The University of Milano Bicocca

THE PhD PROGRAM DIMET

The PhD Program in Translational and Molecular Medicine (DIMET) is an inter-departmental project between the School of Medicine and the School of Science, organized by the University of Milano-Bicocca.

DIMET - Dr. Ilaria PANZERI - A.A. 2013-14

PhD

PROGRAM IN TRANSLATIONAL AND MOLECULAR MEDICINE

DIMET

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

Deciphering the role of regulatory noncoding RNAs in human CD4+ T lymphocytes differentiation through functional and biochemical studies

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

Dr. Ilaria Panzeri

Matr. No. 067721

XXVII CYCLE
ACADEMIC YEAR
2013-2014

Table of contents

Introduction

The revolution of regulatory noncoding RNAs

At the beginning of our century, the results of the human genome project highlighted the complexity of our genome. What emerged was that the fraction of the genome that is informative is higher than we expected. Subsequent analysis revealed that the vast majority of informative sequences does not encode for proteins. Indeed 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]. From an evolutionary point of view, the genome size is in close relationship with coding potential in prokaryotes, which have haploid genomes primarily composed by protein-coding sequences (~88%). Conversely in eukaryotes a correlation lacks between protein-coding gene number and organismal complexity. These observations are likely explained by the evolution of a more sophisticated architecture to control gene expression that includes the expansion of non-coding regulatory RNAs (ncRNAs)[3]. Thus we should clearly reassess the centrality of protein-coding RNAs in favor of non-coding ones.

Non-coding RNAs with fundamental functions within cells are known since the discovery of the first transfer RNA (tRNA)[4] and comprise also ribosomal RNAs (rRNAs). Nonetheless the interest toward non-coding RNAs with regulatory functions arose with the discovery of the first human micro-RNA, let-7[5]. In order to apply a

theoretical framework to the transcriptome, regulatory ncRNAs are usually classified based on their dimension: “small” ncRNAs being less than 200 nucleotides in length and “long” or “large” ncRNAs (lncRNAs) ranging from more than 200 to tens of thousands nucleotides (table 1).

ncRNA*	No. of known transcripts[‡]	Transcript lengths (nucleotides; nt)[‖]	Functions
Precursors to short RNAs			
miRNA	1,756	>1,000	Precursors to short (21–23 nt) regulatory RNAs
snoRNA	1,521	>100	Precursors to short (60–300 nt) RNAs that help to chemically modify other RNAs
snRNA	1,944	1,000	Precursors to short (150 nt) RNAs that assist in RNA splicing
piRNA	89	Unknown	Precursors to short (25–33 nt) RNAs that repress retrotransposition of repeat elements
tRNA	497	>100	Precursors to short (73–93 nt) transfer RNAs
Long ncRNAs			
Antisense ncRNA	5,446	100–>1,000	Mostly unknown, but some are involved in gene regulation through RNA interference
Enhancer ncRNA (eRNA)[§]	>2,000	>1,000	Unknown
Enhancer ncRNA (meRNA)[‖]	Not fully documented	As variable as the length of mRNAs	Unknown, but they resemble alternative gene transcripts
Intergenic ncRNA	6,742	10^2–10^5	Mostly unknown, but some are involved in gene regulation
Pseudogene ncRNA	680	10^2–10^4	Mostly unknown, but some are involved in regulation of miRNA
3′ UTR ncRNA	12	>100	Unknown

*miRNA, microRNA; snoRNA, small nucleolar RNA; snRNA, small nuclear RNA; piRNA, piwi-interacting RNA; tRNA, transfer RNA; antisense ncRNA, transcripts mapping and overlapping coding and non-coding RNAs; enhancer ncRNA (eRNAs and meRNAs), transcripts that initiate within regions that regulate specific genes; intergenic ncRNA, transcripts that map to genome regions between annotated genes; pseudogene ncRNA, transcripts that come from processed or unprocessed pseudogenes; 3′ UTR ncRNA, 3′-untranslated regions of ncRNA.

Table 1 - Main classes and functions of mammalian ncRNAs[6].

Further complicating the picture, lncRNAs seem to be the preferred substrate for the generation of small RNAs[7]. Both classes can be further classified according to their position relative to known sequences of the genome, like in the case of promoter-associated (PASRs) or transcription initiation small RNAs (tiRNAs) and so on. In particular, long non-coding RNAs are usually classified relative to neighboring protein-coding genes. They can be defined as “sense” if they are transcribed from the same strand of the protein-coding gene or “antisense” if the opposite is true. They can be “divergent” if their promoter and the one of the coding transcript are in close proximity and located in a head to head fashion. They can be “exonic” or “intronic” if they overlap one or more exons, or an intron of the

protein-coding gene respectively. Instead they can be “intergenic” if they lie within a sequence between two protein-coding genes (figure 1).

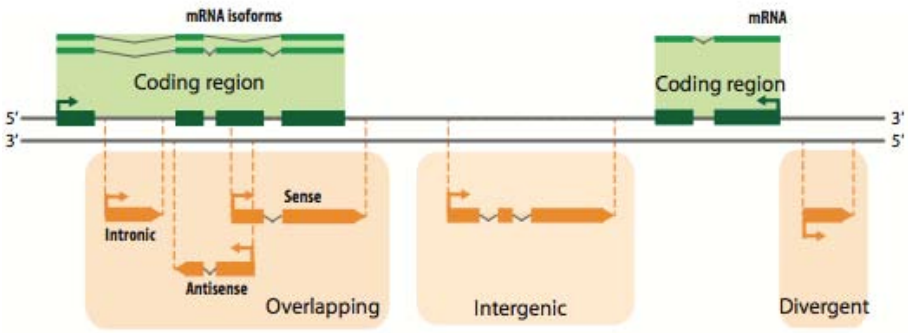


Figure 1 - LncRNAs (orange) classification respect to neighbouring coding regions (green)[8].

This last category can be better defined as long “intervening” non-coding RNAs (lincRNAs), giving that the DNA that codify for these ncRNAs is *per se* a gene. In this thesis we will focus on this last category, which is probably the most studied given that the location of these lncRNAs avoids complications deriving from the overlap with other genes. The majority of known lincRNAs is generated by the same transcriptional machinery of mRNAs. This means that transcribed lincRNAs genomic sequences are marked by RNA polymerase II occupancy and histone modifications, such as H3K4me3 at promoters and H3K36me3 within gene bodies, that are shared with active protein-coding genes[9]. They are capped by methylguanosine at their 5’, spliced and polyadenylated, even if the widespread representation of this last property among known lincRNAs could be partially due to the RNA sequencing strategies used for their identification[10, 11]. Indeed, broader analysis identified

about 39% of lncRNAs to have at least one of the six most common poly(A) motifs, compared to 51% for coding transcripts[1]. These properties imply that there are few distinctive biochemical features that allow the distinction of lincRNAs from protein-coding mRNAs. Among them, lncRNAs have unusual exon structure, with mostly 2-5 exons. Intriguingly, lincRNAs are significantly more likely to overlap repetitive elements and particularly RNA-derived transposable elements (TEs). These last account for about 30% of human lncRNAs nucleotides, often in proximity of their transcriptional start site (TSS), which could suggest that TEs could be important drivers of lncRNAs evolution. Nonetheless, the main difference between lincRNAs and protein-coding genes relies by definition in their coding potential: lncRNAs does not possess open reading frames (ORFs), as evaluated based on: the conservation of ORFs codons[12], ORFs length, the presence of known protein domains, *in vitro* translation[13, 14] and ribosome footprinting assays[15, 16]. However these conceptual constraints are terribly artificial: short, noncanonical peptides have been found to arise from small ORFs within ncRNA[17, 18, 19, 20]; lincRNAs genes can also codify for proteins and have a double function[21] and ultimately, the coding potential does not necessarily exclude a function as RNA also for known mRNAs[22]. Evolution makes boundaries between coding and non-coding genes fainter, as ncRNAs can evolve by pseudogenization. This event can follow the disruption of the ancestral ORF, but not of the untranslated regulatory regions (UTRs) in protein-coding genes duplicates[23] or can arise without duplication, but from co-option of ancestral genes to different, non-coding functions[24]. The boundary between coding and non-

coding is even less defined when ncRNAs arise from joining of coding and non-coding exons through alternative splicing[25, 26], from untranslated regions of mRNAs[27, 28], or from the opposite strand of the overlapping protein-coding gene[29]. Strikingly, more than a half of protein coding genes in mammals have a complementary noncoding transcript[30]. These findings further challenge our “linear” model of the genome, prompting a re-evaluation of current dogma and genes definitions. Genomic regions indeed are far more complex than previously thought: genes can be used for different purposes and different functional elements can co-locate intermingling coding and non-coding regions.

The interest toward lincRNAs has been rapidly growing and their expressions have been quantitated in many different tissues and cell types by high-throughput sequencing (RNA-seq). These efforts retrieved catalogues with little overlap, so that the number of known lincRNAs is still growing, in contrast with the number of known protein-coding genes that has been remarkably stable over years. Indeed, lincRNAs are far more cell-specific than mRNAs, generally less but also more dynamically expressed at various differentiation stages. As mentioned before, such tissue-specificity has been linked to the enrichment of transposable elements in proximity to lincRNAs TSS[31, 32].

These unique properties hint to lincRNAs involvement as fine tuners in cell fate determination, maintenance of cell identity[33], pluripotency[34], commitment and differentiation[35, 36], as demonstrated in many examples. Also lincRNAs are functionally involved in cell

growth[37], apoptosis[38], development[39], imprinting[40] and dosage compensation[41] in almost every cellular context (figure 2).

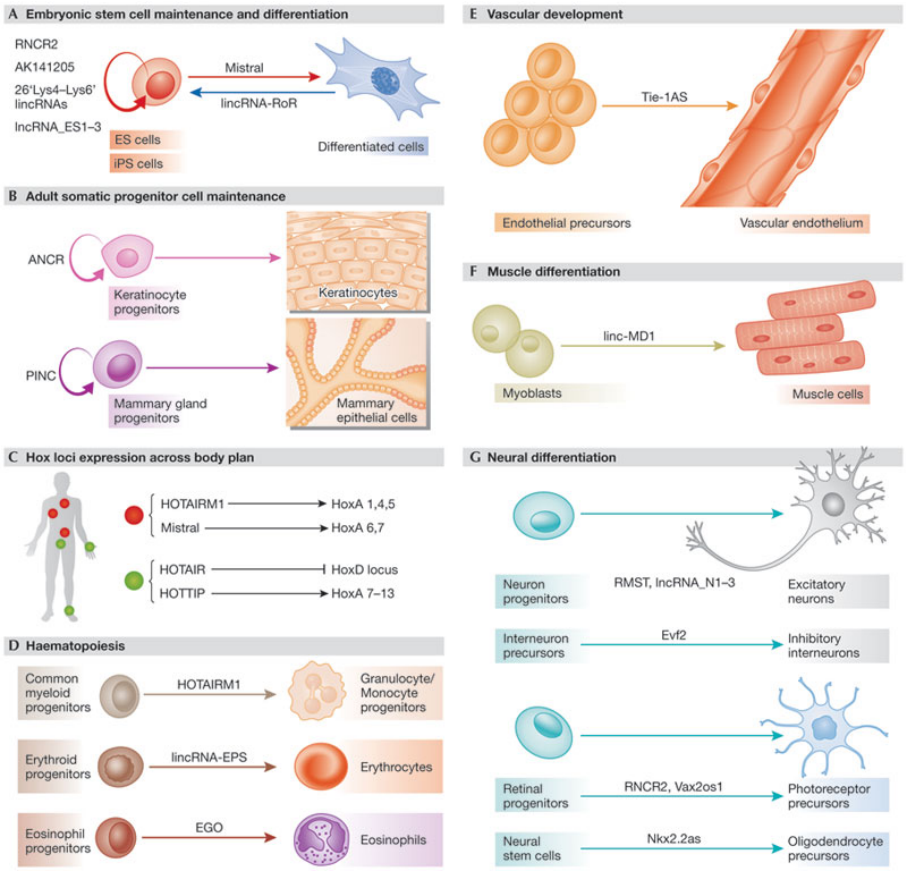


Figure 2 - Regulation of mammalian cell differentiation by lncRNAs[42].

LincRNAs act in these fundamental processes interacting with chromatin or DNA modifiers and transcription factors modulating gene expression; competing with microRNAs acting as sponges; modulating subcellular trafficking, translation, splicing, post-transcriptional modifications and likely through many other mechanisms still to be discovered (figure 3 and table 2, 3).

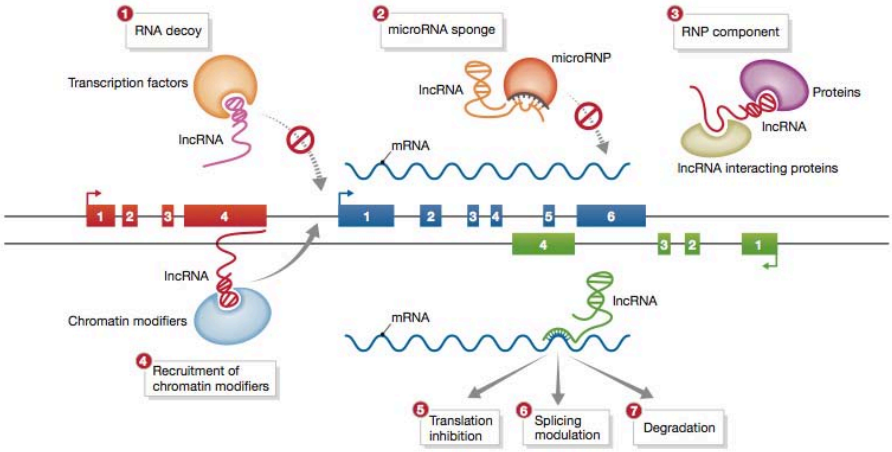


Figure 3 - Mechanisms of lncRNAs function[42].

lncRNA	Function	Mechanism
Regulation of protein activity		
GAS5	Repression of glucocorticoid receptor-mediated transcription	DNA mimicry
EVF2	Transcriptional activation of DLX2 targets	Activation of DLX2
CCND1 promoter RNA	Repression of *CCND1* transcription	Allosteric activation of TLS
NRON	Repression of NFAT-mediated transcription	Inhibition of transcription factor nucleocytoplasmic shuttling
15q11-q13 sno-lncRNA	Regulation of alternative splicing	Inhibition of FOX2 function
rncs-1	Inhibition of Dicer-mediated repression	Sequestration of Dicer or accessory double-stranded RNA-binding proteins
sfRNA	Stabilization of viral and host mRNAs	Inhibition of XRN1-mediated mRNA degradation
gadd7	Inhibition of TDP43-mediated regulatory events	Sequestration of TDP43
Organization of protein complexes		
HOTAIR	Repression at the *HOXD* locus	Recruitment of PRC2 and LSD1
KCNQ1OT1	Imprinting at the *KCNQ1* cluster	Recruitment of PRC2 and G9A
ANRIL	Repression at the *INK4b–ARF–INK4a* locus	Recruitment of PRC1 and PRC2
TERC	Addition of telomeric repeats to the ends of chromosomes	Organizational scaffold for telomerase components and template for repeat addition
SRP RNA	Directing of proteins to the ER	Organizational scaffold for SRP components
NEAT1	Assembly of paraspeckles	Nucleation of subnuclear domains

Table 2 - LncRNAs-mediated regulation of proteins[2]

lncRNA	Function	Mechanism
Regulation of mRNA transcription		
XIST	X inactivation	Chromatin-mediated repression
HOTAIR	Repression at the *HOXD* locus	Chromatin-mediated repression
HOTTIP	Activation at the *HOXA* locus	Chromatin-mediated activation
KCNQ1OT1	Imprinting at the *KCNQ1* cluster	Chromatin-mediated repression
ANRIL	Repression at the *INK4b–ARF–INK4a* locus	Chromatin-mediated repression
AIRN	Imprinting at the *IGF2R* cluster	Chromatin-mediated repression, transcription interference
IME4 antisense	Repression of *IME4* mRNA	Transcription interference
IRT1	Repression of *IME1* mRNA	Chromatin-mediated repression
GAL10 lncRNA	Repression of *GAL1* and *GAL10* mRNAs	Chromatin-mediated repression
PHO84 antisense	Repression of *PHO84* mRNA	Chromatin-mediated repression
ICR1	Repression of *FLO11* mRNA	Modulation of transcription factor recruitment
PWR1	Activation of *FLO11* mRNA	Modulation of transcription factor recruitment
SRG1	Repression of *SER3* mRNA	Nucleosome remodelling
fbp1 ncRNA	Activation of *fbp1*	Chromatin remodelling
LINOCR	Activation of lysozyme mRNA	Nucleosome remodelling
Alu repeat-containing RNA	Transcriptional repression during heat shock	Inhibition of Pol II
HSR1	Activation of the HSF1 transcription factor	Allosteric activation together with eEF1A
Non-coding *DHFR*	Transcriptional repression of *DHFR*	Inhibition of pre-initiation complex formation
GAS5	Repression of glucocorticoid receptor-mediated transcription	DNA mimicry
EVF2	Transcriptional activation of DLX2 targets, transcriptional repression of MeCP2 targets	Recruitment of DLX2 or MeCP2
CCND1 promoter RNA	Repression of *CCND1* transcription	Allosteric activation of TLS
NRON	Repression of NFAT-mediated transcription	Inhibition of transcription factor nucleocytoplasmic shuttling
Regulation of mRNA processing		
Neuroblastoma *MYC* (NAT)	Inhibition of neuroblastoma *MYC* intron 1 splicing	Unknown mechanism involving the inhibition of splicing via RNA–RNA duplex formation
Rev-ErbAalpha	Inhibition of the *c-ErbAalpha 2* splice isoform	Unknown mechanism involving the inhibition of splicing via RNA–RNA duplex formation
ZEB2 (NAT)	Activation of *ZEB2* translation	Unknown mechanism involving regulated splicing of an IRES-containing intron
MALAT1	Ser/Arg splicing factor regulation	Scaffolding of subnuclear domains
Sas10 mRNA 3′ UTR	Repression of *Rnp4F* mRNA	Unknown mechanism involving RNA editing
Modulation of mRNA post-transcriptional regulatory pathways		
Antisense *UCHL1*	Upregulation of UCHL1 protein production	SINE2B element-mediated translational upregulation
KCS1 antisense	Production of truncated KCS1 protein	Unknown mechanism involving base pairing
1/2-sbsRNA 1	Down-regulation of *SERPINE1* and *FLJ21870* mRNAs	Staufen-mediated decay through Alu element base pairing
BACE1AS	Up-regulation of *BACE1*	Stabilization of *BACE1* mRNA by blocking miRNA-induced repression
LINCMD1	Control of muscle differentiation through upregulation of MAML1 and MEF2C transcription factors	Sequestration of miRNAs
HULC	Downregulation of miRNA-mediated repression	Sequestration of miRNAs
PTENP1 pseudogene	Upregulation of PTEN	Sequestration of miRNAs
IPS1	Downregulation of miRNA-mediated repression	Sequestration of miRNAs
CDR1as	Downregulation of miRNA-mediated repression	Sequestration of miRNAs

