://www.statmethods.net/advstats/cluster.html).

Figures

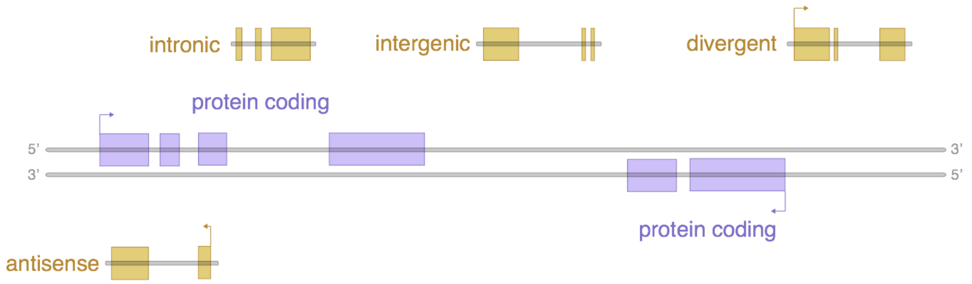


Fig.1 Long non coding RNAs are classified according to the genomic context in which they are located [5]

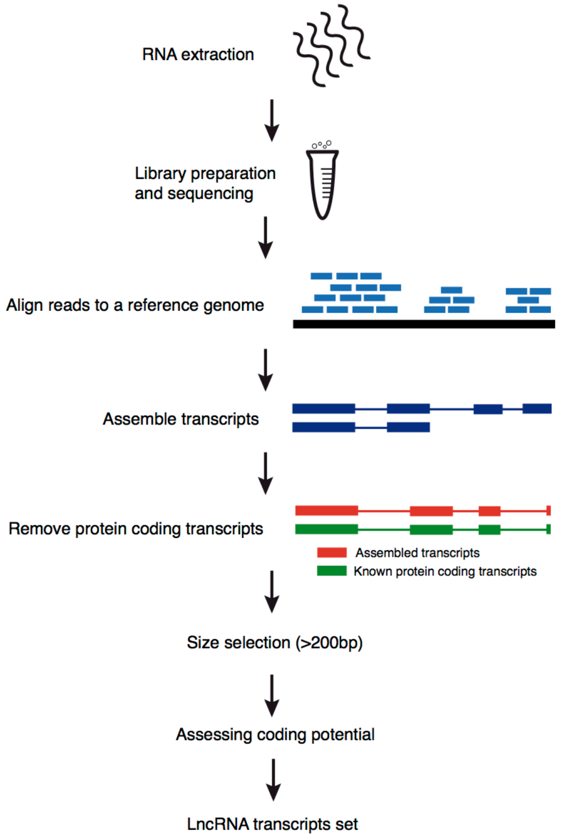
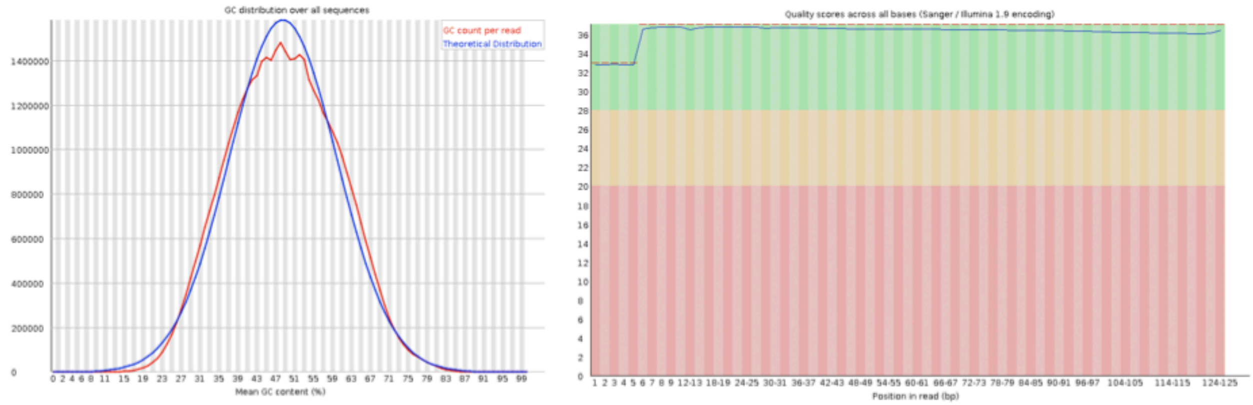
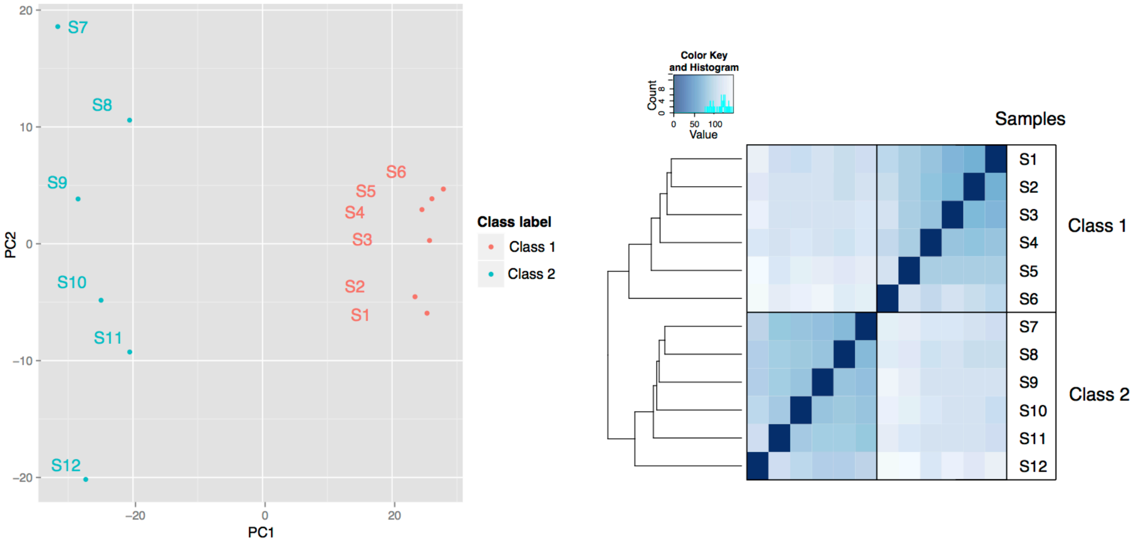

RNA extraction
Library preparation and sequencing
Align reads to a reference genome
Assemble transcripts
Remove protein coding transcripts
🟥 Assembled transcripts
🟩 Known protein coding transcripts
Size selection (>200bp)
Assessing coding potential
LncRNA transcripts set

Fig.2: Overview of the pipeline for the identification of lncRNAs. Schematics of the workflow described in the 'Methods' section.



Panel 1: Quality control after trimming using FastQC output (high-quality data): **A)** Per-sequence GC content is shown as a density graph and compared to a random Gaussian function. **B)** Boxplot graphs representing quality scores over the entire length of reads in the sample. After the trimming step only high-quality reads are expected to be found in output.



Panel 2: Exploratory data analysis of raw RNA-seq samples performed using DESeq2. **A)** Principal Component Analysis (PCA) performed on rlog-normalized expression data reveals a separation between samples belonging to different biological classes ('labels'). **B)** Similar results are obtained by performing hierarchical clustering on rlog-normalized expression data.

References

1. Pagani, M., Rossetti, G., Panzeri, I., Candia, P., Bonnal, R. J., Rossi, R. L., et al. (2013). Role of microRNAs and long□non□coding RNAs in CD4+ T□cell differentiation. *Immunological reviews*, *253*(1), 82-96.
2. Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., et al. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development*, *25*(18), 1915-1927.
3. Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nature genetics*.
4. Ranzani, V., Rossetti, G., Panzeri, I., Arrigoni, A., Bonnal, R. J., Curti, S., et al. (2015). The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of T cell differentiation by linc-MAF-4. *Nature immunology*, *16*(3), 318-325.
5. Rinn, J. L., & Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annual review of biochemistry*, *81*.
6. Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., et al. (2005). Identification of hundreds of conserved and nonconserved human microRNAs. *Nature genetics*, *37*(7), 766-770.

7. Diederichs, S. (2014). The four dimensions of noncoding RNA conservation. *Trends in Genetics*, *30*(4), 121-123.

8. Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, *31*(1), 46-53.

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

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

11. Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., et al. (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, *30*(12), 1660-1666.

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

13. Lin, M. F., Jungreis, I., & Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, *27*(13), i275-i282.

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

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

16. Li, A., Zhang, J., & Zhou, Z. (2014). PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC bioinformatics*, *15*(1), 311.

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

18. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, btu170.

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

20. Anders, S., Pyl, P. T., & Huber, W. (2014). HTSeq–A Python framework to work with high-throughput sequencing data. *Bioinformatics*, btu638.

21. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, *15*(12), 550.
22. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths□Jones, S., et al. (2004). The Pfam protein families database. *Nucleic acids research*, *32*(suppl 1), D138-D141.
23. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, *28*(5), 511-515.
24. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, *17*(1), pp. 10-12.
25. Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biol*, *11*(10), R106.
26. Mattick, J. S., & Rinn, J. L. (2015). Discovery and annotation of long noncoding RNAs. *Nature structural & molecular biology*, *22*(1), 5-7.

Chapter 5

Long non-coding RNAs expression in tumor-infiltrating lymphocytes (unpublished data)[1]

The involvement of T regulatory cells in cancer progression has been clearly delineated in the past few years, as treatments that are targeted to Treg cells' depletion have become routinely used [4,5].

Despite increasing attention has been driven in the last few years towards Til Tregs by the effectiveness of cancer immunotherapy, the mechanisms that contribute to generate Tumor infiltrating Tregs and their role in cancer progression still remain elusive [3].

The whole picture is further complicated by the different roles that Treg cells play in different contexts and tissues. Though the presence of Treg cells in NSCLC (non-small cells lung cancer) has been linked to a poor prognosis, Treg cells infiltration of colorectal cancer mass has been found to reduce tumor's aggressiveness and it is associated with a favourable prognosis [6].

In this context, an in-depth understanding of the functional features of tumor infiltrating Treg cell populations may lead to the comprehension of their role in tumors control and allow the identification of novel

[1] *This chapter represents an update on a follow-up work regarding the involvement of lncRNAs in the modulation of tumor-infiltrating regulatory T lymphocytes*

therapeutic targets for the effective modulation of these cells in cancer patients.

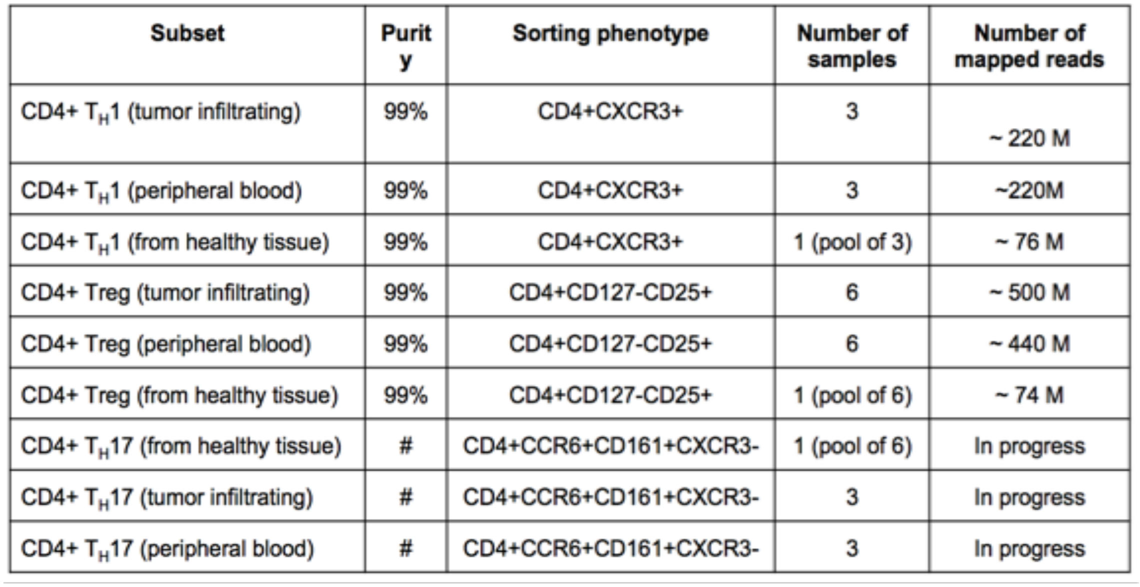
In light of what has been reported in the introductory section of this thesis regarding their role in differentiative and regulatory processes, lncRNAs have indeed the potential of being important actors that preserve and alter T helper cell functions.