Table 3 - LncRNAs-mediated regulation of gene expression[2]

LincRNAs exert these roles thanks to their intrinsic propensity to fold into thermodynamically stable secondary and higher orders structures that function as interaction modules. Each module can fold independently from another, forming bonds at the level of Watson-Crick, Hoogstein and ribose face[43, 44]. These RNAs can rapidly shift between diverse stable structural conformation, allowing allosteric transitions that can act as switches in response to environmental stimuli. They are also processed faster than mRNA, given that they must not be translated, allowing a rapid response to signals. LincRNAs can also be regulated via more than a hundred different nucleotide modifications, like in the case of tRNAs, rRNAs and snoRNAs[45, 46], that modulate their function and probably their structure. RNAs can generate multiple modules within their structure, allowing the interaction with multiple players, the reception of multiple stimuli and the generation of multiple outputs. The pairing required is extremely flexible, such as in the case of microRNAs, and allows mismatches, bulges and wobblings[47]. Many of these interaction modules derive from repetitive elements, such as transposons that took advantage of the fewer constraints that lincRNAs sequences have compared to protein coding genes[1, 48]. Indeed, lncRNAs rate of sequence evolution is higher relative to protein coding genes, even if these transcripts exhibit in any case evolutionary signatures of functionality. They evolved under modest but detectable selective pressure, accumulating fewer substitutions than neutrally evolving sequences[49, 50]. Likely, conservation of relatively small units of lncRNAs sequences (estimated to be less than 5%) could be sufficient to preserve their function, considering the already mentioned modular

structure[51]. This could be the reason why actual evolutionary tools fail to detect low level and scattered selective constraint within these loci[51].

Through such a plastic and versatile structure, lincRNAs can exert their functions binding to proteins, other RNAs[52] and probably also DNA, even if there is still little evidence on the existence of DNA:RNA triplex[53, 54]. In particular, lincRNAs can act as scaffolds, bridging together different molecules in a coordinated hub, like in the case of NEAT1: a highly abundant lncRNA that controls sequestration of proteins involved in the formation of paraspeckles, nuclear domains associated with mRNA retention and pathologically enriched in influenza and herpes viruses infections[55, 56]. LincRNA can also act as guides, recruiting proteins at specific loci: this has been hypothesized in the case of recombination events that mediate genetic diversity in developing lymphocytes as class switch (CSR) and V(D)J recombinations that seem to be mediated by sense and antisense transcripts that dictates the locations of combinatorial events[57, 58, 59]. Again lincRNAs can act as control devices or riboswitches in response to extracellular stimuli. For example, they can act as decoys, precluding pre-existing interactions such as GAS-5 RNA that detach glucocorticoid receptor from its responsive elements in conditions of growth-arrest[60, 61] (figure 4).

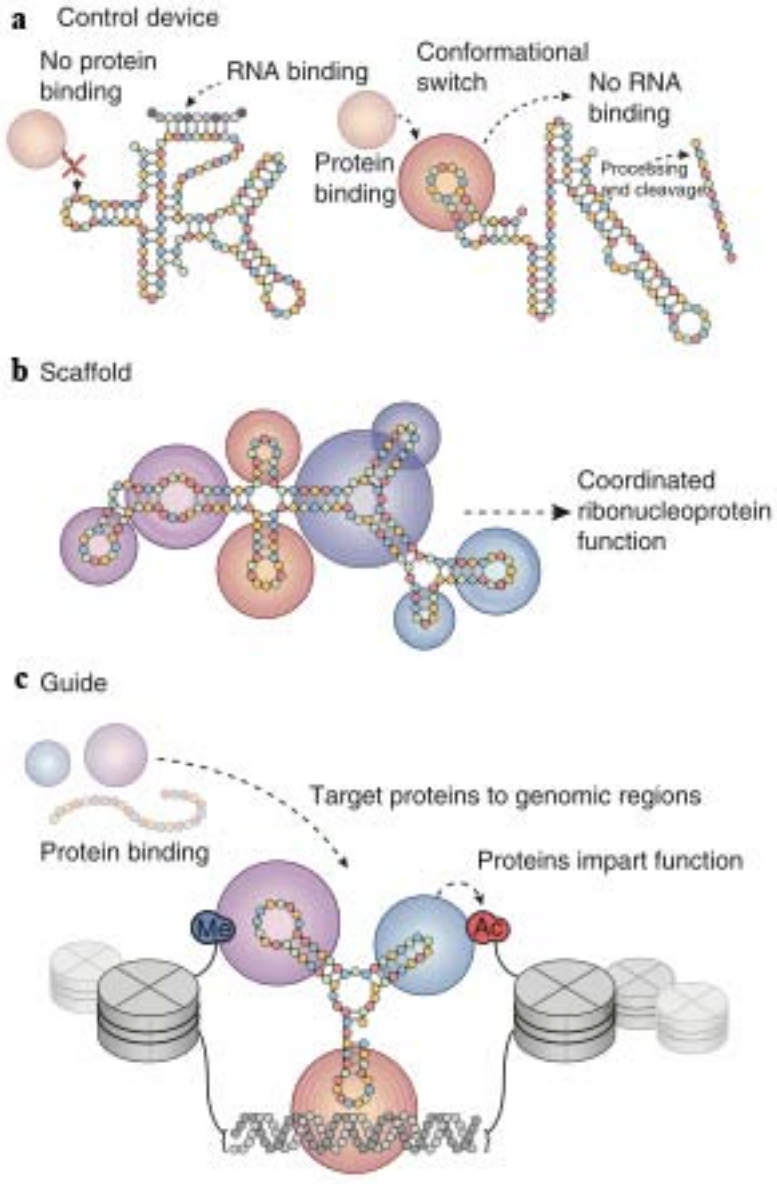


Figure 4 - The versatile and modular structure of lncRNAs allows them to act as: (a) allosteric inhibitors, preventing proteins from binding or detaching them from interactors (decoy); (b) scaffolds, binding multiple proteins; (c) guides, recruiting proteins[62].

Nonetheless, the regulatory potential of lincRNAs has been better characterized in the context of the epigenetic regulation of transcription that ultimately defines the cell transcriptome.

Epigenetics roles for long noncoding RNAs

Histones and DNA modifications together with the tridimensional chromosomes conformation within the nucleus define, at least in part, the epigenetic landscape of the cell. This extremely dynamic context modulates gene expression and dictates the final transcriptional output in response to environmental stimuli. By definition, these modifications are then propagated throughout cell divisions. This process is important in every moment of cell life, but particularly during differentiation. Indeed, every cell within our body harbor the same genome, but every cell acquires a particular phenotype according to intrinsic and extrinsic clues that ultimately defines its epigenome and therefore its fate during differentiation. Epigenetics also defines to what extent this fate can be irreversible or plastic.

Human lymphocytes are an interesting model system for understanding the basis of cell fate specification and plasticity. Indeed, although traditionally the broad range of effector lymphocytes has been referred to as constituted by distinct lineages, it has become increasingly clear that these cells also have notable features of plasticity. Differentiation of naïve cells into specific helper subsets requires the integration of extrinsic cues that converge into cell-intrinsic changes in the epigenetic landscape on the genome. The

interest within the field has been focused on the regulation of prototypical cytokine genes for each subset such as *Ifng* gene for T_H1 or *Il-4* for T_H2 $CD4^+$ lymphocytes. Much work has been done in both cases to define the complex genetic structure of these loci and the *cis* regulatory elements bound by transcription factors and chromatin modifiers promoting or repressing their transcription[63, 64, 65, 66, 67]. The importance of the setting of epigenetic memory at these fundamental loci has been underlined also by treatment with DNA methylation inhibitors[68, 69] or histones deacetylases inhibitors[70, 71, 72] and by deletion of DNA methyltransferases[73, 74, 75], which caused respectively: constitutive production of IFN-γ, enhanced production of both T_H1 and T_H2 prototypic cytokines and inability to activate the proper expression pattern of cytokines. The same is true for deletion of components of Trithorax group (TrxG) or Polycomb repressive complex (PRC) that dictates active or repressive epigenetic marks at fundamental loci for proper T helper cell differentiation, such as *Il-4*, *Il-5*, *Il-13* and *Gata3*[76, 77, 78, 79, 80, 81]. The pattern of chromatin marks is conventional for signature cytokines: active marks are present at prototypical cytokines whereas repressive marks restrain the expression of antagonistic molecules. However master regulators and other transcription factors usually referred to as definers of lineage-specific identity are characterized by bivalent poised domains, in which both active and repressive chromatin marks are present[82, 83]. This histone epigenetic status is peculiar also to promoters in embryonic stem cells, where it poises the expression of key developmental genes thus allowing their timely activation in presence of differentiative signals and concomitantly precluding expression in

their absence[84]. Indeed, while the expression of master transcription factor is quite rapid, cell divisions are required for cytokine loci to become accessible or conversely repressed[85, 86]. GATA3 and T-bet/STAT proteins initiate the epigenetic changes at IFN-γ and IL-4 loci that follow the initial activation of naïve T cells and differentiation toward T_H1 and T_H2 cell fate. These observations imply that T helper cells harbor both clear-cut and plastic epigenetic marks. Nonetheless we must consider that even epigenetically clearly defined cytokines genes can be expressed or repressed in unexpected context, as reported in T_H1 cells converted in IL-4-producing cells during strong T_H2-polarizing helminth infections[87] or stable T_H1/ T_H2 hybrid cells derived after parasite infections[88]. Therefore other players must be involved to define the degree of plasticity of lymphocytes in response to these ever-changing environmental conditions.

LincRNAs have been linked to epigenetic control of gene expression since the first studies regarding the well-known Xist transcript, involved in X chromosome inactivation in eutherians. Many other lincRNAs have been associated to chromatin or DNA modifiers and even transcription factors, thanks to specific mechanistic studies or high-througput screenings[89, 90, 91, 92]. This interplay can be observed across a broad range of eukaryotic organisms, suggesting that the epigenetic role of lincRNAs is conserved, even if their mere sequence conservation is often limited. It seems that lincRNAs could act as scaffolds, physically associating with proteins that modify chromatin either activating or repressing gene expression. Thanks to the already discussed structural properties of RNA, lincRNAs could organize multiple players in spatially and

temporally concerted actions[91]. Not only: thanks to their ability to base pair with other nucleic acids, they could recruit these modifiers at specific loci, therefore conferring them specificity of action[52]. This property has been an unsolved question, given that chromatin modifiers do not possess intrinsic bias toward consensus sequences, at least in mammals, while in *Drosophila* these 'docking sites' are well defined[93, 94]. Interestingly, while many of these enzymes lack DNA binding properties, they instead possess RNA binding motifs[95, 96, 97].

The majority of reported lincRNAs are linked to repression of gene transcription, in particular by interacting with Polycomb Group (PcG) proteins. The first examples of a direct interaction with Polycomb Repressive Complex 2 (PRC2) are the already mentioned Xist[98] and Kcnq1ot1, expressed only in the mammalian paternal chromosome and involved in the silencing of 8-10 protein-coding genes[99]. In both these cases, lincRNAs are strictly required for the enrichment of PRC2-associated proteins and for the trimethylation of the lysine 27 of histone H3 at specific loci. Furthermore, lincRNAs have been found to act as scaffolds and modulate PcG bodies: foci of PcG proteins are aggregated rather then dispersed in nuclei[100, 101, 102]. Indeed, NEAT2 and TUG1 promote relocation of growth-control genes at these subnuclear structures in response to mitogenic signal, therefore likely facilitating the concerted repression/activation of the transcription units[103]. Many other protein complexes have been found to interact with lincRNAs, the majority targeting histones, either methylases or demethylases, but other involved in DNA methylation. Indeed lincRNAs can bind proteins part of the Thritorax Group (TrxG)[36, 104, 105, 106, 107], that antagonize PcG-mediated silencing[108, 109].

Interestingly, an antisense lncRNA has been recently involved in recruiting a regulator of DNA demethylation at a specific promoter[110]. This process remains still largely unknown, being referred to as passive for a long time and only recently associated to active enzymatic reactions, via TET family of methylcytosine dioxygenases[111, 112]. Even in this case, one of the unsolved questions has been how locus-specificity can be achieved. Particularly, DNA demethylation is often restricted to few dinucleotides at the TSS. Though, the precise mechanism though which lncRNAs could direct DNA or chromatin modification has never been described. Indeed in all reported examples, correlations have been described between lincRNA-modifiers associations and loss of modification after lincRNA gene silencing. LincRNAs are supposed to confer binding specificity to modifiers and recruiting them either *in cis* or *in trans* (figure 5). In the first case, lincRNAs could act directly on sites where they are synthesized without needing to leave DNA. The current hypothesis suggests that the 5' region of the nascent transcript could bind proteins while the 3' is transcriptionally lagging, being still tethered to chromatin by RNA polymerase[113]. This model is particularly intriguing as through this mechanism lincRNAs could exert an allele-specific effect, like in the well-studied case of Xist. *In trans* regulation is instead achieved when lincRNAs act modulating genes across great distances or even on different chromosomes[114].

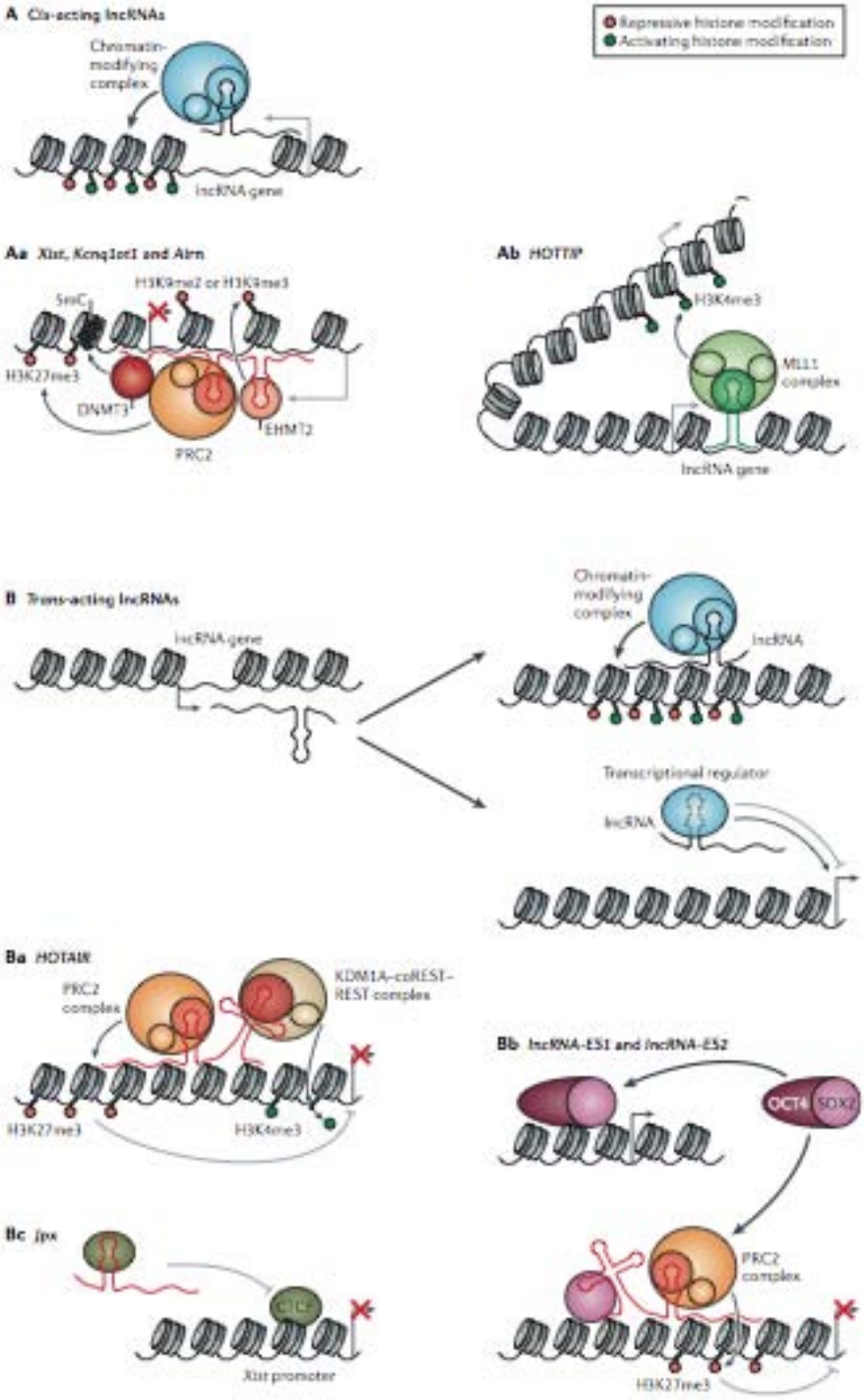


Figure 5 - Models of nuclear lncRNAs function with examples[115].