In the next paragraph, a brief description of the preliminary results of the characterization by RNA-seq of tumor-infiltrating (NSLC and CRC) Tregs, Th1, Th17 is reported.

Long non-coding RNAs expression in tumor infiltrating lymphocytes

An in-depth characterization and understanding of the molecular mechanisms underlying the functional features of tumor-infiltrating lymphocytes may lead to a comprehension of their role in tumor immune escape and allow the identification of new therapeutic targets for the effective modulation of these cells in cancer. Since very little is known on the expression of lncRNAs in TILs, CD4+ Th1, Th17 and Tregs cells infiltrating both tumor and the adjacent healthy tissue as well as lymphocytes from lymphoid tissues and peripheral blood of Non-Small-Cell-Lung cancer patients were isolated. These cells were analysed by RNA-seq and a set of IncRNAs that are specifically expressed in TIL subsets has been defined.

The number of RNA-seq samples produced along with selection markers used and purity achieved are reported below [these preliminary results refer to work performed on NSCLC, while characterization of CRC is still ongoing]:



Subset	Purity	Sorting phenotype	Number of samples	Number of mapped reads
CD4+ T_H1 (tumor infiltrating)	99%	CD4+CXCR3+	3	~ 220 M
CD4+ T_H1 (peripheral blood)	99%	CD4+CXCR3+	3	~220M
CD4+ T_H1 (from healthy tissue)	99%	CD4+CXCR3+	1 (pool of 3)	~ 76 M
CD4+ Treg (tumor infiltrating)	99%	CD4+CD127-CD25+	6	~ 500 M
CD4+ Treg (peripheral blood)	99%	CD4+CD127-CD25+	6	~ 440 M
CD4+ Treg (from healthy tissue)	99%	CD4+CD127-CD25+	1 (pool of 6)	~ 74 M
CD4+ T_H17 (from healthy tissue)	#	CD4+CCR6+CD161+CXCR3-	1 (pool of 6)	In progress
CD4+ T_H17 (tumor infiltrating)	#	CD4+CCR6+CD161+CXCR3-	3	In progress
CD4+ T_H17 (peripheral blood)	#	CD4+CCR6+CD161+CXCR3-	3	In progress

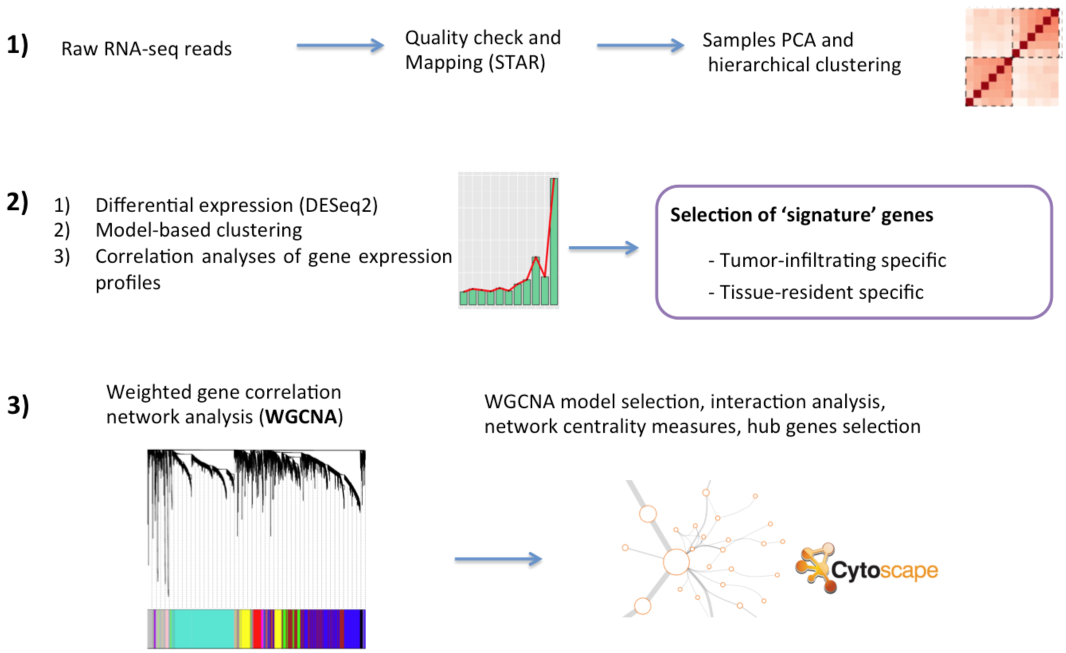
Table 1: RNA-seq samples analyzed of tumor-infiltrating (NSCLC), healthy lung tissue and patients' peripheral blood lymphocytes.

The analyses of RNA-seq data produced in this experiment have been carried out according to the workflow reported in fig.1.

Briefly, quality controls on raw reads and mapping to the reference genome have been followed by exploratory data analysis (samples PCA and hierarchical clustering) to ensure that the examined biological classes (Til, Pb and healthy tissue) are homogeneous. A differential expression analysis has been conducted to identify 'variable' genes using DESeq2 [7]. Then, model-based clustering [8] has been conducted on genes found to be differentially expressed, and clusters have been correlated to specificity for biological classes of interest.

In fig.2 normalized expression values of genes that are specifically expressed in tumor-infiltrating Tregs are represented as violin plots and bar charts.

Bioinformatic workflow for the analysis of
Tumor-infiltrating lymphocytes



1) Raw RNA-seq reads → Quality check and Mapping (STAR) → Samples PCA and hierarchical clustering

2)
1) Differential expression (DESeq2)
2) Model-based clustering
3) Correlation analyses of gene expression profiles

→ **Selection of 'signature' genes**

- Tumor-infiltrating specific
- Tissue-resident specific

3) Weighted gene correlation network analysis (**WGCNA**) → WGCNA model selection, interaction analysis, network centrality measures, hub genes selection

Fig.1: Schematics of the bioinformatics workflow used for the analysis of tumor-infiltrating lymphocytes

In order to identify co-regulated gene networks that may be responsible for important regulatory processes in tumor-infiltrating lymphocytes, Weighted Gene Correlation Network Analysis (WGCNA) [9] will be performed on the same RNA-seq data.

WGCNA will reinforce and complement the results obtained with model-based clustering, and will add a new layer of knowledge-based (Gene Ontology) information to the pipeline.

Taken together, these preliminary results show that it is possible to identify transcripts that are specifically expressed in tumor-infiltrating Tregs (both protein coding genes and lncRNAs), thereby constituting a pool of potential biomarkers that could be used in therapeutic settings to target Til Tregs.

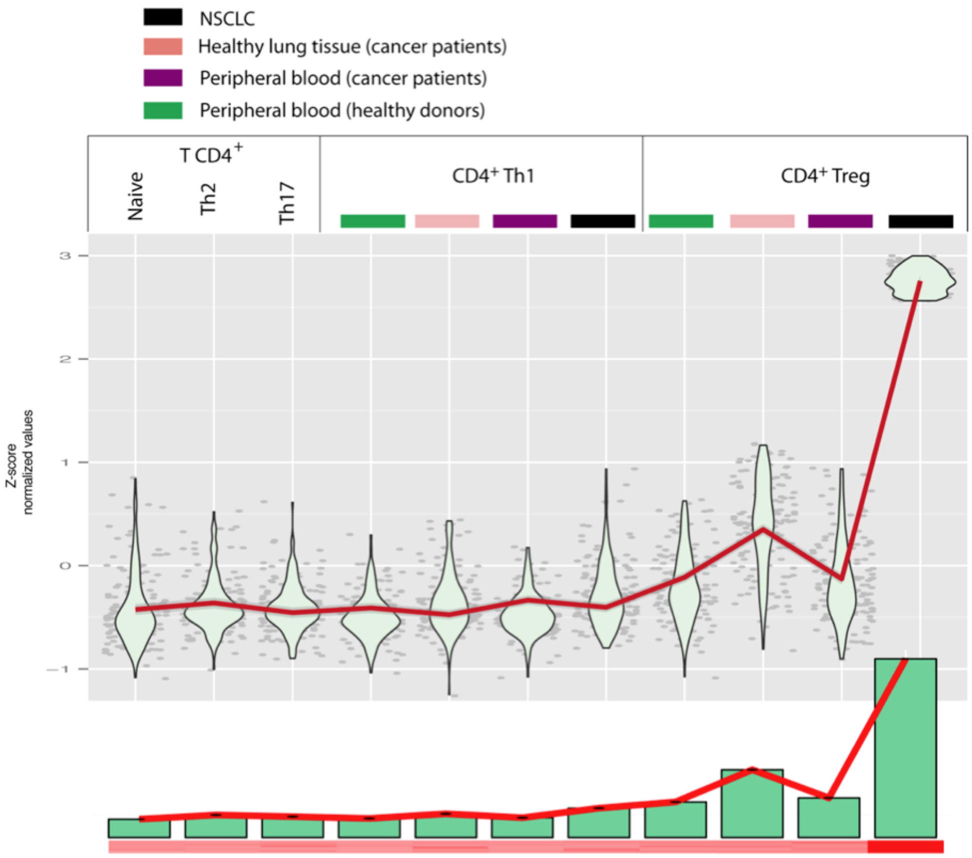


Fig.2: Normalized expression profiles of genes that have high specificity (Pearson correlation >0.85) for tumor-infiltrating lymphocytes (selection is based on model-based clustering results). Starting from the top downward: - violin plot of average expression values and trend line fit - same values showed as barchart - heatmap representation

Single cell characterization of tumor infiltrating Tregs)

Over the past few years, many studies have been conducted that addressed the differentiation of Tregs in the thymus, homeostasis of the Treg cell compartment, and cellular and molecular mechanisms of Treg cell–mediated immunosuppression [10]. In these studies, the Treg population has often been regarded as homogeneous, and the mechanisms and effects studied were averaged over the cells residing in a particular tissutal district (for example the spleen or lymph nodes).

In time, it became apparent that this approach missed out important aspects that underlie Treg cells generation and acquisition of functionalities. In fact, the Foxp3+CD4+ compartment is characterized by considerable heterogeneity [11], which can only be fully appreciated by using transcriptomic analysis methods that narrow their sensitivity range down to a single-cell level.

Data analysis challenges in single-cell transcriptomics

Single-cell RNA-seq data analysis poses unique computational challenges that necessitate the adaptation of existing workflows, as well as the development of entirely new *ad hoc* analytical approaches.

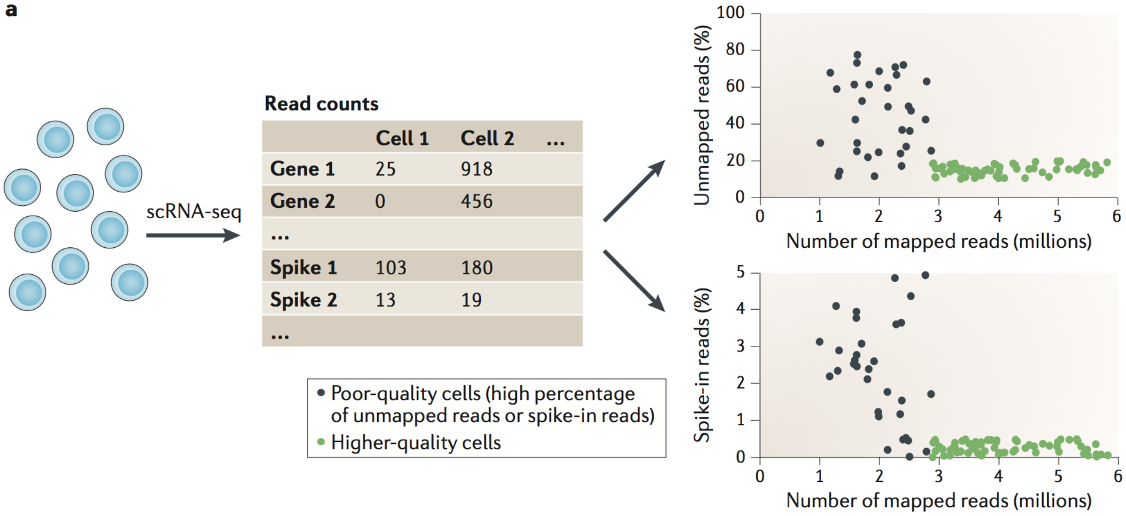
Briefly, the initial step of a single-cell RNA-seq protocol consists in the isolation of individual cells by exploiting microfluidics-based systems before lysing the cell, capturing the polyadenylated fraction of mRNA molecules and obtaining cDNA by reverse transcription. Then, the cDNA is amplified using PCR to obtain enough material to perform RNA-seq profiling [1].