Regarding this dichotomy, we must underline once again its artificiality. Indeed chromosomes fold into complex, three-dimensional territories together with specialized subnuclear bodies. These foci are enriched for proteins that are part of the transcriptional or splicing machinery and for regulators of these processes[116, 117]. These structures are not static, but on the contrary large-scale chromosomal repositioning is observed in response to environmental stimuli or during differentiation, that is dependent on the active remodeling of the nucleoskeleton[118, 119, 120, 121]. The dynamic folding of the genome into higher order structure encompasses loci belonging to the same chromosome, even hundreds of kilobases apart, or different ones, bringing together regions that are distant if we consider the genome as linear. Therefore in this context, is extremely difficult to discern what regulations are *in cis* or *in trans*, especially when they involve long distance interactions. Intriguingly, lincRNAs have been found that regulate the formation of subnuclear structures, such as NEAT1, required for paraspeckles nucleation[122]. LncRNAs can also affect directly the three-dimensional organization of chromosomes enhancing the function of proteins involved in looping formation, like the insulator protein CTCF[123]. There are also many examples of lincRNAs involved in three-dimensional local chromatin looping that brings together the ncRNA gene with the region that it regulates within the same chromosome[36, 124, 125] (figure 6). Recently, a lincRNA called Firre has been shown to recruit specific gene loci located on different chromosomes, acting as a docking station for organizing *trans*-chromosomal associations. Consistently, genetic deletion of Firre leads to a loss of proximity of several *trans*-interactions[126]. A

peculiar type of lincRNA has been described that is transcribed from enhancer regions (eRNAs or activating lincRNAs: ncRNA-a). Classic enhancer elements therefore likely act through transcription of these lincRNAs that upregulates expression at promoters via the recruitment of Mediator complex[124, 127]. Finally, there is increasing evidence that even promoters could be transcribed[128], producing lincRNAs probably involved in the enhancer-promoter loop that was hypothesized years ago but never fully resolved[129].

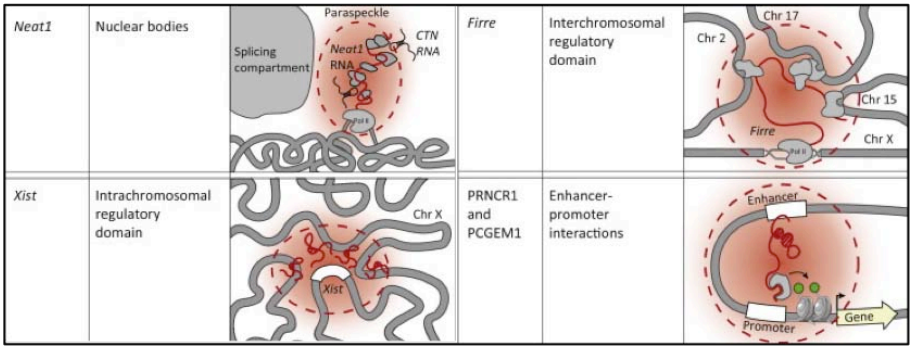


Figure 6 - LncRNAs can shape 3D nuclear structure[130].

Recently, the idea that lincRNAs could act as guides, and particularly as PRC2-recruiters, has been discussed. *In vitro* binding assays revealed a promiscuous RNA-binding activity by PRC2. Correlative analysis reveal that the fraction of EZH2-associated transcripts in WT compared to *EZH2*$^{-/-}$ cells correlates positively with active genes and negatively to repressed ones. ChIP-seq highlights that RNA-associated-PRC2 is never deposited to promoter regions of active genes though there is a small fraction of genes enriched both for EZH2 and H3K36me3 or H3K4me3 marks, in absence of H3K27me3. These observations led to the “junk-mail” hypothesis that

promiscuous RNA binding by PRC2 allows the identification of spurious transcription derived from not fully silenced genes, already partially decorated by H3K27me3. This could allow PRC2 to restore the repression. Conversely, if these genes are decorated by active marks (H3K4me3 and H3K36me3), RNA-mediated PRC2 binding to nucleosome could be inhibited and therefore the expression maintained[131]. These results are in line with a re-evaluation of PRC2 binding pattern, especially in embryonic stem-cells[132, 133]. Nonetheless, these genome-wide correlation studies should not necessarily be regarded to as in conflict with previously mentioned functional studies neither in the case of PRC2 nor with other RNA-interactors. Indeed, the broad snapshot they depict could fail to appreciate functional relationships reported in specific cases. Another recent study indeed goes into details regarding RNA binding efficiencies of the key components of PRC2. It seems that while EZH2 alone is able to bind RNA in a nonselective fashion, the PRC2 complex as a whole clearly discriminates between specific and nonspecific RNAs. Interestingly, binding of RNA to PRC2 reduces its methyltransferase activity while JARID2 can negatively modulate the interaction, increasing the catalytic activity of the complex[134]. Conversely, another paper gives a hint indicating JARID2 as the recruiter of PRC2 via lncRNAs binding[135]. Recently another paper identifies a novel player that regulates PRC2 activity: ATRX is a high-affinity RNA-binding protein that directly interacts with Xist RNA to promote the loading of PRC2 *in vivo*. The loss of ATRX leads to a global redistribution of PRC2 and a derepression of repressed genes[136]. These highly debated studies highlights once again the impact that lncRNAs have on the

entire genome. The comprehension of the mechanisms that regulate the recruitment of chromatin remodeling complexes by lncRNAs as well as the other fundamental roles they play is therefore of key importance.

Long noncoding RNAs in the immune system

The immune system is an extraordinary context for the study of the role of lincRNAs in differentiation. Indeed, upon antigen stimuli, naïve $CD4^+$ T cells differentiate into distinct T helper subsets that were traditionally referred as lineages and defined by a prototypic set of expressed cytokines and master transcription factors (TFs). Recently this relative simple scenario, although useful, has been subjected to debate. $CD4^+$ T cells demonstrated to exhibit substantial plasticity and it has become increasingly clear that they can change the pattern of cytokines and transcription factors according to the milieu they encounter through their life[137]. Not only, in some cases they can concomitantly express other cytokines and transcription factors together with their prototypical set. Best examples include IL-10, once thought to specifically identify T_H2 and now known to be produced also by T_H1, T_{regs} and T_H17 cells[138] and IFN-γ, the classic T_H1 cytokine, frequently released by T_H17 cells simultaneously with IL17[139, 140]. Regarding master transcription factors, T_{regs} can express Foxp3 (their prototypical TF), but also RORγt (Th17 TF) and Runx3[141, 142, 143] and Tfh cells can differentiate from Foxp3 positive cells also expressing Bcl6 (their specific TF)[144, 145]. In this context, lincRNAs have a fundamental role in governing flexibility and

plasticity or maintenance of cell identity, together with lineage-specific transcription factors and other ncRNAs. In particular, what is emerging from literature is that ncRNAs typically act as fine-tuners of fate choices and this seems to be true not only in the immune system. Nonetheless, in the case of $CD4^+$ T cell subsets that are specified but not fully determined, subtle changes in extrinsic signals can reverberate through responsive ncRNAs inducing changes that alter cell phenotype[8, 146, 147]. Usually, the stability of lineage identity is achieved through the implementation and inheritance of epigenetic modification, but as mentioned before lincRNAs can act directly on histone and DNA modifiers redefining this context. Conversely, lincRNAs can also buffer this situation in other conditions, acting as maintainers of cell identity. In the cellular system, lincRNAs can be regarded as minor nodes in a huge interconnected network[148]: they usually interact with few other players. This condition allows them to be more flexible and sensitive to variations without disrupting the whole network integrity[149]. This is true both in a very short period, as cells can easily and rapidly adapt to environment, and also in long evolutionary periods, as lincRNAs are among the fastest evolving sequences in the genome[49, 150, 151, 152]. Conversely, master transcription regulators can be referred as highly connected hubs, which confers robustness to the network. Indeed very few protein-coding genes have been lost from worms to human and mutations are most often pathological[153, 154].

Several single-case or genome-wide studies on lincRNAs in the murine immune system are now available in literature, whereas

only few studies have been conducted until now in the human context (table 4).

LncRNA	Model system	Observation
Innate immune response		
Multiple	Coronavirus infection in mouse lung	RNA-seq demonstrated widespread differential expression of lncRNAs following lung infection with severe acute respiratory syndrome coronavirus in four mouse strains (129/S1, CAST, PWK, and WSB)
Multiple	LPS-stimulated mouse macrophages	Identification of multiple lincRNAs and eRNAs using pol II and H3K36me3 epigenetic marks. Eight of 11 lincRNAs were validated by qRT-PCR
LincRNA-Cox2	LPS-stimulated mouse bone marrow-derived dendritic cells	Identification of 20 lincRNAs including *lincRNA-Cox2* using deposition of epigenetic marks of active transcription (H3K4me3 at their promoters and H3K36me3 within the transcribed region)
LincRNA-Cox2	$Pam3CSK_4$-stimulated mouse bone marrow-derived macrophages	Revealed that *lincRNA-Cox2* repressed the expression of 787 genes in non-stimulated cells and the increased expression of 713 genes following exposure to $Pam3CSK_4$. The actions of *lincRNA-Cox2* were mediated through interaction with hnRNP-A/B and hnRNP-A2/B1
THRIL	$Pam3CSK_4$-stimulated human monocytic THP-1 cells	Microarray analysis identified 169 differentially expressed lincRNAs including down-regulation of antisense lncRNA *THRIL* (TNFα and hnRNPL related immunoregulatory lincRNA). *THRIL* was shown to regulate both basal and $Pam3CSK_4$-stimulated gene expression through an interaction with hnRNPL
Lethe	TNFα-stimulated mouse embryonic fibroblasts	RNA-seq identified 112 lncRNAs and 54 transcribed pseudogenes that were differentially expressed including *Rps15a-ps4* (renamed *Lethe*). *Lethe* was induced in response to IL1β and dexamethasone and shown to interact and block the binding of the RelA (p65) subunit of NF-κB
NEAT1	Poly(IC)- or influenza-stimulated HeLa and human epithelial A549 cells	Increased *NEAT1* expression induced the formation of paraspeckle formation. Redistribution of SFPQ from the *CXCL8* promoter to the paraspeckles following *NEAT1* binding leads to increased *CXCL8* expression
Ptprj-as1	LPS-stimulated mouse bone marrow-derived macrophages	Induced in response to LPS
IL1β-RBT48 and *IL1β-eRNA*	LPS-stimulated human monocytes and monocytic THP-1 cells	RNA-seq identified 76 eRNAs, 40 lincRNAs, 65 antisense RNAs, and 35 regions of bidirectional transcription (RBTs) that are differentially expressed. *IL1β-RBT48* and *IL1β-eRNA* were shown to regulate LPS-induced IL1β and CXCL8 expression
Unnamed	LPS-stimulated K562 leukemias cells	Multiple lncRNAs were located upstream of *TNF* and shown to negatively regulate *TNF* expression, possibly through binding to the transcriptional repressor, LRRFIP1 [leucine rich repeat (in FLII) interacting protein 1]
Lnc-IL7R	LPS-stimulated monocytic THP1 cells	*Lnc-IL7R* is transcribed from the 3′-UTR of *IL7R* in the sense orientation. Induced following LPS stimulation and negatively regulates IL7R, IL8, IL-6, VCAM-1, and E-selectin expression, a process associated with diminished H3K27me3 levels
PACER	PMA- and LPS-stimulated human U937 monocytic cell line	*PACER* (p50-associated COX-2 extragenic RNA) is expressed upstream of the *Cox2* promoter and positively regulates COX2 production. *PACER* binds to, and drives the release of, the repressive p50 dimer of NF-κB from the *Cox2* promoter
Lnc-DC	Differentiation of human and mouse dendritic cells	*Lnc-DC* (LOC645638) is required for monocyte differentiation into dendritic cells (DC). *Lnc-DC* promotes phosphorylation and activation of STAT3, a transcription involved in DC differentiation, by blocking its dephosphorylation by SHP1
Adaptive immune response		
Multiple	Human $CD8^+$ T cells	Microarray studies identified 100s of lymphoid-specific lncRNAs and showed differential expression during $CD8^+$ T cell activation and following differentiation into $CD8^+$ memory and effector T cells

LncRNA	Model system	Observation
NTT	Human T cell lines	*NTT* (noncoding transcript in $CD4^+$ T cells) was identified in activated T cells
Gas5	Human primary T cells and T cell lines (CEM-C7 and Jurkat)	*Gas5* (growth arrest specific transcript 5) levels increase upon growth arrest and inhibit cell-cycle progression and promote apoptosis
Gas5	Human primary T cells	Inhibition of T cell proliferation through the mTOR antagonist rapamycin is mediated by upregulation of *Gas5*
NRON	Human Jurkat T cell	*NRON* (noncoding repressor of NFAT) blocked the nucleocytoplasmic transport and therefore the transcriptional activity of NFAT through interaction with multiple proteins including members of the importin-β superfamily
NRON	Human Jurkat T cells and mouse T cells	*NRON* shown to attenuate NFAT dephosphorylation and thereby block NFAT nuclear translocation, activation, and induction of IL-2
NeST	Transgenic mouse infected with *Salmonella* and Theiler's virus	Overexpression of *NeST* (Nettoie Salmonella pas Theilers's) was shown to increase clearance of bacterial *Salmonella* infection but reduce resistance to the mouse Theiler's picornavirus. *NeST* induced the expression of IFN-γ through an interaction with WD repeat domain 5 (WDR5), a core subunit of the MLL histone H3 lysine 4 (H3K4) methyltransferase complex
LincR-Ccr2-5′AS	Mouse $CD4^+$ T_H2 cells	ftRNA-seq studies identified 1524 lincRNAs in 42 mouse T cell subsets. *LincR-Ccr2-5′AS* was located at the 5′ end of *Ccr2* in $CD4^+$ T_H2 cells and was shown to regulate both the induction and suppression of gene expression during T_H2 differentiation. *LincR-Ccr2-5′AS* is also implicated in chemokine-mediated signalling including cell migration
Multiple	Mouse T and B cells	LncRNAs shown to regulate chromatin remodelling associated with variable, diversity, and joining (V(D)J) recombination required to produce antigen receptors (Ig or TCR)
Multiple	Mouse B cells	Transcription of antisense and sense lncRNAs is associated with looping of V_H regions into close proximity with the DJ_H region during recombination in pro-B cells, a process that occurs within transcription factories
Pathogen-associated		
PAN	KSHV-infected B cell lines	PAN (polyadenylated nuclear) RNA expression from KSHV was shown to modulate host cell response including downregulation of IFN_γ, IL18, and α-interferon 16
PAN	KSHV-infected B- and T cell lines	PAN RNA-mediated suppression of host genes is mediated through polycomb repression complex 2 (PRC2)-mediated histone methylation

Table 4 - LncRNAs and immune response[155].

Nonetheless there are significant differences between experimental animal models and human, both regarding immunologic responses[156] and ncRNAs[157, 158]. In particular, lincRNAs are really fast-evolving elements as demonstrated by the fact that over 80% of the human lncRNAs that arose in the primate lineage, only 3% are conserved across tetrapods and most mammalian lncRNAs lack known orthologs outside vertebrates[159]. In detail, even between mouse and human, lncRNAs are poorly conserved[160, 161, 162]. Despite their rapid evolution, lncRNAs are selected more than neutral sequences

and in particular more than intergenic regions, but significantly less than mRNAs[50, 159, 163]. It must be underlined that the conservation rate reported could be overestimated: substitution rates are derived from whole-genome alignment and based on the assumption that even segment of homologies imply that that segment belongs to the same RNA class, but this is not necessarily the case. Indeed it could be that in another genome context that lncRNA gene segment is transcribed and processed as part of a protein-coding RNA[164]. A striking example is *Hotair* that is involved in the regulation of the highly conserved cluster of Hox genes[36]. The human lincRNA is conserved in the mouse genome[165], nonetheless only the 3' region is effectively part of the murine homolog[152]. Taking into account these considerations, it is of crucial importance to study lincRNAs specifically within the human in immune system, but this field is still poorly addressed. The majority of the studies focused on the innate immune system[166, 167, 168] or analysed pathological situations, such as cancer-related lncRNAs[169, 170] or responses to specific infections[171, 172, 173, 174], mostly in mice. The first functional study focused on the adaptive immune system, and in particular on T_H1 and T_H2 lymphocytes, involved a lincRNA that is selectively expressed in T_H1 cells via Stat4 and T-bet, both in mouse and human. It participates in the induction of IFN-γ expression strictly in response to T_H1 differentiation program and not in other cellular contexts. These results highlight once again the complexity of the gene expression regulatory network and the specificity of action of lincRNAs[175]. Another paper found a lincRNA specifically expressed in primary T_H2, instead, and hypothesized its coregulation with GATA3[176]. GAS-5 is degraded in optimal growth conditions, but it

accumulates contributing to growth arrest in starving conditions[61]. In this situation it competes with glucocorticoid receptors (GR) DNA-binding sequences, suppressing GR-mediated transcription[177]. Broader studies have been performed on the $CD8^+$ T cell transcriptome[178], and recently on $CD4^+$ T lymphocytes[179], but still on mice models. In B cells, chromatin remodeling associated with V(D)J recombination has been potentially linked to a widespread antisense intergenic transcription that occurs in the variable (V) region of the immunoglobulin heavy chain (Igh) locus[180, 181]. So far no studies have been published that performed a deep transcriptomic analysis on human primary lymphocytes from healthy donors, identifying lncRNAs fundamental for differentiation processes. These few examples are just clues of the importance that lincRNA could have also for the proper function of the human immune system and prompt to a deeper analysis of their role in this particularly intriguing context.