In order to control for amplification bias and technical noise, we need to incorporate quantitative standards that facilitate the comparisons of gene expression levels across cells. To this end, extrinsic spike-in molecules are added to the lysate extracted from each cell (the most widely used spike-in mix is the External RNA Control Consortium [ERCC], a set of 92 synthetic spikes based on bacterial sequences) [12]. As the same volume of spike-in mix is added to each sample, the concentrations of single spike-ins should also be the same, and final quantitative estimates produced with RNA-seq should reflect this. Within the spike-in mix, molecules have different lengths and concentrations, so it is possible to assess the dynamic range of the experiment and derive normalization factors for each sample/cell. A statistical approach based on these assumptions that estimates and controls for statistical noise in single-cell RNA-seq experiments has been proposed by Brennecke et al. [2]. Briefly, spike-ins concentration estimates produced with RNA-seq are fitted to a function (using a GLM of the gamma family) that constitutes a reference for the technical noise that is present in the data. Genes for which expression estimates variance significantly exceed the effects due to technical noise, are selected as ‘differentially expressed’ across studied conditions.

Samples quality control: although software for quality control is routinely used for bulk raw RNA-seq data, single-cell RNA-seq presents specific aspects that need to be taken into account. For instance, it is important to determine whether the RNA in each captured cell is degraded.

This is an extremely important part of the analysis of scRNA-seq data, as many of the cells captured may contain degraded RNA (for example, because the cell is stressed [2]) and should therefore be discarded before downstream analysis.

To this end, an evaluation of the total percentage of mapped reads along with a comparison with the proportion of reads mapping to spike-in molecules can be useful. Moreover, cells with aberrant expression patterns can be spotted by using unsupervised analyses (such as PCA or hierarchical clustering) (fig.3).



	Cell 1	Cell 2	...
Gene 1	25	918	
Gene 2	0	456	
...			
Spike 1	103	180	
Spike 2	13	19	
...			

Fig.3: Basic quality control steps for single-cell RNA-seq data [1]

Normalization: In addition to the aforementioned normalization techniques based on the use of spike-ins, it is important to account for differences in the mRNA content between cells (fig.4)

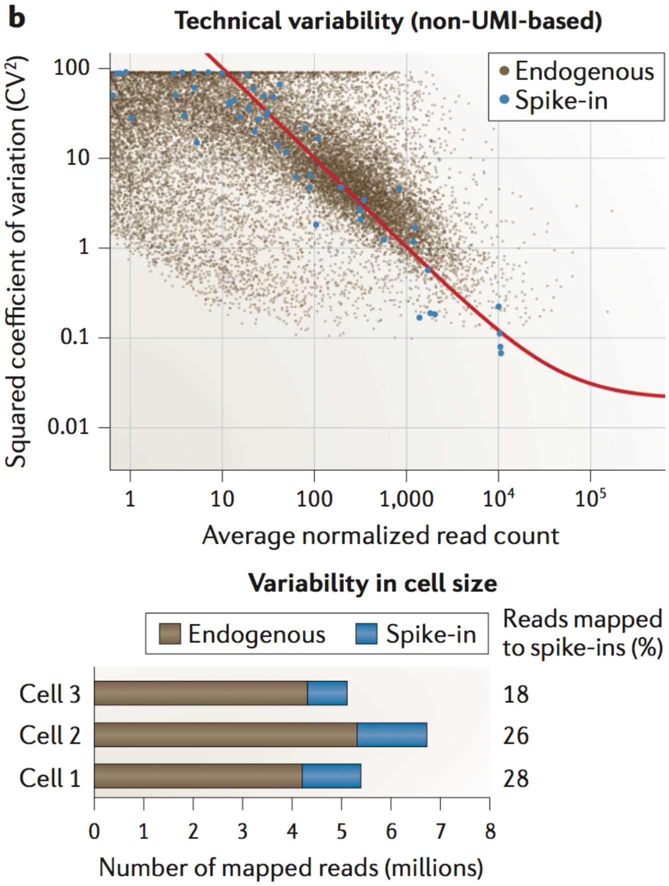


Fig.4: Normalization based on ERCC spike-ins for single-cell RNA-seq data [1]

Downstream analysis: as previously stated, the power of single-cell RNA-seq relies on the possibility to characterize the heterogeneity of cellular populations. This is achieved by the means of unsupervised techniques such as hierarchical clustering, dimensionality-reduction approaches (singular values decomposition [SVD], principal component analysis [PCA]), and more recently neural network-based approaches (self-organizing maps [SOM]).

Nevertheless, single-cell data are inherently noisy and group estimates may be biased if robust data normalization is not performed prior to clustering. Hence, it is important to account for confounding factors (e.g. cell cycle and differentiation state) and to select only highly variable genes for cell type characterization.

References

1) Stegle, Oliver, Sarah A Teichmann, and John C Marioni. "Computational and analytical challenges in single-cell transcriptomics." *Nature Reviews Genetics* 16.3 (2015): 133-145.
2) Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature methods*, *10*(11), 1093-1095.
3) Burzyn, D., Benoist, C., & Mathis, D. (2013). Regulatory T cells in nonlymphoid tissues. *Nature immunology*, *14*(10), 1007-1013.
4) Hodi, F. S., O'Day, S. J., McDermott, D. F., Weber, R. W., Sosman, J. A., Haanen, J. B., et al. (2010). Improved survival with ipilimumab in patients with metastatic melanoma. *New England Journal of Medicine*, *363*(8), 711-723.
5) Simpson, T. R., Li, F., Montalvo-Ortiz, W., Sepulveda, M. A., Bergerhoff, K., Arce, F., et al. (2013). Fc-dependent depletion of tumor-infiltrating regulatory T cells co-defines the efficacy of anti–CTLA-4 therapy against melanoma. *The Journal of experimental medicine*, *210*(9), 1695-1710.
6) Wilke, C. M., Wu, K., Zhao, E., Wang, G., & Zou, W. (2010). Prognostic significance of regulatory T cells in tumor. *International Journal of Cancer*, *127*(4), 748-758.

7) Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, *15*(12), 550.

8) Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, *97*(458), 611-631.

9) Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, *9*(1), 559.