Scope of the thesis

In this thesis we investigated the transcriptome of human lymphocytes and in particular the expression of specific long intergenic non-coding RNAs (lincRNAs) expressed by thirteen lymphocytes subsets. We focused our attention on a T_H1-specific lincRNA that we called linc-MAF-4 due to its proximity to MAF gene. We provided evidences of the role of linc-MAF-4 in the maintenance of T_H1 cell identity via an epigenetic-mediated MAF dowregulation.

References

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

2. Geisler S, Coller J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nature reviews Molecular cell biology* 2013, **14**(11)**:** 699-712.

3. Taft RJ, Pheasant M, Mattick JS. The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays : news and reviews in molecular, cellular and developmental biology* 2007, **29**(3)**:** 288-299.

4. Hoagland MB, Keller EB, Zamecnik PC. Enzymatic carboxyl activation of amino acids. *The Journal of biological chemistry* 1956, **218**(1)**:** 345-358.

5. Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, *et al.* Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* 2000, **408**(6808)**:** 86-89.

6. Kowalczyk MS, Higgs DR, Gingeras TR. Molecular biology: RNA discrimination. *Nature* 2012, **482**(7385)**:** 310-311.

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

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

9. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, *et al.* Chromatin signature reveals over a thousand highly

conserved large non-coding RNAs in mammals. *Nature* 2009, **458**(7235)**:** 223-227.

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

11. Dieci G, Fiorino G, Castelnuovo M, Teichmann M, Pagano A. The expanding RNA polymerase III transcriptome. *Trends in genetics : TIG* 2007, **23**(12)**:** 614-622.

12. Lin R, Maeda S, Liu C, Karin M, Edgington TS. A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. *Oncogene* 2007, **26**(6)**:** 851-858.

13. Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS biology* 2007, **5**(5)**:** e106.

14. Lanz RB, McKenna NJ, Onate SA, Albrecht U, Wong J, Tsai SY, *et al.* A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. *Cell* 1999, **97**(1)**:** 17-27.

15. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, *et al.* Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell reports* 2014, **8**(5)**:** 1365-1379.

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

17. Banfai B, Jia H, Khatun J, Wood E, Risk B, Gundling WE, Jr., *et al.* Long noncoding RNAs are rarely translated in two human cell lines. *Genome research* 2012, **22**(9)**:** 1646-1657.

18. Lee S, Liu B, Lee S, Huang SX, Shen B, Qian SB. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**(37)**:** E2424-2432.

19. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, *et al.* Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature chemical biology* 2013, **9**(1)**:** 59-64.

20. Andrews SJ, Rothnagel JA. Emerging evidence for functional peptides encoded by short open reading frames. *Nature reviews Genetics* 2014, **15**(3)**:** 193-204.

21. Ulveling D, Francastel C, Hube F. When one is better than two: RNA with dual functions. *Biochimie* 2011, **93**(4)**:** 633-644.

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

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

24. Duret L, Chureau C, Samain S, Weissenbach J, Avner P. The Xist RNA Gene Evolved in Eutherians by Pseudogenization of a Protein-Coding Gene. *Science (New York, NY)* 2006, **312**(5780)**:** 1653-1655.

25. Bussotti G, Notredame C, Enright AJ. Detecting and comparing non-coding RNAs in the high-throughput era. *International journal of molecular sciences* 2013, **14**(8)**:** 15423-15458.

26. Nigro JM, Cho KR, Fearon ER, Kern SE, Ruppert JM, Oliner JD, *et al.* Scrambled exons. *Cell*, **64**(3)**:** 607-613.

27. Beltran M, Puig I, Pena C, Garcia JM, Alvarez AB, Pena R, *et al.* A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes & development* 2008, **22**(6)**:** 756-769.

28. Martick M, Horan LH, Noller HF, Scott WG. A discontinuous hammerhead ribozyme embedded in a mammalian messenger RNA. *Nature* 2008, **454**(7206)**:** 899-902.

29. Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, *et al.* Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nature biotechnology* 2012, **30**(1)**:** 99-104.

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

31. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, *et al.* Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS genetics* 2013, **9**(4)**:** e1003470.

32. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome biology* 2012, **13**(11)**:** R107.

33. Lu X, Sachs F, Ramsay L, Jacques P-É, Göke J, Bourque G, *et al.* The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat Struct Mol Biol* 2014, **21**(4)**:** 423-425.

34. Lin N, Chang KY, Li Z, Gates K, Rana ZA, Dang J, *et al.* An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. *Molecular cell* 2014, **53**(6)**:** 1005-1019.

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

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

37. Yin D, He X, Zhang E, Kong R, De W, Zhang Z. Long noncoding RNA GAS5 affects cell proliferation and predicts a poor prognosis in patients with colorectal cancer. *Medical oncology (Northwood, London, England)* 2014, **31**(11)**:** 253.

38. Rossi MN, Antonangeli F. LncRNAs: New Players in Apoptosis Control. 2014, **2014:** 473857.

39. Standaert L, Adriaens C, Radaelli E, Van Keymeulen A, Blanpain C, Hirose T, *et al.* The long noncoding RNA Neat1 is required for mammary gland development and lactation. *RNA (New York, NY)* 2014.

40. Autuoro JM, Pirnie SP, Carmichael GG. Long Noncoding RNAs in Imprinting and X Chromosome Inactivation. *Biomolecules* 2014, **4**(1)**:** 76-100.

41. Gendrel AV, Heard E. Noncoding RNAs and Epigenetic Mechanisms During X-Chromosome Inactivation. *Annual review of cell and developmental biology* 2014, **30:** 561-580.

42. Hu W, Alvarez-Dominguez JR, Lodish HF. Regulation of mammalian cell differentiation by long non-coding RNAs. *EMBO reports* 2012, **13**(11)**:** 971-983.

43. Cruz JA, Westhof E. The dynamic landscapes of RNA architecture. *Cell* 2009, **136**(4)**:** 604-609.

44. Lescoute A, Westhof E. Topology of three-way junctions in folded RNAs. *RNA (New York, NY)* 2006, **12**(1)**:** 83-93.

45. Jia G, Fu Y, Zhao X, Dai Q, Zheng G, Yang Y, *et al.* N6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nature chemical biology* 2011, **7**(12)**:** 885-887.

46. Squires JE, Patel HR, Nousch M, Sibbritt T, Humphreys DT, Parker BJ, *et al.* Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic acids research* 2012, **40**(11)**:** 5023-5033.

47. Chi SW, Hannon GJ, Darnell RB. An alternative mode of microRNA target recognition. *Nat Struct Mol Biol* 2012, **19**(3)**:** 321-327.

48. Mariner PD, Walters RD, Espinoza CA, Drullinger LF, Wagner SD, Kugel JF, *et al.* Human Alu RNA Is a Modular Transacting Repressor of mRNA Transcription during Heat Shock. *Molecular cell* 2008, **29**(4)**:** 499-509.

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

50. Ponjavic J, Ponting CP, Lunter G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome research* 2007, **17**(5)**:** 556-565.

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

52. Engreitz Jesse M, Sirokman K, McDonel P, Shishkin AA, Surka C, Russell P, *et al.* RNA-RNA Interactions Enable Specific Targeting of Noncoding RNAs to Nascent Pre-mRNAs and Chromatin Sites. *Cell* 2014, **159**(1)**:** 188-199.

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

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

55. Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, Chess A, *et al.* An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Molecular cell* 2009, **33**(6)**:** 717-726.

56. Imamura K, Imamachi N, Akizuki G, Kumakura M, Kawaguchi A, Nagata K, *et al.* Long noncoding RNA NEAT1-dependent SFPQ relocation from promoter region to paraspeckle mediates IL8 expression upon immune stimuli. *Molecular cell* 2014, **53**(3)**:** 393-406.

57. Abarrategui I, Krangel MS. Noncoding transcription controls downstream promoters to regulate T-cell receptor alpha recombination. *The EMBO journal* 2007, **26**(20)**:** 4380-4390.

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

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

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

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

62. Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol* 2013, **20**(3)**:** 300-307.

63. Balasubramani A, Mukasa R, Hatton RD, Weaver CT. Regulation of the Ifng locus in the context of T-lineage specification and plasticity. *Immunological reviews* 2010, **238**(1)**:** 216-232.

64. Collier SP, Henderson MA, Tossberg JT, Aune TM. Regulation of the Th1 Genomic Locus from Ifng through Tmevpg1 by T-bet. *Journal of immunology (Baltimore, Md : 1950)* 2014, **193**(8)**:** 3959-3965.

65. Chang S, Aune TM. Dynamic changes in histone-methylation 'marks' across the locus encoding interferon-gamma during the differentiation of T helper type 2 cells. *Nature immunology* 2007, **8**(7)**:** 723-731.

66. Schieck M, Sharma V, Michel S, Toncheva AA, Worth L, Potaczek DP, *et al.* A polymorphism in the TH 2 locus control region is associated with changes in DNA methylation and gene expression. *Allergy* 2014, **69**(9)**:** 1171-1180.

67. Tykocinski LO, Hajkova P, Chang HD, Stamm T, Sozeri O, Lohning M, *et al.* A critical control element for interleukin-4 memory expression in T helper lymphocytes. *The Journal of biological chemistry* 2005, **280**(31)**:** 28177-28185.

68. Ballas ZK. The use of 5-azacytidine to establish constitutive interleukin 2-producing clones of the EL4 thymoma. *Journal of immunology (Baltimore, Md : 1950)* 1984, **133**(1)**:** 7-9.

69. Young HA, Ghosh P, Ye J, Lederer J, Lichtman A, Gerard JR, *et al.* Differentiation of the T helper phenotypes by analysis of the methylation state of the IFN-gamma gene. *Journal of immunology (Baltimore, Md : 1950)* 1994, **153**(8)**:** 3603-3610.

70. Bird JJ, Brown DR, Mullen AC, Moskowitz NH, Mahowald MA, Sider JR, *et al.* Helper T cell differentiation is controlled by the cell cycle. *Immunity* 1998, **9**(2)**:** 229-237.

71. Morinobu A, Kanno Y, O'Shea JJ. Discrete roles for histone acetylation in human T helper 1 cell-specific gene expression. *The Journal of biological chemistry* 2004, **279**(39)**:** 40640-40646.

72. Valapour M, Guo J, Schroeder JT, Keen J, Cianferoni A, Casolaro V, *et al.* Histone deacetylation inhibits IL4 gene expression in T cells. *The Journal of allergy and clinical immunology* 2002, **109**(2)**:** 238-245.

73. Hutchins AS, Mullen AC, Lee HW, Sykes KJ, High FA, Hendrich BD, *et al.* Gene silencing quantitatively controls the function of a developmental trans-activator. *Molecular cell* 2002, **10**(1)**:** 81-91.

74. Lee PP, Fitzpatrick DR, Beard C, Jessup HK, Lehar S, Makar KW, *et al.* A critical role for Dnmt1 and DNA methylation in T cell development, function, and survival. *Immunity* 2001, **15**(5)**:** 763-774.

75. Makar KW, Perez-Melgosa M, Shnyreva M, Weaver WM, Fitzpatrick DR, Wilson CB. Active recruitment of DNA methyltransferases regulates interleukin 4 in thymocytes and T cells. *Nature immunology* 2003, **4**(12)**:** 1183-1190.

76. Yamashita M, Hirahara K, Shinnakasu R, Hosokawa H, Norikane S, Kimura MY, *et al.* Crucial role of MLL for the maintenance of memory T helper type 2 cell responses. *Immunity* 2006, **24**(5)**:** 611-622.

77. Onodera A, Yamashita M, Endo Y, Kuwahara M, Tofukuji S, Hosokawa H, *et al.* STAT6-mediated displacement of polycomb by trithorax complex establishes long-term maintenance of GATA3 expression in T helper type 2 cells. *The Journal of experimental medicine* 2010, **207**(11)**:** 2493-2506.

78. Kimura M, Koseki Y, Yamashita M, Watanabe N, Shimizu C, Katsumoto T, *et al.* Regulation of Th2 cell differentiation by mel-18, a mammalian polycomb group gene. *Immunity* 2001, **15**(2)**:** 275-287.

79. Yamashita M, Kuwahara M, Suzuki A, Hirahara K, Shinnaksu R, Hosokawa H, *et al.* Bmi1 regulates memory CD4 T cell survival via repression of the Noxa gene. *The Journal of experimental medicine* 2008, **205**(5)**:** 1109-1120.

80. Koyanagi M, Baguet A, Martens J, Margueron R, Jenuwein T, Bix M. EZH2 and histone 3 trimethyl lysine 27 associated with Il4 and Il13 gene silencing in Th1 cells. *The Journal of biological chemistry* 2005, **280**(36)**:** 31470-31477.

81. Su IH, Dobenecker MW, Dickinson E, Oser M, Basavaraj A, Marqueron R, *et al.* Polycomb group protein ezh2 controls actin polymerization and cell signaling. *Cell* 2005, **121**(3)**:** 425-436.

82. Wei G, Wei L, Zhu J, Zang C, Hu-Li J, Yao Z, *et al.* Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. *Immunity* 2009, **30**(1)**:** 155-167.

83. Hirahara K, Vahedi G, Ghoreschi K, Yang XP, Nakayamada S, Kanno Y, *et al.* Helper T-cell differentiation and plasticity: insights from epigenetics. *Immunology* 2011, **134**(3)**:** 235-245.

84. Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. *Cell* 2007, **128**(4)**:** 669-681.

85. Grogan JL, Mohrs M, Harmon B, Lacy DA, Sedat JW, Locksley RM. Early transcription and silencing of cytokine genes underlie polarization of T helper cell subsets. *Immunity* 2001, **14**(3)**:** 205-215.

86. Mullen AC, High FA, Hutchins AS, Lee HW, Villarino AV, Livingston DM, *et al.* Role of T-bet in Commitment of TH1 Cells Before IL-12-Dependent Selection. *Science (New York, NY)* 2001, **292**(5523)**:** 1907-1910.

87. Panzer M, Sitte S, Wirth S, Drexler I, Sparwasser T, Voehringer D. Rapid in vivo conversion of effector T cells into Th2 cells during helminth infection. *Journal of immunology (Baltimore, Md : 1950)* 2012, **188**(2)**:** 615-623.

88. Peine M, Rausch S, Helmstetter C, Frohlich A, Hegazy AN, Kuhl AA, *et al.* Stable T-bet(+)GATA-3(+) Th1/Th2 hybrid cells arise in vivo, can develop directly from naive precursors, and limit immunopathologic inflammation. *PLoS biology* 2013, **11**(8)**:** e1001633.

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

90. Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, *et al.* Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Molecular cell* 2010, **40**(6)**:** 939-953.

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

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

93. Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R, Tolhuis B, *et al.* Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. *PLoS biology* 2009, **7**(1)**:** e13.

94. Kassis JA, Brown JL. Polycomb group response elements in Drosophila and vertebrates. *Advances in genetics* 2013, **81:** 83-118.

95. Bernstein E, Allis CD. RNA meets chromatin. *Genes & development* 2005, **19**(14)**:** 1635-1655.

96. Jeffery L, Nakielny S. Components of the DNA methylation system of chromatin control are RNA-binding proteins. *The Journal of biological chemistry* 2004, **279**(47)**:** 49479-49487.

97. Hiragami-Hamada K, Fischle W. RNAs - physical and functional modulators of chromatin reader proteins. *Biochimica et biophysica acta* 2014, **1839**(8)**:** 737-742.

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

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

100. Saurin AJ, Shiels C, Williamson J, Satijn DP, Otte AP, Sheer D, *et al.* The human polycomb group complex associates with pericentromeric heterochromatin to form a novel nuclear domain. *The Journal of cell biology* 1998, **142**(4)**:** 887-898.

101. Li HB, Ohno K, Gui H, Pirrotta V. Insulators target active genes to transcription factories and polycomb-repressed genes to polycomb bodies. *PLoS genetics* 2013, **9**(4)**:** e1003436.

102. Cheutin T, Cavalli G. Polycomb silencing: from linear chromatin domains to 3D chromosome folding. *Current opinion in genetics & development* 2014, **25**(0)**:** 30-37.

103. Yang L, Lin C, Liu W, Zhang J, Ohgi KA, Grinstein JD, *et al.* ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. *Cell* 2011, **147**(4)**:** 773-788.

104. Grote P, Wittler L, Hendrix D, Koch F, Wahrisch S, Beisaw A, *et al.* The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Developmental cell* 2013, **24**(2)**:** 206-214.

105. Yang YW, Flynn RA, Chen Y, Qu K, Wan B, Wang KC, *et al.* Essential role of lncRNA binding for WDR5 maintenance of active chromatin and embryonic stem cell pluripotency. *eLife* 2014, **3:** e02046.

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

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

108. Ringrose L, Paro R. Polycomb/Trithorax response elements and epigenetic memory of cell identity. *Development (Cambridge, England)* 2007, **134**(2)**:** 223-232.

109. Steffen PA, Ringrose L. What are memories made of? How Polycomb and Trithorax proteins mediate epigenetic memory. *Nature reviews Molecular cell biology* 2014, **15**(5)**:** 340-356.

110. Arab K, Park Yoon J, Lindroth Anders M, Schäfer A, Oakes C, Weichenhan D, *et al.* Long Noncoding RNA TARID Directs Demethylation and Activation of the Tumor Suppressor TCF21 via GADD45A. *Molecular cell* 2014, **55**(4)**:** 604-614.

111. Ponnaluri VK, Maciejewski JP, Mukherji M. A mechanistic overview of TET-mediated 5-methylcytosine oxidation. *Biochemical and biophysical research communications* 2013, **436**(2)**:** 115-120.

112. Iyer LM, Abhiman S, Aravind L. Natural history of eukaryotic DNA methylation systems. *Progress in molecular biology and translational science* 2011, **101:** 25-104.

113. Lee JT. Epigenetic regulation by long noncoding RNAs. *Science (New York, NY)* 2012, **338**(6113)**:** 1435-1439.

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

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

116. Cremer T, Cremer C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews Genetics* 2001, **2**(4)**:** 292-301.

117. Reddy KL, Zullo JM, Bertolino E, Singh H. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature* 2008, **452**(7184)**:** 243-247.

118. Kosak ST, Skok JA, Medina KL, Riblet R, Le Beau MM, Fisher AG, *et al.* Subnuclear Compartmentalization of Immunoglobulin Loci During Lymphocyte Development. *Science (New York, NY)* 2002, **296**(5565)**:** 158-162.

119. Kumaran RI, Thakar R, Spector DL. Chromatin Dynamics and Gene Positioning. *Cell* 2008, **132**(6)**:** 929-934.

120. Meaburn KJ, Cabuy E, Bonne G, Levy N, Morris GE, Novelli G, *et al.* Primary laminopathy fibroblasts display altered genome organization and apoptosis. *Aging cell* 2007, **6**(2)**:** 139-153.

121. Amendola M, van Steensel B. Mechanisms and dynamics of nuclear lamina–genome interactions. *Current Opinion in Cell Biology* 2014, **28**(0)**:** 61-68.

122. Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, Chess A, *et al.* An Architectural Role for a Nuclear

Noncoding RNA: NEAT1 RNA Is Essential for the Structure of Paraspeckles. *Molecular cell* 2009, **33**(6)**:** 717-726.

123. Yao H, Brick K, Evrard Y, Xiao T, Camerini-Otero RD, Felsenfeld G. Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. *Genes & development* 2010, **24**(22)**:** 2543-2555.

124. Lai F, Orom UA, Cesaroni M, Beringer M, Taatjes DJ, Blobel GA, *et al.* Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* 2013, **494**(7438)**:** 497-501.

125. Xiang JF, Yin QF, Chen T, Zhang Y, Zhang XO, Wu Z, *et al.* Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. *Cell research* 2014, **24**(5)**:** 513-531.

126. Hacisuleyman E, Goff LA, Trapnell C, Williams A, Henao-Mejia J, Sun L, *et al.* Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat Struct Mol Biol* 2014, **21**(2)**:** 198-206.

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

128. Krivega I, Dean A. Enhancer and promoter interactions-long distance calls. *Current opinion in genetics & development* 2012, **22**(2)**:** 79-85.

129. Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 2010, **467**(7314)**:** 430-435.

130. Quinodoz S, Guttman M. Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. *Trends in Cell Biology* 2014, **24**(11)**:** 651-663.

131. Davidovich C, Zheng L, Goodrich KJ, Cech TR. Promiscuous RNA binding by Polycomb repressive complex 2. *Nat Struct Mol Biol* 2013, **20**(11)**:** 1250-1257.

132. Kaneko S, Son J, Shen SS, Reinberg D, Bonasio R. PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat Struct Mol Biol* 2013, **20**(11)**:** 1258-1264.

133. Kaneko S, Son J, Bonasio R, Shen SS, Reinberg D. Nascent RNA interaction keeps PRC2 activity poised and in check. *Genes & development* 2014.

134. Cifuentes-Rojas C, Hernandez Alfredo J, Sarma K, Lee Jeannie T. Regulatory Interactions between RNA and Polycomb Repressive Complex 2. *Molecular cell*, **55**(2)**:** 171-185.

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

136. Sarma K, Cifuentes-Rojas C, Ergun A, del Rosario A, Jeon Y, White F, *et al.* ATRX Directs Binding of PRC2 to Xist RNA and Polycomb Targets. *Cell*, **159**(4)**:** 869-883.

137. O'Shea JJ, Paul WE. Mechanisms Underlying Lineage Commitment and Plasticity of Helper CD4+ T Cells. *Science (New York, NY)* 2010, **327**(5969)**:** 1098-1102.

138. Hedrich C, Bream J. Cell type-specific regulation of IL-10 expression in inflammation and disease. *Immunol Res* 2010, **47**(1-3)**:** 185-206.

139. Chen Z, Tato CM, Muul L, Laurence A, O'Shea JJ. Distinct regulation of interleukin-17 in human T helper lymphocytes. *Arthritis & Rheumatism* 2007, **56**(9)**:** 2936-2946.

140. Wilson NJ, Boniface K, Chan JR, McKenzie BS, Blumenschein WM, Mattson JD, *et al.* Development, cytokine

profile and function of human interleukin 17-producing helper T cells. *Nature immunology* 2007, **8**(9)**:** 950-957.

141. Zhang F, Meng G, Strober W. Interactions among the transcription factors Runx1, ROR[gamma]t and Foxp3 regulate the differentiation of interleukin 17-producing T cells. *Nature immunology* 2008, **9**(11)**:** 1297-1306.

142. Klunker S, Chong MMW, Mantel P-Y, Palomares O, Bassin C, Ziegler M, *et al.* Transcription factors RUNX1 and RUNX3 in the induction and suppressive function of Foxp3+ inducible regulatory T cells. *The Journal of experimental medicine* 2009, **206**(12)**:** 2701-2715.

143. Li L, Patsoukis N, Petkova V, Boussiotis VA. Runx1 and Runx3 are involved in the generation and function of highly suppressive IL-17-producing T regulatory cells. *PloS one* 2012, **7**(9)**:** e45115.

144. Chung Y, Tanaka S, Chu F, Nurieva RI, Martinez GJ, Rawal S, *et al.* Follicular regulatory T cells expressing Foxp3 and Bcl-6 suppress germinal center reactions. *Nat Med* 2011, **17**(8)**:** 983-988.

145. Tsuji M, Komatsu N, Kawamoto S, Suzuki K, Kanagawa O, Honjo T, *et al.* Preferential Generation of Follicular B Helper T Cells from Foxp3+ T Cells in Gut Peyer's Patches. *Science (New York, NY)* 2009, **323**(5920)**:** 1488-1492.

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

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

148. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nature reviews Genetics* 2004, **5**(2)**:** 101-113.

149. Zhang B, Arun G, Mao YS, Lazar Z, Hung G, Bhattacharjee G, *et al.* The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult. *Cell reports* 2012, **2**(1)**:** 111-123.

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

151. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, *et al.* An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 2006, **443**(7108)**:** 167-172.

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

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

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

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

156. Seok J, Warren HS, Cuenca AG, Mindrinos MN, Baker HV, Xu W, *et al.* Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**(9)**:** 3507-3512.

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

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

159. Marques AC, Ponting CP. Intergenic lncRNAs and the evolution of gene expression. *Current opinion in genetics & development* 2014, **27**(0)**:** 48-53.

160. Pang KC, Frith MC, Mattick JS. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends in Genetics* 2006, **22**(1)**:** 1-5.

161. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, *et al.* Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS biology* 2009, **7**(5)**:** e1000112.

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

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

164. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* 2011, **147**(7)**:** 1537-1550.

165. He S, Liu S, Zhu H. The sequence, structure and evolutionary features of HOTAIR in mammals. *BMC evolutionary biology* 2011, **11:** 102.

166. Zhang X, Lian Z, Padden C, Gerstein MB, Rozowsky J, Snyder M, *et al.* A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. *Blood* 2009, **113**(11)**:** 2526-2534.

167. Wright PW, Huehn A, Cichocki F, Li H, Sharma N, Dang H, *et al.* Identification of a KIR antisense lncRNA expressed by progenitor cells. *Genes and immunity* 2013, **14**(7)**:** 427-433.

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

169. Sehgal L, Mathur R, Braun FK, Wise JF, Berkova Z, Neelapu S, *et al.* FAS-antisense 1 lncRNA and production of soluble versus membrane Fas in B-cell lymphoma. *Leukemia* 2014.

170. Xing Z, Lin A, Li C, Liang K, Wang S, Liu Y, *et al.* lncRNA Directs Cooperative Epigenetic Regulation Downstream of Chemokine Signals. *Cell* 2014, **159**(5)**:** 1110-1125.

171. Carpenter S, Aiello D, Atianand MK, Ricci EP, Gandhi P, Hall LL, *et al.* A Long Noncoding RNA Mediates Both Activation and Repression of Immune Response Genes. *Science (New York, NY)* 2013, **341**(6147)**:** 789-792.

172. Rapicavoli NA, Qu K, Zhang J, Mikhail M, Laberge R-M, Chang HY. *A mammalian pseudogene lncRNA at the interface of inflammation and anti-inflammatory therapeutics*, vol. 2, 2013.

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

174. Imamura K, Imamachi N, Akizuki G, Kumakura M, Kawaguchi A, Nagata K, *et al.* Long Noncoding RNA NEAT1-Dependent SFPQ Relocation from Promoter Region to Paraspeckle Mediates IL8 Expression upon Immune Stimuli. *Molecular cell* 2014, **53**(3)**:** 393-406.

175. Collier SP, Collins PL, Williams CL, Boothby MR, Aune TM. Influence of Tmevpg1, a long intergenic noncoding RNA, on the expression of Ifng by Th1 cells. *Journal of immunology (Baltimore, Md : 1950)* 2012, **189**(5)**:** 2084-2088.

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

177. Kino T, Hurt DE, Ichijo T, Nader N, Chrousos GP. Noncoding RNA Gas5 Is a Growth Arrest and Starvation-Associated Repressor of the Glucocorticoid Receptor. *Science signaling* 2010, **3**(107)**:** ra8.

178. Pang KC, Dinger ME, Mercer TR, Malquori L, Grimmond SM, Chen W, *et al.* Genome-Wide Identification of Long Noncoding RNAs in CD8+ T Cells. *The Journal of Immunology* 2009, **182**(12)**:** 7738-7748.

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

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

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

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

Paper submitted to Nature Immunology

Abstract

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

Introduction

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

Besides the well-established role of transcription factors as instructive signals for cell differentiation toward a given lineage, other cues, such as epigenetic modifications, can regulate maintenance of cellular states[4]. In this context non-coding RNAs (ncRNAs) are emerging as a new regulatory layer impacting on both the development and the functioning of the immune system[5, 6]. Among the several classes of ncRNAs that play a specific role in lymphocyte biology, microRNAs are

the best-characterized[7, 8, 9, 10, 11, 12]. As to long intergenic non-coding RNAs (lincRNAs), although thousands of them have been identified in the mammalian genome by bioinformatics analyses of transcriptomic data[13, 14], their functional characterization is still largely incomplete. The functional studies performed so far have shown that lincRNAs contribute to the control of cell differentiation and to the maintenance of cell identity through different modes of action[15]. Nuclear lincRNAs act mainly through their association with chromatin-modifying complexes[16, 17, 18]. Whereas, cytoplasmic lincRNAs can modulate translational control[19] and transcripts stability[20] directly by base pairing with specific targets or indirectly as competing endogenous RNAs[21, 22, 23]. Few examples of functional lincRNAs have been recently described in the mouse immune system. A broad analysis performed by interrogating naïve and memory $CD8^+$ cells purified from mouse spleen with a custom array of lincRNAs reported the identification of 96 lymphoid-specific lincRNAs and suggested a role for lincRNAs in lymphocyte differentiation and activation[24]. The lincRNA NeST has been found to be downregulated during lymphocyte activation in a reciprocal manner to IFN-g and to control susceptibility to Theiler's virus and Salmonella infection in mice through epigenetic regulation of the IFN-g locus[25, 26]. More recently, mouse lincRNA-Cox2 has been reported to be induced downstream Toll-like receptor signalling and to mediate the activation and repression of distinct sets of immune target genes involved in inflammatory responses[27]. Another study on mouse thymocytes and mature peripheral T cells allowed the identification of lincRNAs with specific cell expression

pattern during T cell differentiation and of a $CD4^+$ T_H2 specific lincRNA - LincR-Ccr2-5'AS - involved in the regulation of $CD4^+$ T_H2 lymphocytes migration[28]. Although these studies highlight the relevance of lincRNAs in regulating immune responses, a thorough analysis of their expression profile and functional role in the human immune system is still lacking.

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

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

Results

LincRNAs identify human lymphocyte subsets better than protein coding genes

To assess lincRNA expression in human primary lymphocytes, RNA was extracted from thirteen lymphocyte cell subsets (Table 1) purified from peripheral blood mononuclear cells (PBMCs) of five healthy donors[12]. The polyadenylated RNA fraction was then analysed by paired-end RNA sequencing obtaining about 1.7 billion mapped reads. In order to enrich for transcripts deriving from "bona fide" active genes we applied an expression threshold ("0.21" FPKM) defined through the integration of RNAseq and chromatin state ENCODE project data[29]. We found a total of 31,902 expressed genes (including both protein coding and non coding genes) in the 13 subsets (Table 1 and Supplementary Fig. 1a), of which 4,201 were lincRNAs annotated in public resources[13, 30] (Fig. 1a). In order to identify novel lincRNAs expressed in primary human lymphocytes, we used three *de novo* transcriptome reconstruction strategies that are based on the combination of two different sequence mappers, TopHat and Star[31, 32], with two different tools for *de novo* transcripts assembly, Cufflinks and Trinity[33, 34]. LincRNAs were identified within the newly described transcripts exploiting the following process: *i*) selection of transcripts longer than 200 nucleotides and multiexonic, which did not overlap with protein coding genes (thus counting out unreliable single-exon fragments assembled from RNA-seq); *ii*) exclusion of transcripts that contain a conserved protein-coding region

and transcripts with ORFs that contain protein domains catalogued in Pfam protein family database[35]; *iii*) exploitation of PhyloCSF, a comparative genomics method that assesses multispecies nucleotide sequence alignment based on a formal statistical comparison of phylogenetic codon models[36], which efficiently identifies non-coding RNAs as demonstrated by ribosome profiling experiments[37]. Finally we defined a stringent *de novo* lincRNA set including those genes for which at least one lincRNA isoform was reconstructed by two assemblers out of three. Through this conservatively multi-layered analysis we identified 563 novel lincRNAs genes, increasing by 11.8% the number of lincRNAs expressed in human lymphocytes. The different classes of RNAs are evenly distributed among different lymphocytes subsets (Supplementary Fig. 1b) and the ratio of already annotated and newly identified lincRNAs is similar across different chromosomes (Supplementary Fig. 1c) and across various lymphocyte subsets (Supplementary Fig. 1d). As previously observed in different cell types[13, 33], also in human lymphocytes lincRNAs are generally expressed at lower levels than protein coding genes (Supplementary Fig. 1e). However, when transcripts were divided based on their expression in cell-specific and non specific (Supplementary Fig. 1f), we found that cell specific lincRNAs and cell specific protein coding genes, display similar expression levels (Supplementary Fig. 1e-g).

Lymphocytes subsets display very different migratory abilities and effector functions, yet they are very closely related from the differentiation point of view. As lincRNAs are generally more tissue

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

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

Identification of lincRNA expression signatures of human lymphocyte subsets

Next, we interrogated our dataset for the presence of lincRNAs signatures in the different lymphocyte subsets. We therefore looked for lincRNAs differentially expressed ($p<0.05$; non-parametric Kruskal-

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

It is important to note that, the newly identified lincRNAs defined as signatures are more expressed (Fig. 2c) and more cell-specific (Supplementary Fig. 2b-m) than the already annotated lincRNAs defined as signatures. The representative data in Fig. 2b refer to the $CD4^+$ T_H1 cell subset; similar results were obtained for all the other subsets (Supplementary Fig. 2b-m).

Finally, to confirm and extend our signature data, we assessed the expression of $CD4^+$ T_H1 lincRNAs by RT-qPCR in a new set of independent samples of primary human $CD4^+$ naïve, T_{reg} and T_H1 cells, as well as in naïve $CD4^+$ T cells that were activated *in vitro* and induced to

differentiate toward T_H1 or T_H2 cells. Specific subset expression was confirmed for 90% of the $CD4^+$ T_H1 signature lincRNAs (Fig. 2d). Moreover, 90% of $CD4^+$ T_H1 signature lincRNAs that are expressed in resting $CD4^+$ T_H1 cells purified *ex vivo*, are highly expressed also in naïve $CD4^+$ T cells differentiated under T_H1 polarizing conditions *in vitro*, whereas they are poorly expressed in naïve $CD4^+$ T cells that are differentiated towards T_H2 *in vitro* (Fig. 2e). As a corollary to these findings, we observed by RNA-seq that $CD4^+$ naïve signature lincRNAs are mostly down-regulated during differentiation towards T_H0 cells *in vitro*, when T_H1, T_H2 and T_H17 signature lincRNAs are mostly up-regulated (Supplementary Fig. 2a).