10) Josefowicz, S. Z., Lu, L., & Rudensky, A. Y. (2012). Regulatory T cells: mechanisms of differentiation and function. *Annual review of immunology*, *30*, 531-564.

11) Feuerer, M., Hill, J. A., Mathis, D., & Benoist, C. (2009). Foxp3+ regulatory T cells: differentiation, specification, subphenotypes. *Nature immunology*, *10*(7), 689-695.

12) Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., et al. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome research*, *21*(9), 1543-1551.

Chapter 6

Conclusions

CD4+ T lymphocytes orchestrate immune responses by differentiating into various subsets of effector T cells. We addressed the role of regulatory long non-coding RNAs in T-cell differentiation and plasticity performing a comprehensive transcriptome analysis by RNA-seq of thirteen highly purified human primary lymphocyte subsets and identified more than 500 new lincRNAs. We found that lincRNAs are preferentially expressed in specific lymphocyte subsets and that their expression patterns change during T-cell differentiation. Furthermore, we functionally characterized linc-MAF-4, a Th1 CD4+ signature lincRNA, and found that linc-MAF-4 down-regulation increases the expression levels of transcription factor MAF and skews CD4+ differentiating cells towards a Th2 like expression profile. After assessing the role of lncRNAs in primary T cells from healthy donors, we seek to characterize their involvement in regulatory processes for tumor infiltrating lymphocytes (TIL).

An in-depth characterization and understanding of the molecular mechanisms underlying the functional features of TIL may lead to a comprehension of their role in tumor immune escape and allow the identification of new therapeutic targets for the effective modulation of these cells in cancer. Since very little is known on the expression of lncRNAs in TILs, we isolated CD4+ Th1, Th17 and Tregs cells infiltrating both tumor and healthy tissue as well as lymphocytes from lymphoid Issues and peripheral blood of Non-Small- Cell-Lung and Colorectal cancer patients. We analysed these cells by RNA-seq and defined a set of lncRNAs that are specifically expressed in TIL subsets (Chapter 5).

Translational perspectives

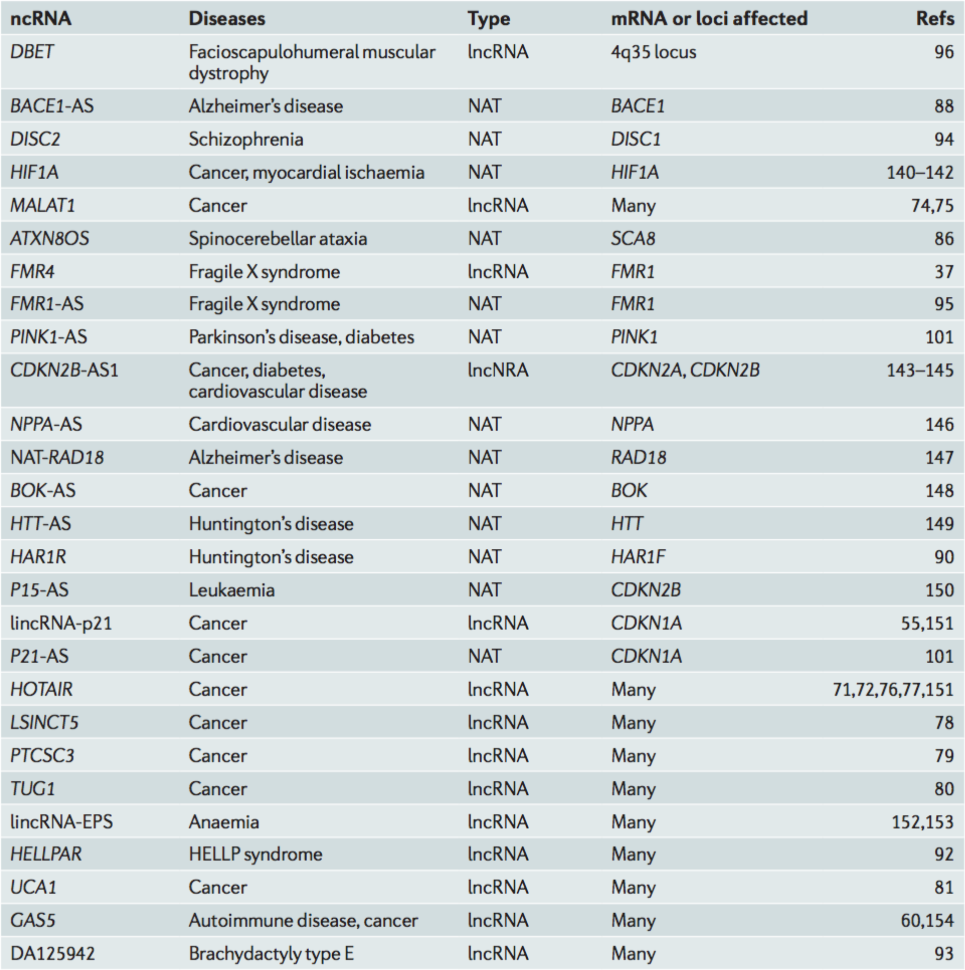
Recent studies on the functions of lncRNAs and their implications in the modulation of cellular plasticity have driven increasing attention to their potential to contribute to the development of novel therapeutical strategies.

Many lncRNAs have already been linked to various disease processes, and cancer features prominently among these (Table 1) [1]. For example, ANRIL (CDKN2B AS RNA 1) is an antisense lncRNA that is overexpressed in prostate cancer tissues, whose loss of expression was associated with a reduction in cellular lifespan and in the upregulation of both INK4A (which is encoded by CDKN2A) and INK4B (which is encoded by CDKN2B) [2]. Polymorphisms in the ANRIL locus also show a statistically significant correlation with acute lymphoblastic leukemia, while it is reported that several SNPs may alter its expression and/or splicing (this too is believed to be linked to alteration of INK4b-ARF-INK4a gene expression) [10]. Thus, antisense therapies could be developed to selectively target the disease-driven ANRIL variants and specifically inhibit the proliferation of malignant cells.

Moreover, high expression levels of HOTAIR have been noted in different metastatic tumors (including prominently breast, liver and the gastrointestinal tract), and have been mainly linked to a poor prognosis [3-5].

Other lncRNAs have been implicated in the biology of cancer, among which - MALAT1 was found to have a role in hepatoblastoma and non-small-cell lung cancer [6], - long stress-induced non-coding transcript 5 (LSINCT5) was related to the onset of breast and ovarian cancer [7], - papillary thyroid carcinoma susceptibility candidate 3 (PTCSC3) was found to be implicated in papillary thyroid carcinoma [8] and TUG1 in bladder urothelial carcinoma [9].

These studies provided solid evidence of the possibility to use lncRNAs as (prognostic) biomarkers. Although a clear demonstration of their functional role in cancer progression is still lacking, it is opinion of many [1] that in the future lncRNAs will be used to target relevant molecular pathways, thereby influencing cells' fate with smart drugs.



ncRNA	Diseases	Type	mRNA or loci affected	Refs
DBET	Facioscapulohumeral muscular dystrophy	lncRNA	4q35 locus	96
BACE1-AS	Alzheimer's disease	NAT	BACE1	88
DISC2	Schizophrenia	NAT	DISC1	94
HIF1A	Cancer, myocardial ischaemia	NAT	HIF1A	140–142
MALAT1	Cancer	lncRNA	Many	74,75
ATXN8OS	Spinocerebellar ataxia	NAT	SCA8	86
FMR4	Fragile X syndrome	lncRNA	FMR1	37
FMR1-AS	Fragile X syndrome	NAT	FMR1	95
PINK1-AS	Parkinson's disease, diabetes	NAT	PINK1	101
CDKN2B-AS1	Cancer, diabetes, cardiovascular disease	lncNRA	CDKN2A, CDKN2B	143–145
NPPA-AS	Cardiovascular disease	NAT	NPPA	146
NAT-RAD18	Alzheimer's disease	NAT	RAD18	147
BOK-AS	Cancer	NAT	BOK	148
HTT-AS	Huntington's disease	NAT	HTT	149
HAR1R	Huntington's disease	NAT	HAR1F	90
P15-AS	Leukaemia	NAT	CDKN2B	150
lincRNA-p21	Cancer	lncRNA	CDKN1A	55,151
P21-AS	Cancer	NAT	CDKN1A	101
HOTAIR	Cancer	lncRNA	Many	71,72,76,77,151
LSINCT5	Cancer	lncRNA	Many	78
PTCSC3	Cancer	lncRNA	Many	79
TUG1	Cancer	lncRNA	Many	80
lincRNA-EPS	Anaemia	lncRNA	Many	152,153
HELLPAR	HELLP syndrome	lncRNA	Many	92
UCA1	Cancer	lncRNA	Many	81
GAS5	Autoimmune disease, cancer	lncRNA	Many	60,154
DA125942	Brachydactyly type E	lncRNA	Many	93

Table.1: Selected characterized lncRNAs with potential roles in human diseases [1]

Prospective strategies for targeting lncRNAs

As discussed above, lncRNAs seem to have critical roles in cancer and modulating their functions may promote anticancer effects. For this reason, different technologies have been proposed [11] to alter the expression levels of lncRNAs that have the potential to set the basis for lncRNA-based cancer therapies.

- RNA interference (RNAi) based techniques are arguably the most popular methods to inhibit lncRNAs in cancer cells, as both siRNAs and shRNAs exhibit great RNA selectivity and knockdown efficiency [11]. SiRNAs target RNA molecules via complementary interaction to the nucleic acid sequence, and after the integration into the active RNA-induced silencing complex (RISC), siRNAs direct post-transcriptional silencing of RNA targets. siRNAs are fully complementary to their RNA targets whom is then cleaved and subsequently degraded.

Recent studies have demonstrated the viability of this approach, as depletion of HOTAIR by siRNAs decreased matrix invasiveness of breast cancer cells [12] and inhibited tumor growth of pancreatic cancer xenograft [13].

- Ribozymes are naturally occurring RNA molecules having intrinsic catalytic activity that can be exploited to selectively cause the degradation of target RNA molecules. Among different known types of ribozymes, hammerhead ribozyme (HamRz) has caught major interest as it shows good target inhibitory effect while having the smallest RNA endoribonucleolytic motif [14]. Anticancer activity of ribozymes has been demonstrated by Pavco et al. [15], and it is current opinion that they may compensate for the limitations of siRNAs design due to differences in target recognition [16].

- the AntagoNAT design: NATs belong to a large class of lncRNAs that have transcripts complementary to other RNAs (their relevance is demonstrated by the fact that about one fifth of the known human genes is overlapped).

Trascriptional de-repression of a target gene can performed by specifically inhibiting the function of NATs using single stranded oligonucleotides designed to strand-specifically block the interaction of the antisense transcript with the sense gene mRNA and/or degrade the antisense transcript.

This approach was originally introduced in 2005 [18], and oligonucleotides that are designed to inhibit NAT function in this manner have been named ‘antagoNATs’ [17].

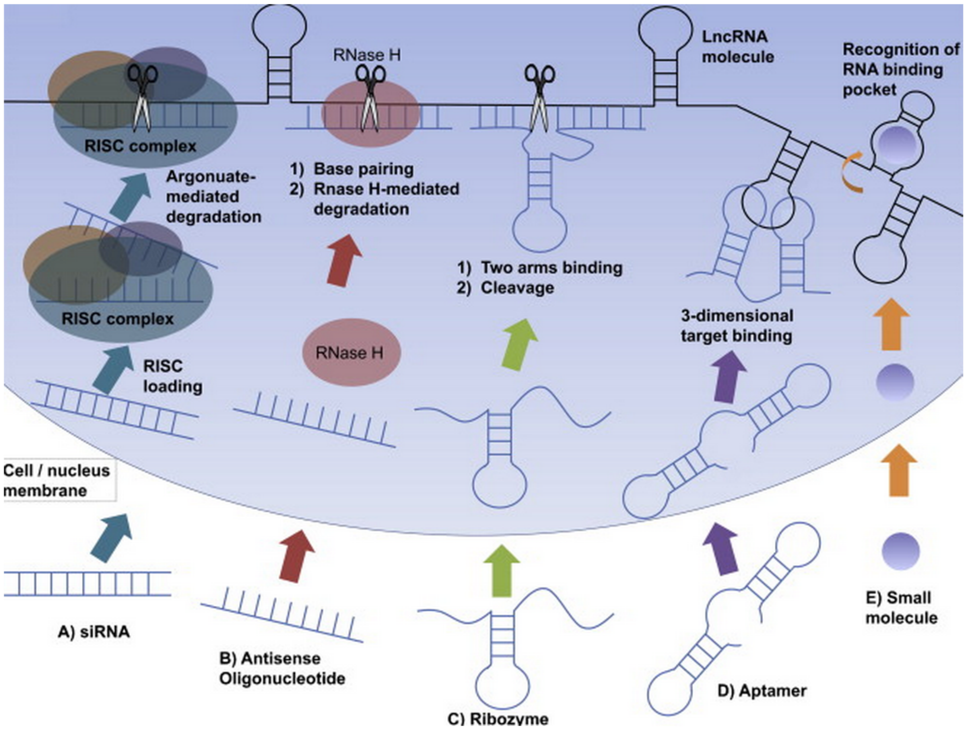


Fig.2: Mechanisms of lncRNA-targeting agents [11]