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

Linc-MAF-4 downregulation skews $CD4^+$ T cell differentiation towards T_H2

As lincRNAs have been reported to influence the expression of neighbouring genes[25, 26, 28, 39], we asked whether protein coding genes proximal to lymphocytes signature lincRNAs were involved in key cell-functions. To this purpose we used the FatiGO tool from the Babelomics suite for functional enrichment analysis[40] and found that protein coding genes neighbouring to signature lincRNAs are enriched for Gene

Ontology terms strongly correlated with lymphocyte T cell activation (Fig. 3a), pointing to a possible role of signature lincRNAs in important lymphocyte functions. In order to obtain proof of concept of this hypothesis, we chose to characterize in depth linc-MAF-4 (also referred to as linc-MAF-2 in LNCipedia database http://www.lncipedia.org[41]), a T_H1 signature lincRNA, localized 139.5 Kb upstream of the *MAF* gene. *MAF* encodes a transcription factor involved in T_H2 differentiation[42], which is also required for the efficient development of T_H17 cells[43] and controls IL4 transcription in $CD4^+$ T follicular helper cells[44]. Our sequencing data showed that high expression of linc-MAF-4 correlates with low levels of *MAF* transcript in $CD4^+$ T_H1 cells, conversely T_H2 cells have low expression levels of linc-MAF-4 and high levels of MAF transcript. The anti-correlation of expression between lincRNAs and their neighbouring genes is not a common feature of all lincRNAs ([13, 16]), and it is probably restricted to a limited number of cis-acting lincRNAs. This observation is confirmed also in our dataset (data not shown). Moreover, no correlation is observed between the expression linc-MAF-4 and its proximal upstream protein coding genes: CDYL2 and DYNLRB2 (Supplementary Fig. 3a).

The same inverse relation between linc-MAF-4 and MAF is observed when naïve $CD4^+$ T cells are differentiated *in vitro* towards T_H1 or T_H2 cells. In details, Fig. 3b shows that in T lymphocytes differentiating towards T_H1 cells, MAF transcript increases up to day 3 and then drops. Conversely, linc-MAF-4 is poorly expressed for the first three days but then increases progressively. In $CD4^+$ T lymphocytes

differentiating towards T_H2 cells, we found the opposite situation, both MAF transcript and protein levels increase constantly up to day 8 while Iinc-MAF4 remains constantly low (Fig. 3b and Supplementary Fig. 3c), similarly to what observed in $CD4^+$ T lymphocytes differentiating towards T_H17 cells (Supplementary Fig. 3d).

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

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

expression of *GATA3* and *IL4*, two genes characteristic of T_H2 cells, is increased after linc-MAF-4 knock-down (Fig. 3g and Supplementary Fig.3e).

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

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

Since *linc-MAF-4* gene maps in relative proximity (139.5 Kb) to *MAF* gene we asked whether linc-MAF-4 can down-regulate *MAF* transcription, and, we investigated whether their genomic regions could physically interact. Chromosome conformation capture (3C) analysis was exploited to determine relative crosslinking frequencies among regions of interest. We tested the conformation of the *linc-MAF-4* - *MAF* genomic region in differentiated $CD4^+$ T_H1 cells. A common reverse primer mapping within the *MAF* promoter region, was used in combination with a set of primers spanning the locus, and interactions were analysed by PCR. Specific interactions between *MAF* promoter and 5' and 3' end regions of *linc-MAF-4* were detected (Fig. 4a,b and Supplementary Fig. 4a), indicating the existence of an *in cis* chromatin looping conformation that brings *linc-MAF-4* in close proximity to *MAF* promoter. Interestingly, the subcellular fractionation of *in vitro* differentiated $CD4^+$ T_H1 lymphocytes revealed a strong enrichment of linc-MAF-4 in the chromatin fraction (Fig. 4c). Because other chromatin-associated lincRNAs regulate neighbouring genes by recruiting specific chromatin

remodellers, we tested in RNA immunoprecipitation (RIP) assays the interaction of linc-MAF-4 with different chromatin modifiers, including activators and repressors (data not shown), and found a specific enrichment of linc-MAF-4 in the immunoprecipitates of two repressors, EZH2 and LSD1 (Fig. 4d and Supplementary Fig. 4b). In agreement with these findings, we found that linc-MAF-4 knock-down in activated $CD4^+$ naïve T cells reduces both EZH2 and LSD1 levels and correlates with the reduction of EZH2 enzymatic activity at *MAF* promoter as demonstrated by the H3K27me3 reduction at this locus (Fig. 4e). Remarkably, H3K27me3 levels were reduced neither at *MYOD1* promoter region (a known target of EZH2) nor at a region within the chromatin loop between *linc-MAF-4* and *MAF* marked by H3K27me3 (Supplementary Fig. 4c).

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

Discussion

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

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

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

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

The lincRNAs role in differentiation has been described in different cell types[17, 20, 23, 46, 47]. In the mouse immune system it has been found that lincRNAs expression changes during naïve to memory $CD8^+$ T cell differentiation[24] and during naïve $CD4^+$ T cells differentiation into

distinct helper T cell lineages[28]. We show in human primary lymphocytes that activation induced differentiation of $CD4^+$ naïve T cells is associated with increased expression of lincRNAs belonging to the $CD4^+$ T_H1 signature suggesting that upregulation of T_H1 lincRNAs is part of the cell differentiation transcriptional program. Indeed, linc-MAF-4, one of the T_H1 signature lincRNA, is poorly expressed in T_H2 cells and its experimental downregulation skews differentiating T helper cells toward a T_H2 transcription profile. We have found that linc-MAF-4 regulates transcription exploiting a chromatin loop that brings its genomic region close to the promoter of *MAF* gene. We propose that the chromatin organization of this region allows linc-MAF-4 transcript to recruit both EZH2 and LSD1 and modulate the enzymatic activity of EZH2 negatively regulating *MAF* transcription with a mechanism of action similar to that shown for the lincRNAs HOTAIR[48] and MEG3 [49]. We therefore provide a mechanistic proof of concept that lincRNAs can be important regulators of $CD4^+$ T-cell differentiation. Given the number of specific lincRNAs expressed in the different lymphocytes subsets, it can be postulated that many other lincRNAs might contribute to cell differentiation and to the definition of cell identity in human lymphocytes.

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

Online Methods

Purification of primary immunological cell subsets

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

then washed two times, resuspended in FACS washing buffer and analysed by flow cytometry.

RNA isolation and RNA sequencing

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

RNA-seq and publicly available datasets

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

Reference annotation

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

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

A comprehensive catalogue of lincRNAs specifically expressed in human lymphocyte subsets was generated using a *de novo* genome-based transcripts reconstruction procedure with three different approaches. Two aligners were used: TopHat v.1.4.1 and STAR v. 2.2.0. The *de novo* transcriptome assembly was performed on the aligned sequences (samples of the same population were concatenated into one "population alignment") generated by STAR and TopHat using Cufflinks v. 2.1.1 with reference annotation to guide the assembly (-g option) coupled with multi-read (-u option) and fragment bias correction (-b option) to improve

the accuracy of transcripts abundance estimates. With this method, about 30,000-50,000 new transcripts were identified in each lymphocyte population. The third approach employed the genome-guided Trinity software (http://pasa.sourceforge.net/#A_ComprehensiveTranscriptome), which generates novel transcripts performing a local assembly on previously mapped reads from specific location. The Trinity[50] default aligner was substituted with STAR. Each candidate transcript was then processed using the PASA pipeline, which reconstructs the complete transcript and gene structures, resolving incongruences derived from transcript misalignments and alternatively splices events, refining the reference annotation when there are enough evidences and proposing new transcripts and genes in case no previous annotation can explain the new data.

Novel lincRNA genes identification

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

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

De novo data integration

Duplicates among the transcripts identified with the same *de novo* method were resolved using Cuffcompare v2.1.1. In the same way, the resulting three datasets were further merged to generate a non-redundant atlas of lincRNAs in human lymphocytes and only genes identified by at least 2

out of 3 software were considered. A unique name was given to each newly identified lincRNA gene composed by the prefix “linc-” followed by the Ensembl gene name of the nearest protein coding gene (irrespective of the strand). The additional designation “up” or “down” defines the location of the lincRNA with respect to the sense of transcription of the nearest protein coding gene. In addition, either “sense” or “antisense” was added to describe the concordance of transcription between the lincRNA and its nearest coding gene. A numerical counter only of newly identified lincRNAs related to the same protein coding gene is added as suffix (such as ‘linc-geneX-(up|down)-(sense|antisense)_#n’). This final non-redundant catalogue of newly identified lincRNAs includes 4,666 new transcripts referring to 3,005 new genes.

LincRNA signatures definition

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

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

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

different lymphocyte populations: a p-value lower than 0.05 was considered and the lincRNA genes that meet these selection criteria were selected as signature genes.

Gene Ontology Enrichment Analysis

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

Naïve CD4$^+$ T cells siRNA transfection

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

Gene Expression Analysis

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

GSEA (Gene Set Enrichment Analysis)

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

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

RT-qPCR Analysis

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

amplification reaction. The primers used in qPCR are listed in Supplementary Table 2.

Cell fractionation

In vitro differentiated T_H1 cells were resuspended in RLN1 buffer (50 mM Tris-HCl pH 8, 140 mM NaCl; 1.5 mM $MgCl_2$, 0.5% NP-40) supplemented with SUPERase In (Ambion) for 10 minutes on ice. After a centrifugation at 300g for 2 minutes, the supernatant was collected as the cytoplasmic fraction. The pellet was resuspended in RLN2 buffer (50 mM Tris-HCl pH 8, 500 mM NaCl, 1.5 mM $MgCl_2$, 0.5% NP-40) supplemented with RNase inhibitors for 10 minutes on ice. Chromatin was pelletted at maximum speed for 3 minutes. The supernatant represents the nuclear fraction. All the fractions were resuspended in TRIzol (Ambion) to 1 ml and RNA was extracted following the standard protocol.

RNA immunoprecipitation (RIP)

In vitro differentiated T_H1 cells were UV-crosslinked at 400 mJ/cm^2 in ice-cold D-PBS and then pelleted at 1350 g for 5 minutes. The pellet was resuspended in ice-cold lysis buffer (25 mM Tris-HCl, 150 mM NaCl, 0.5% NP-40) supplemented with 0.5 mM β-mercaptoethanol, Protease Inhibitor Cocktail Tablets cOmplete, EDTA-free (Roche) and SUPERase In (Ambion) and left rocking at 4°C until the lysis is complete. Debris was centrifuged at 13000 g for 10'. The lysate was precleared with Dynabeads® Protein G (Novex®) for 30 minutes at 4°C

and then incubated for 2 hours with 7 mg of antibodies specific for EZH2 (Active Motif - 39875); LSD1 (Abcam – ab17721), or HA (Santa Cruz) as mock control. The lysate was coupled with Dynabeads® Protein G (Novex®) for 1 hour at 4°C. Immunoprecipitates were washed for five times with lysis buffer. RNA was then extracted following mirVana miRNA Isolation Kit (Ambion) protocol. Levels of Linc-MAF-4 or of the negative controls b-actin, RNU2.1 and a region upstream the TSS of linc-MAF-4 (linc-MAF-4 control) were assed by RT-qPCR.

Chromatin Immunoprecipitation analysis (ChIP)

In vitro differentiated T_H1 and T_H2 cells were crosslinked in their medium with 1/10 of fresh formaldehyde solution (50 mM Hepes-KOH pH 7.5, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 11% formaldehyde) for 12 minutes. Then they were treated with 1/10 of 1.25 M glycine for 5 minutes and centrifuged at 1350 g for 5 minutes at 4°C. Cell membranes were lysated in LB1 (50 mM Hepes-KOH pH 7.5, 10 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40 and 0.25% Triton X-100) supplemented with Protease Inhibitor Cocktail Tablets cOmplete, EDTA-free (Roche) and Phenylmethanesulfonyl fluoride (Sigma) at 4°C. Nuclei were pelletted at 1350 g for 5 minutes at 4°C and washed in LB2 (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA) supplemented protease inhibitors. Nuclei were again pelleted at 1350 g for 5 minutes at 4°C and resuspended with a syringe in 200 μl LB3 (10 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Na-deoxycholate, 0.5% N-lauroylscarcosine) supplemented with

protease inhibitors. Cell debris were pelleted at 20000 g for 10 minutes at 4°C and a ChIP was set up in LB3 supplemented with 1% Triton X-100, protease inhibitors and antibodies against H3K4me3, H3K27me3 (Millipore), RNA polymerase II STD repeat YSPTSPS, LSD1 (Abcam), EZH2 (Active Motif) or no antibody (as negative control) o/n at 4°C. The day after Dynabeads® Protein G (Novex®) were added at left at 4°C rocking for 2 hours. Then the beads were washed twice with Low salt wash buffer (0.1% SDS, 2 mM EDTA, 1% Triton X-100, 20 mM Tris-HCl pH 8.0, 150 mM NaCl) and with High salt wash buffer (0.1% SDS, 2 mM EDTA, 1% Triton X-100, 20 mM Tris-HCl pH 8.0, 500 mM NaCl). Histones IPs were also washed with a LiCl solution (250 mM LiCl, 1% NP-40, 1 mM EDTA, 10 mM Tris-HCl pH 8.0). All samples were finally washed with 50 mM NaCl in 1X TE. Elution was performed o/n at 65°C in 50 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% SDS. Samples were treated with 0.02 μg/μl RNase A (Sigma) for 2 hours at 37 °C and with 0.04 μg/μl proteinase K (Sigma) for 2 hours at 55°C. DNA was purified with phenol/chloroform extraction.

Chromosome Conformation Capture (3C)

For 3C analysis cells were crosslinked and digested as described for ChIP[53]. Nuclei were resuspended in 500 μl of 1.2X NEB3 buffer (New England BioLabs) with 0.3% SDS and incubated at 37°C for 1h and then with 2% Triton X-100 for another 1h. Digestion was performed with 800U of BglII (New England BioLabs) o/n at 37°C shaking. Digestion was checked loading digested and undigested controls on a 0.6% agarose

gel. Then the sample was incubated with 1.6% SDS for 25 minutes at 65°C and with 1.15X ligation buffer (New England BioLabs) and 1% Triton X-100 for 1 hour at 37°C. Ligation was performed with 1000U of T4 DNA ligase (New England BioLabs) for 8 hours at 16°C and at room temperature for 30 minutes. DNA was purified with phenol-chloroform extraction after RNase A (Sigma) and Proteinase K (Sigma) digestion. As controls, BACs corresponding to the region of interested were digested with 100U BglII in NEB3 buffer in 50 μl o/n at 37°C. Then fragments were ligated with 400U T4 DNA ligase o/n at room temperature in 40 μl. PCR products amplified with GoTaq Flexi (Promega) for BACs and samples were run on 2.5% agarose gels and quantified with ImageJ software. Primers are listed in Supplementary Table 3.

Accession numbers

ArrayExpress accession: E-MTAB-2319

Reviewer account: Username: Reviewer_E-MTAB-2319

Password: ppkieb1o

Author contribution

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

Acknowledgments

We would like to thank C. Cheroni for support in statistical analysis; M. Moro and MC. Crosti for technical assistance with cell sorting; D. Gabellini, S.Biffo, P. Della Bona and A. Lanzavecchia for discussions and critical revision of the manuscript; B. J. Haas and A. Dobin for helping the integration of genome guided Trinity with STAR aligner. The INGM Bioinformatic facility for support. Google Summer of Code Project for supporting Clayton Wheeler in the development of https://github.com/csw/bioruby-maf.

This study was supported by: the Flagship CNR-MIUR grant “EPIGEN”, CARIPLO grant n° 2013-0955, AIRC grant n° IG2013-ID14596, ERC Advanced Grant n° 269022 to S.A, ERC Consolidator Grant n° 617978 to M.P, and by an unrestricted grant of the “Fondazione Romeo ed Enrica Invernizzi”.

Figure and Table Legends

Table 1. Purification and RNA-sequencing of human primary lymphocyte subsets

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

Figure 1. Identification of lincRNAs expressed in human lymphocyte subsets

(a) RNA-seq data generated from 63 lymphocyte samples were processed according to two different strategies: quantification of lincRNAs already annotated in public resources and *de novo* Genome Based Transcripts Reconstruction for the quantification of new lincRNAs expressed in human lymphocytes. Three methods for the identification of new transcripts were adopted: Reference Annotation Based assembly by Cufflinks with two different aligners (TopHat and STAR) and an approach that integrates Trinity and PASA software. Only transcripts reconstructed by at least two assemblers were considered. Novel transcripts were filtered with a computational analysis pipeline to select for lincRNAs. The number of lincRNA genes and transcripts identified in lymphocytes subsets is indicated.

(b) Expression profiles of lincRNA and protein coding genes across 13 human lymphocyte subsets according to K-Means clusters definition. The black line represents the mean expression of the genes belonging to the same cluster. The peaks of expression profiles refer to the populations reported in legend according to numbering.