Selective targeting of lncRNAs through AsiCs technology

According to what has been elucidated in Chapter 5 in the context of tumor-infiltrating lymphocytes, the downregulation of lncRNAs essential for the maintenance of Treg functions might be an efficient way to target Treg cells at tumor sites, and therefore to unleash local tumor specific effector T cells.

LncRNAs modulation can be achieved through the use of small interfering RNAs (siRNA), whose therapeutic potential has been tested in many disease models. In recent years, technologies that mediate targeted delivery of siRNAs have evolved to produce solutions that have high therapeutic efficacy and safety (for example aptamer-siRNA-chimeras [AsiCs] [16]). Aptamers are short single-stranded structured oligonucleotides that can bind a wide range of targets with high affinity and specificity. One caveat of this approach is the fact that a double specificity needs to be defined (a target lncRNA and the receptor that mediates the aptamer intake). Thus, an important step of this approach would be the identification of surface molecules specific for tumour infiltrating Treg cells.

The AsiCs approach has been shown to be effective in different experimental settings directed to: CD4 [13], prostate-specific membrane antigen [14] and HIV-gp120 [15]. In the works documenting these cases, AsiCs are efficiently transfected in cells bearing the recognized surface receptor and they are then able to knock-down gene expression.

These evidences demonstrate that Asics technology could be used to target lncRNAs that are specifically expressed in tumor-infiltrating Tregs, and that this may contribute to the modulation of the immune response inside the tumoral microenvironment.

References

1) Wahlestedt, Claes. "Targeting long non-coding RNA to therapeutically upregulate gene expression." *Nature reviews Drug discovery* 12.6 (2013): 433-446.

2) Yap, Kyoko L et al. "Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a." *Molecular cell* 38.5 (2010): 662-674.

3) Gupta, Rajnish A et al. "Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis." *Nature* 464.7291 (2010): 1071-1076.

4) Yang, Zhe et al. "Overexpression of long non-coding RNA HOTAIR predicts tumor recurrence in hepatocellular carcinoma patients following liver transplantation." *Annals of surgical oncology* 18.5 (2011): 1243-1250.

5) Kogo, Ryunosuke et al. "Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers." *Cancer research* 71.20 (2011): 6320-6326.

6) Schmidt, Lars Henning et al. "The long noncoding MALAT-1 RNA indicates a poor prognosis in non-small cell lung cancer and induces migration and tumor growth." *Journal of Thoracic Oncology* 6.12 (2011): 1984-1992.

7) Silva, Jessica M et al. "LSINCT5 is over expressed in breast and ovarian cancer and affects cellular proliferation." *RNA biology* 8.3 (2011): 496-505.

8) Jendrzejewski, Jaroslaw et al. "The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type." *Proceedings of the National Academy of Sciences* 109.22 (2012): 8646-8651.

9) Han, Yonghua et al. "Long intergenic non□coding RNA TUG1 is overexpressed in urothelial carcinoma of the bladder." *Journal of surgical oncology* 107.5 (2013): 555-559.

10) Iacobucci, Ilaria et al. "A polymorphism in the chromosome 9p21 ANRIL locus is associated to Philadelphia positive acute lymphoblastic leukemia." *Leukemia research* 35.8 (2011): 1052-1059.

11) Li, Chi Han, and Yangchao Chen. "Targeting long non-coding RNAs in cancers: progress and prospects." *The international journal of biochemistry & cell biology* 45.8 (2013): 1895-1910.

12) Gupta, Rajnish A et al. "Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis." *Nature* 464.7291 (2010): 1071-1076.

13) Kim, Kyounghyun et al. "HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer." *Oncogene* 32.13 (2013): 1616-1625.

14) Ruffner, Duane E, Gary D Stormo, and Olke C Uhlenbeck. "Sequence requirements of the hammerhead RNA self-cleavage reaction." *Biochemistry* 29.47 (1990): 10695-10702.

15) Pavco, Pamela A et al. "Antitumor and antimetastatic activity of ribozymes targeting the messenger RNA of vascular endothelial growth factor receptors." *Clinical Cancer Research* 6.5 (2000): 2094-2103.

16) Tedeschi, Lorena et al. "Hammerhead ribozymes in therapeutic target discovery and validation." *Drug discovery today* 14.15 (2009): 776-783.

17) Modarresi, Farzaneh et al. "Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation." *Nature biotechnology* 30.5 (2012): 453-459.

18) Kota, Janaiah et al. "Therapeutic microRNA delivery suppresses tumorigenesis in a murine liver cancer model." *Cell* 137.6 (2009): 1005-1017.

19) Wheeler, L. A., Trifonova, R., Vrbanac, V., Basar, E., McKernan, S., Xu, Z., et al. (2011). Inhibition of HIV transmission in human cervicovaginal explants and humanized mice using CD4 aptamer-siRNA chimeras. *The Journal of clinical investigation*, *121*(6), 2401.

20) Dassie, J. P., Liu, X., Thomas, G. S., Whitaker, R. M., Thiel, K. W., Stockdale, K. R., et al. (2009). Systemic administration of optimized aptamer-siRNA chimeras promotes regression of PSMA-expressing tumors. *Nature biotechnology*, *27*(9), 839-846.

21) Zhou, J., Li, H., Li, S., Zaia, J., & Rossi, J. J. (2008). Novel dual inhibitory function aptamer–siRNA delivery system for HIV-1 therapy. *Molecular Therapy*, *16*(8), 1481-1489.

22) McNamara, J. O., Andrechek, E. R., Wang, Y., Viles, K. D., Rempel, R. E., Gilboa, E., et al. (2006). Cell type–specific delivery of siRNAs with aptamer-siRNA chimeras. *Nature biotechnology*, *24*(8), 1005-1015.