(c) Specificity of lincRNAs and protein coding genes. Rows and columns are ordered based on a K-Means clustering of lincRNAs and protein coding genes across 13 human lymphocyte populations. Colour intensity represents the Z-score log_2-normalized raw FPKM counts estimated by Cufflinks. 79% of lincRNAs genes and 39% of protein coding genes are assigned to specific clusters. See also Supplementary Fig. 1h.

(d) As in (c), performed on receptors and metabolic processes genes.

Figure 2. Definition of lincRNA signatures in human lymphocyte subsets

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

(b) $CD4^+$ T_H1 signature lincRNAs extracted from panel (A). The barcode on the left indicates already annotated lincRNAs (white) and newly

described lincRNAs (brick red). For newly described lincRNAs name, 'S' and 'AS' indicates 'sense' and 'antisense' respectively.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

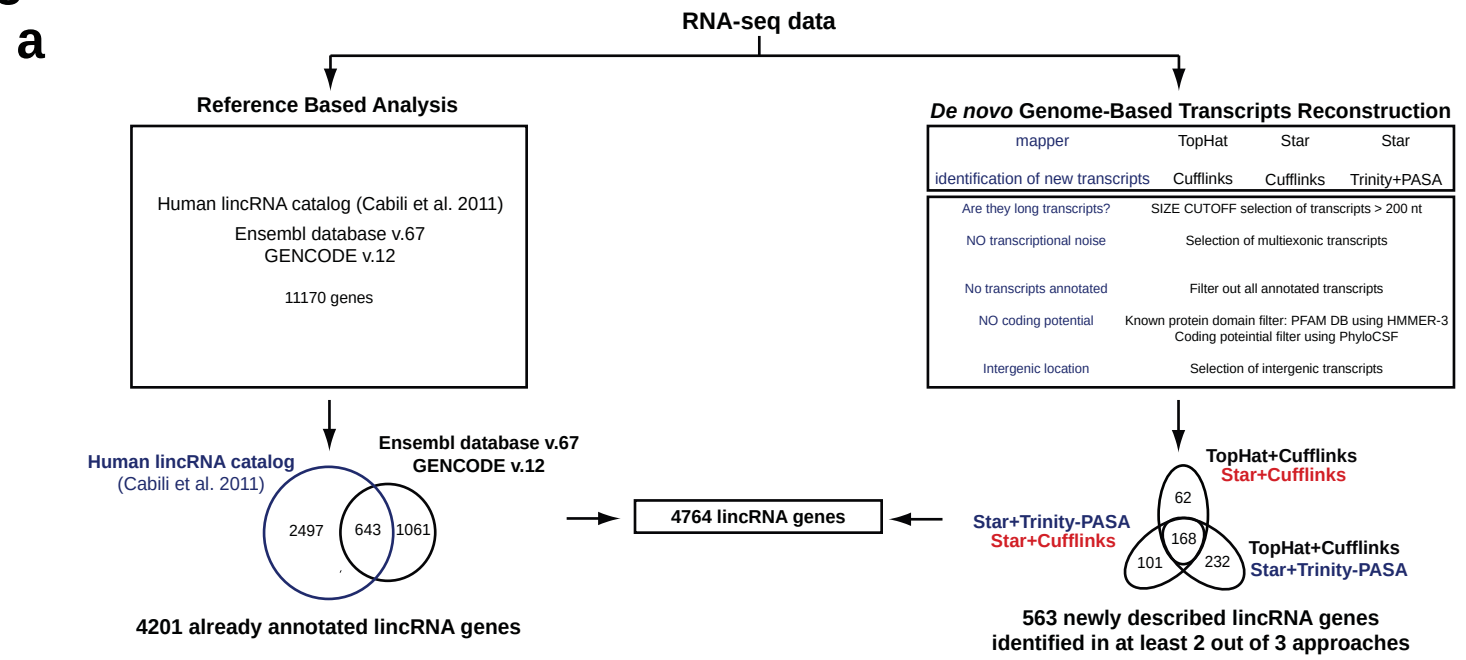
(d) RIP assay for LSD1 and EZH2 in $CD4^+$ naïve T cells differentiated in T_H1 polarizing conditions (day 8). The enrichment of linc-MAF-4 is relative to mock. β-actin, RNU2.1 and a region upstream the TSS of linc-MAF-4 were chosen as controls (average of six independent experiments ± SEM). The statistical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

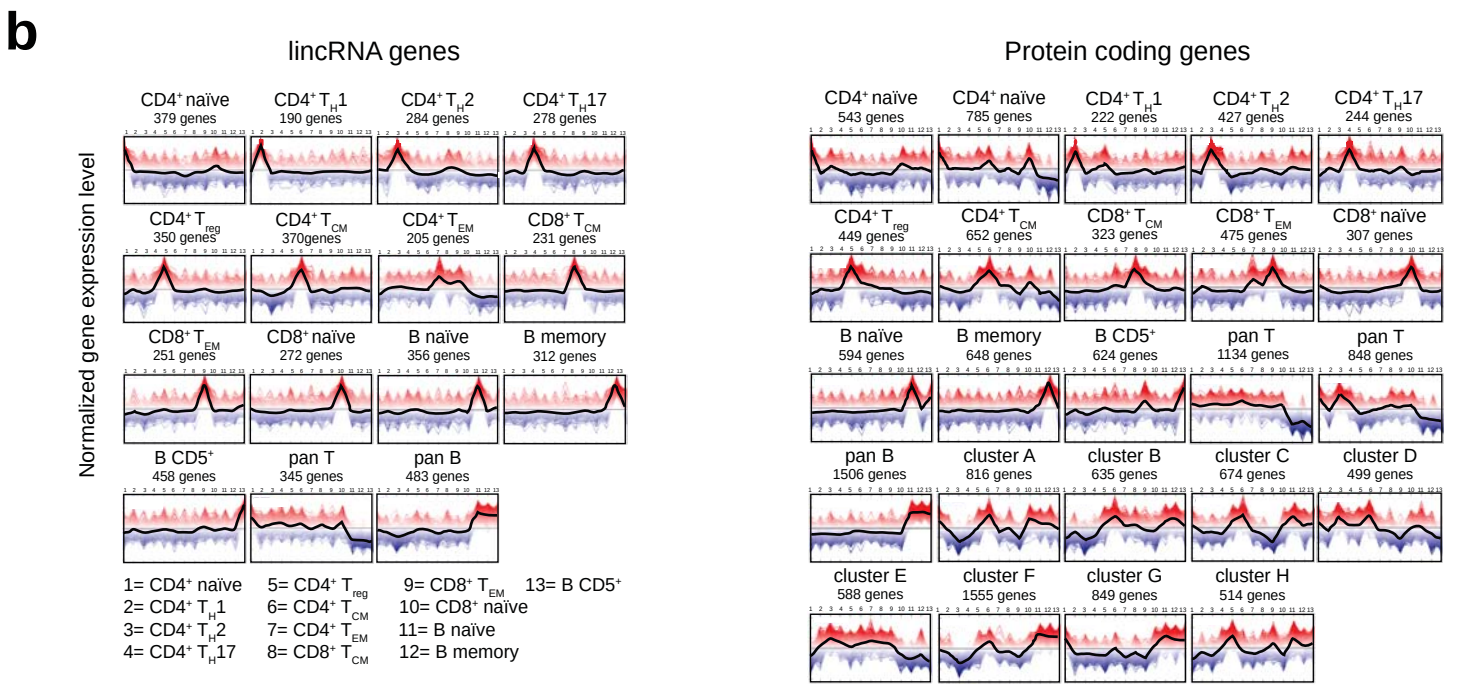
(e) ChIP-qPCR analysis of EZH2, H3K27me3 and LSD1 occupancy at *MAF* locus in activated $CD4^+$ naïve T cells transfected with linc-MAF-4

siRNA (black) or ctrl siRNA (white) (average of at least three independent experiments ± SEM). One-tailed t-test * $p < 0.05$.

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

Figure 1


a
RNA-seq data
Reference Based Analysis
Human lincRNA catalog (Cabili et al. 2011)
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
(Cabili et al. 2011)
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
TopHat+Cufflinks
Star+Trinity-PASA
62
168
101
232
563 newly described lincRNA genes
identified in at least 2 out of 3 approaches


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

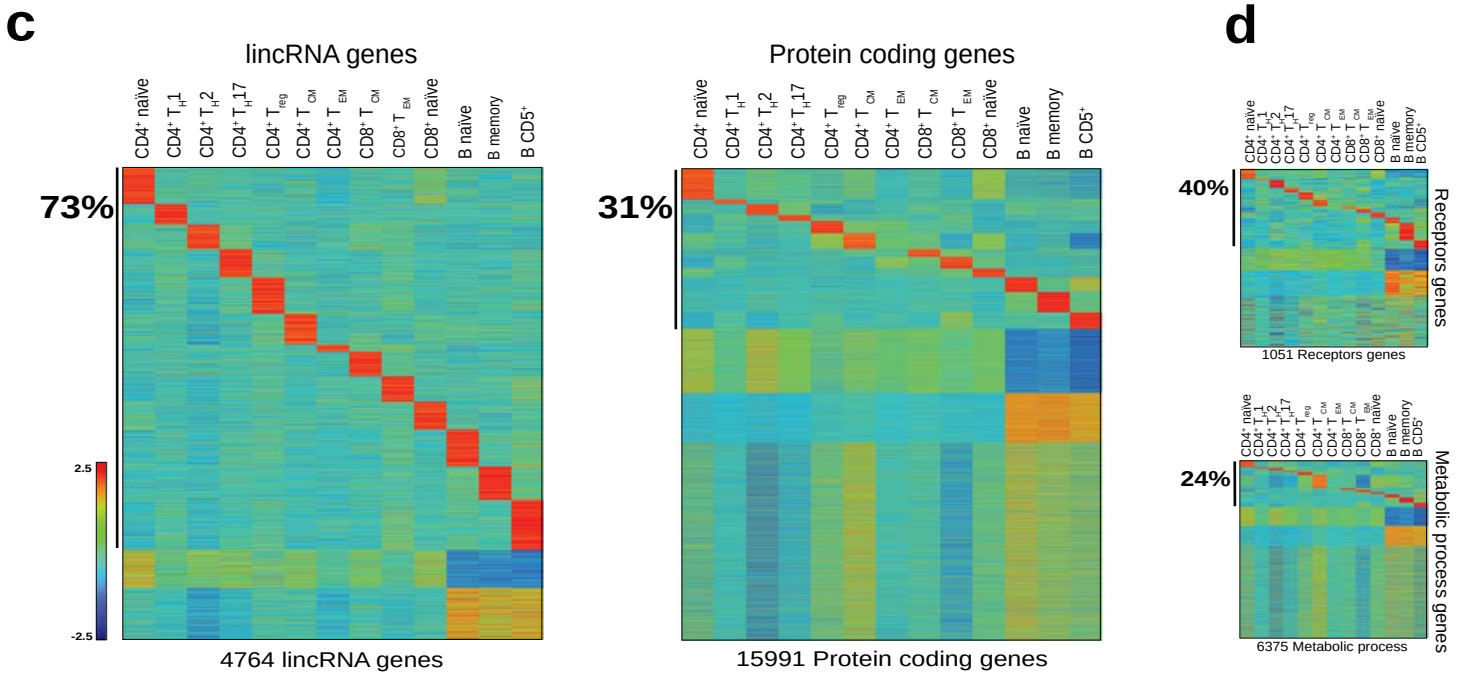

c
lincRNA genes
Protein coding genes
d
73%
31%
40%
24%
2.5
-2.5
4764 lincRNA genes
15991 Protein coding genes
1051 Receptors genes
6375 Metabolic process
Receptors genes
Metabolic process genes

Figure 2

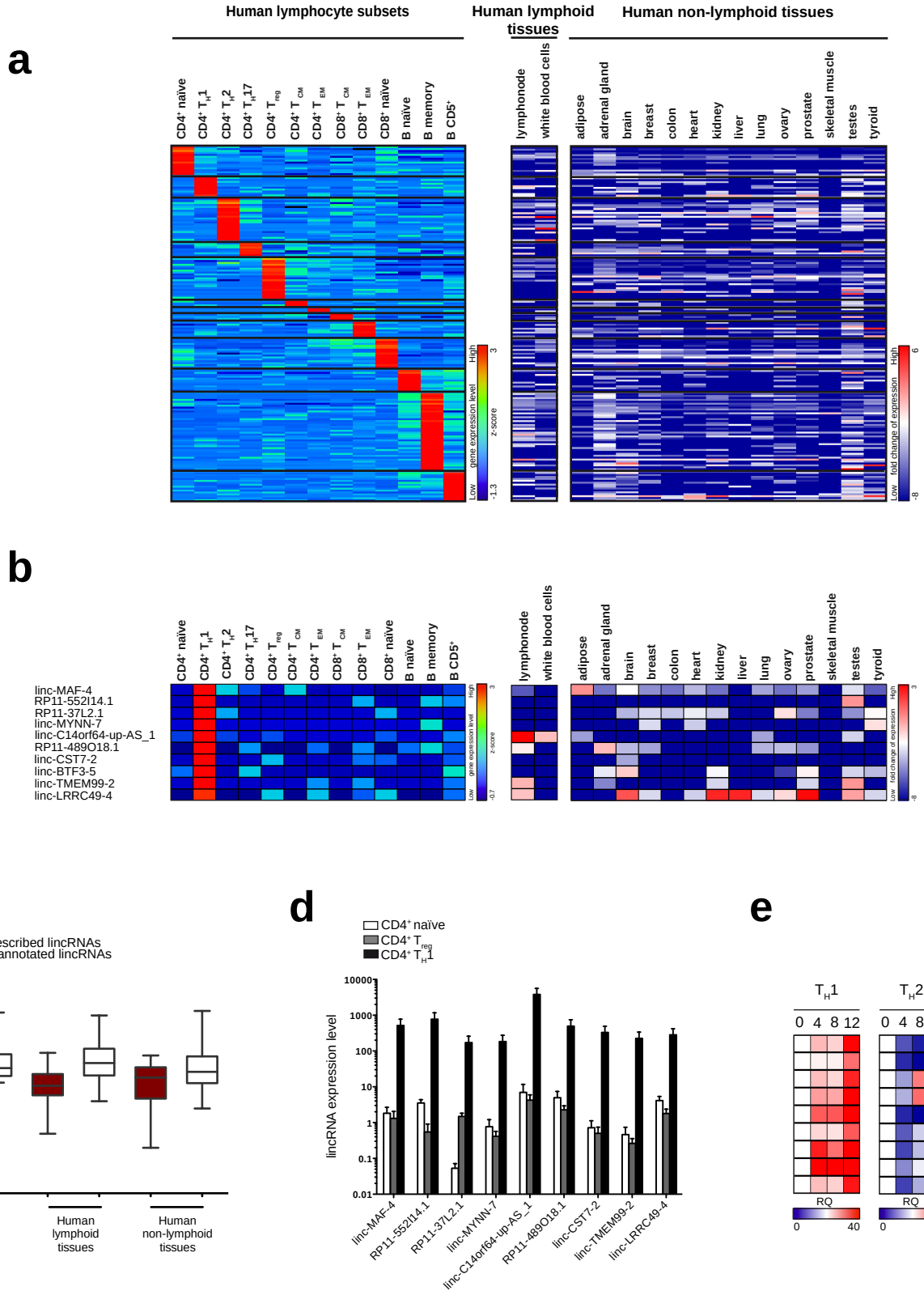

a
Human lymphocyte subsets
Human lymphoid tissues
Human non-lymphoid tissues
CD4+ naïve
CD4+ TH1
CD4+ TH2
CD4+ TH17
CD4+ Treg
CD4+ T CM
CD4+ T EM
CD8+ T CM
CD8+ T EM
CD8+ naïve
B naïve
B memory
B CD5+
lymphonode
white blood cells
adipose
adrenal gland
brain
breast
colon
heart
kidney
liver
lung
ovary
prostate
skeletal muscle
testes
tyroid
High
gene expression level
z-score
Low
3
-1.3
fold change of expression
6
-8
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
-0.7
c
newly described lincRNAs
already annotated lincRNAs
Gene expression level
(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 level
10000
1000
100
10
1
0.1
0.01
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
0 4 8 12
days
RQ
0
40
2

Figure 3

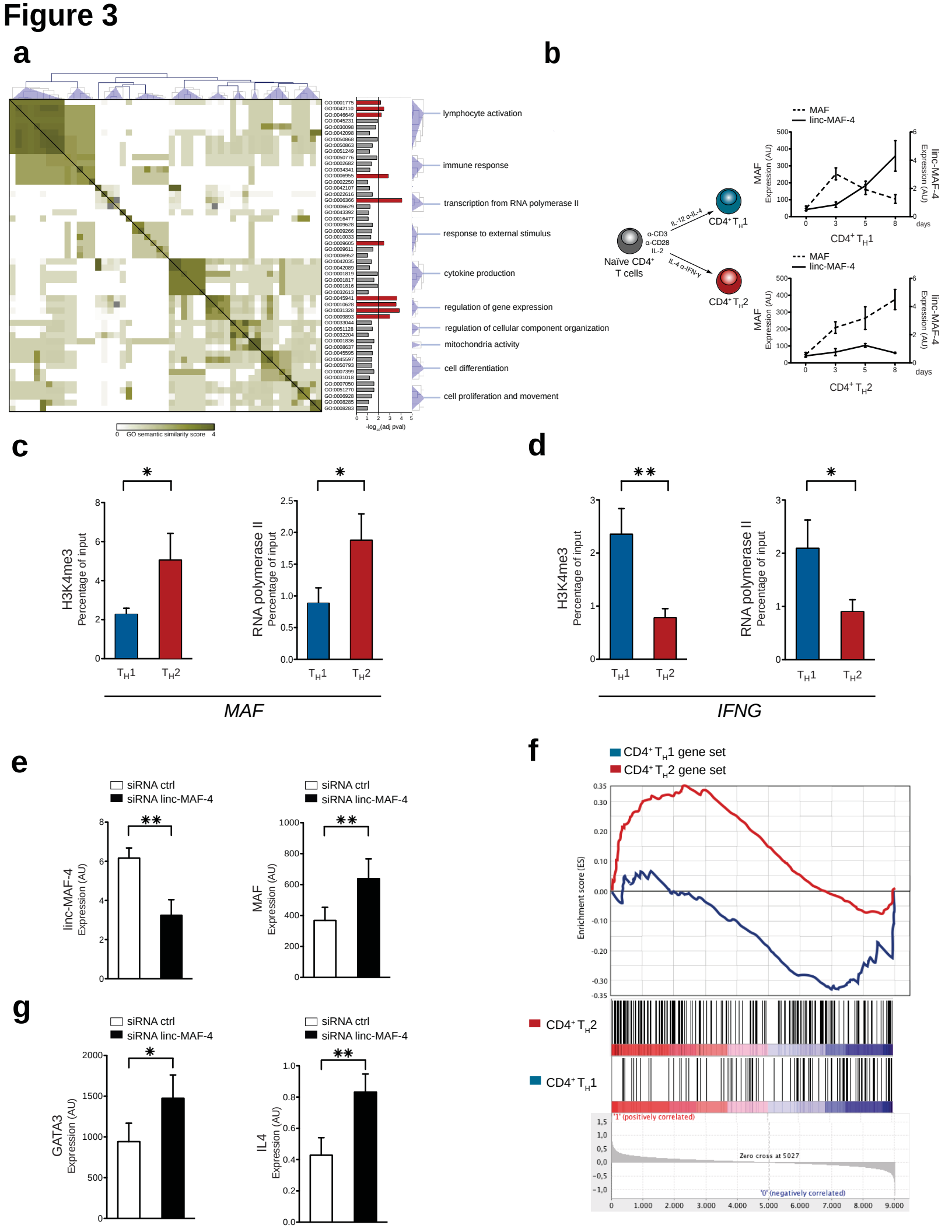
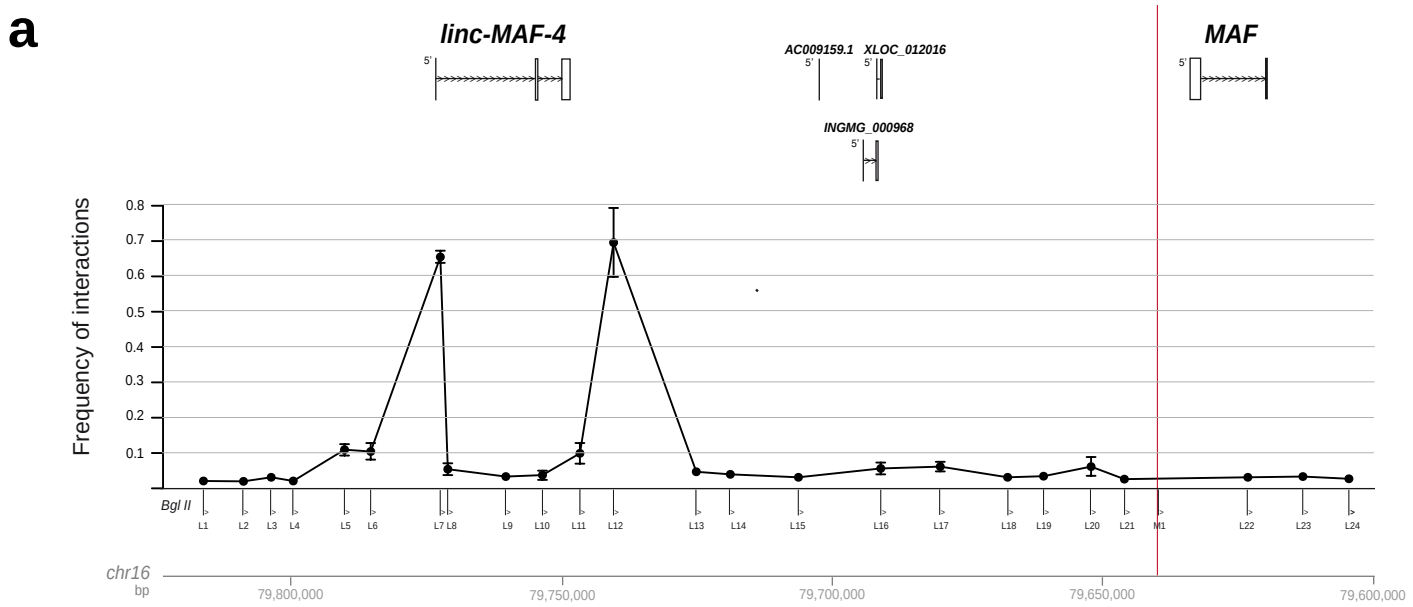
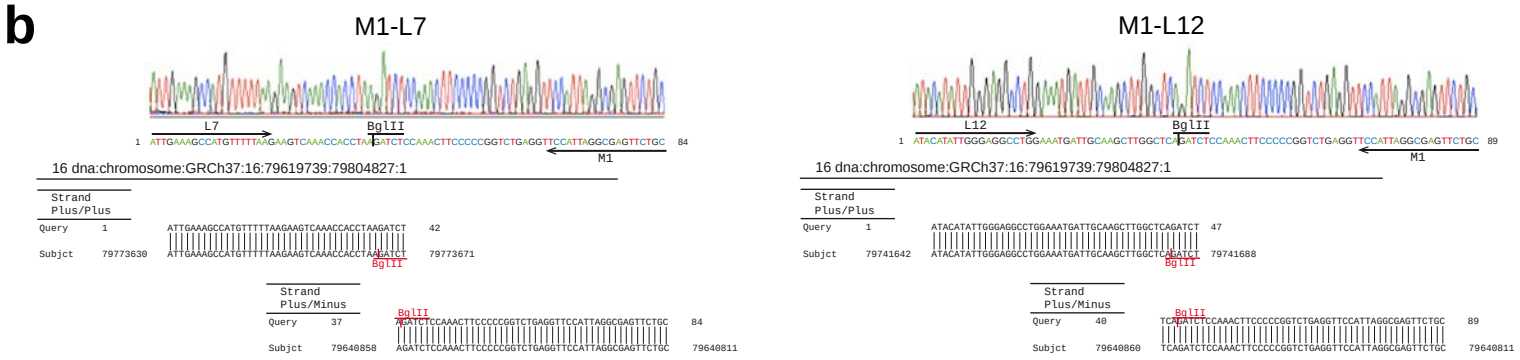
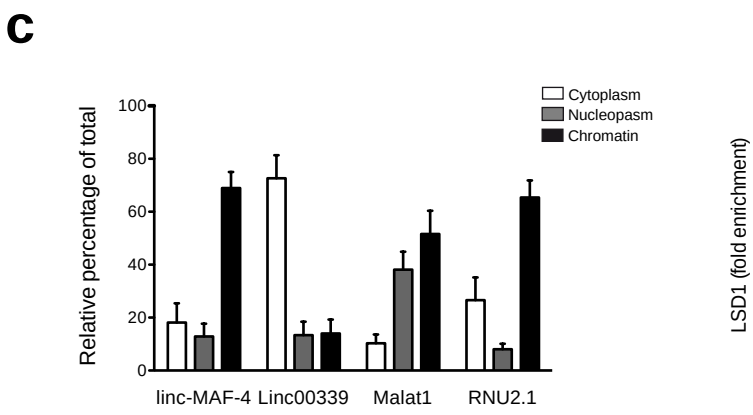
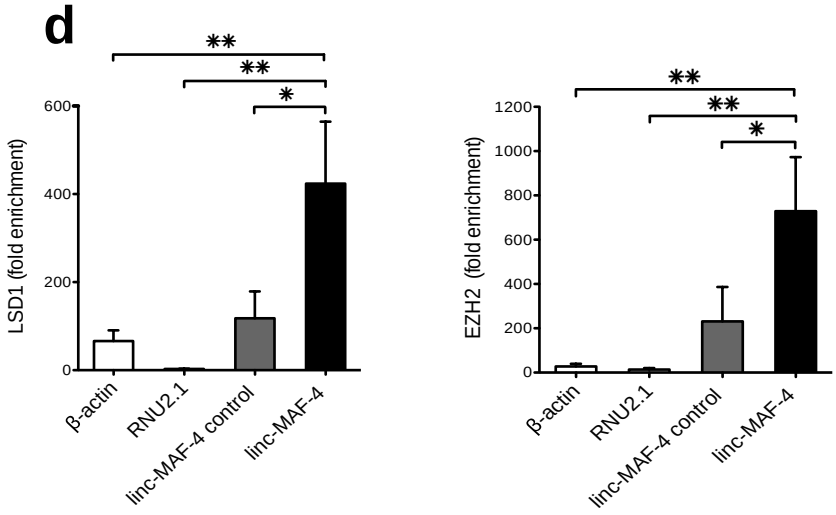

a
lymphocyte activation
immune response
transcription from RNA polymerase II
response to external stimulus
cytokine production
regulation of gene expression
regulation of cellular component organization
mitochondria activity
cell differentiation
cell proliferation and movement
GO semantic similarity score
-log10(adj pval)
b
Naïve CD4+ T cells
α-CD3
α-CD28
IL-2
IL-12 α-IL-4
IL-4 α-IFN-γ
CD4+ TH1
CD4+ TH2
MAF
linc-MAF-4
MAF Expression (AU)
linc-MAF-4 Expression (AU)
days
c
H3K4me3 Percentage of input
RNA polymerase II Percentage of input
TH1
TH2
MAF
d
H3K4me3 Percentage of input
RNA polymerase II Percentage of input
TH1
TH2
IFNG
e
siRNA ctrl
siRNA linc-MAF-4
linc-MAF-4 Expression (AU)
MAF Expression (AU)
f
CD4+ TH1 gene set
CD4+ TH2 gene set
Enrichment score (ES)
CD4+ TH2
CD4+ TH1
'1' (positively correlated)
Zero cross at 5027
'0' (negatively correlated)
g
siRNA ctrl
siRNA linc-MAF-4
GATA3 Expression (AU)
IL4 Expression (AU)

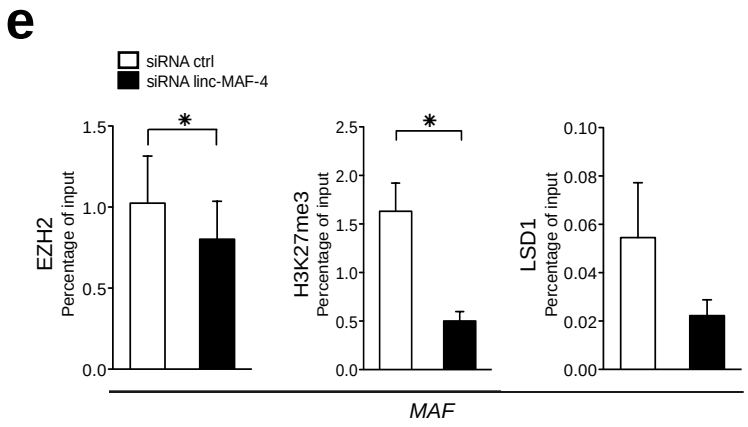
Figure 4

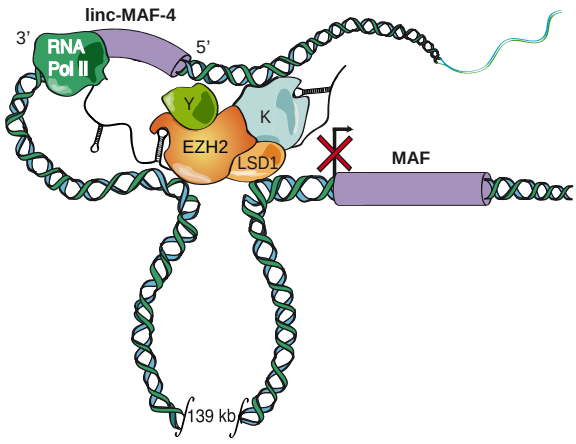

a
linc-MAF-4
AC009159.1
XLOC_012016
MAF
INGMG_000968
Frequency of interactions
Bgl II
chr16
bp
79,800,000
79,750,000
79,700,000
79,650,000
79,600,000


b
M1-L7
M1-L12
BglII
16 dna:chromosome:GRCh37:16:79619739:79804827:1
Strand
Plus/Plus
Plus/Minus


c
Relative percentage of total
Cytoplasm
Nucleoplasm
Chromatin
linc-MAF-4
Linc00339
Malat1
RNU2.1

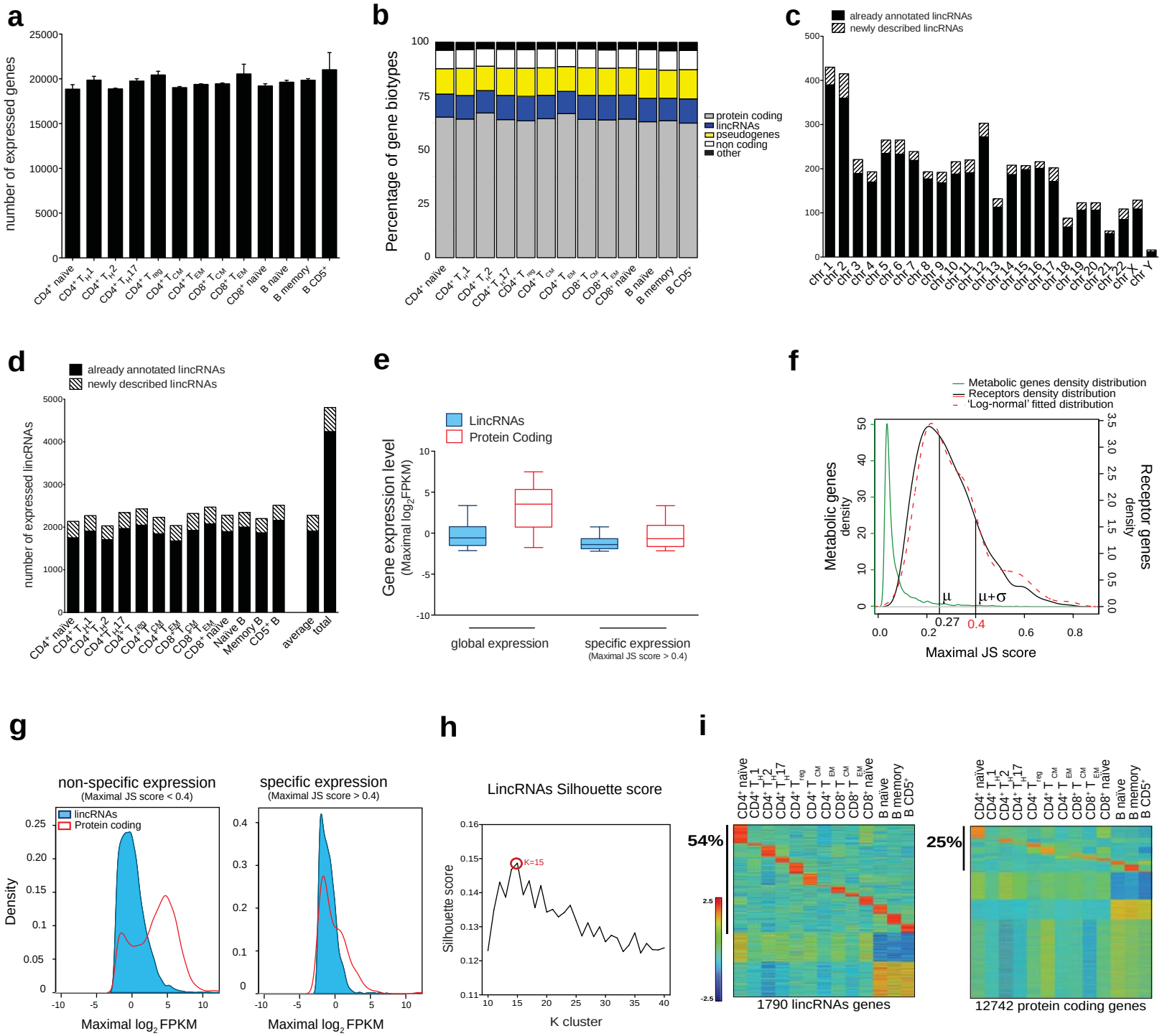

d
LSD1 (fold enrichment)
EZH2 (fold enrichment)
β-actin
RNU2.1
linc-MAF-4 control
linc-MAF-4


e
siRNA ctrl
siRNA linc-MAF-4
EZH2
Percentage of input
H3K27me3
LSD1
MAF


f
linc-MAF-4
RNA Pol II
Y
K
EZH2
LSD1
MAF
139 kb

Subset	Purity (%)	Sorting phenotype	Genes
$CD4^+$ naïve	99,8 ± 0,1	$CD4^+$ $CCR7^+$ $CD45RA^+$ $CD45RO^-$	20061
$CD4^+$ T_H1	99,9 ± 0,05	$CD4^+$ $CXCR3^+$	20855
$CD4^+$ T_H2	99,7 ± 0,3	$CD4^+$ $CRTH2^+$ $CXCR3^-$	19623
$CD4^+$ T_H17	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

Supplementary Figure 1

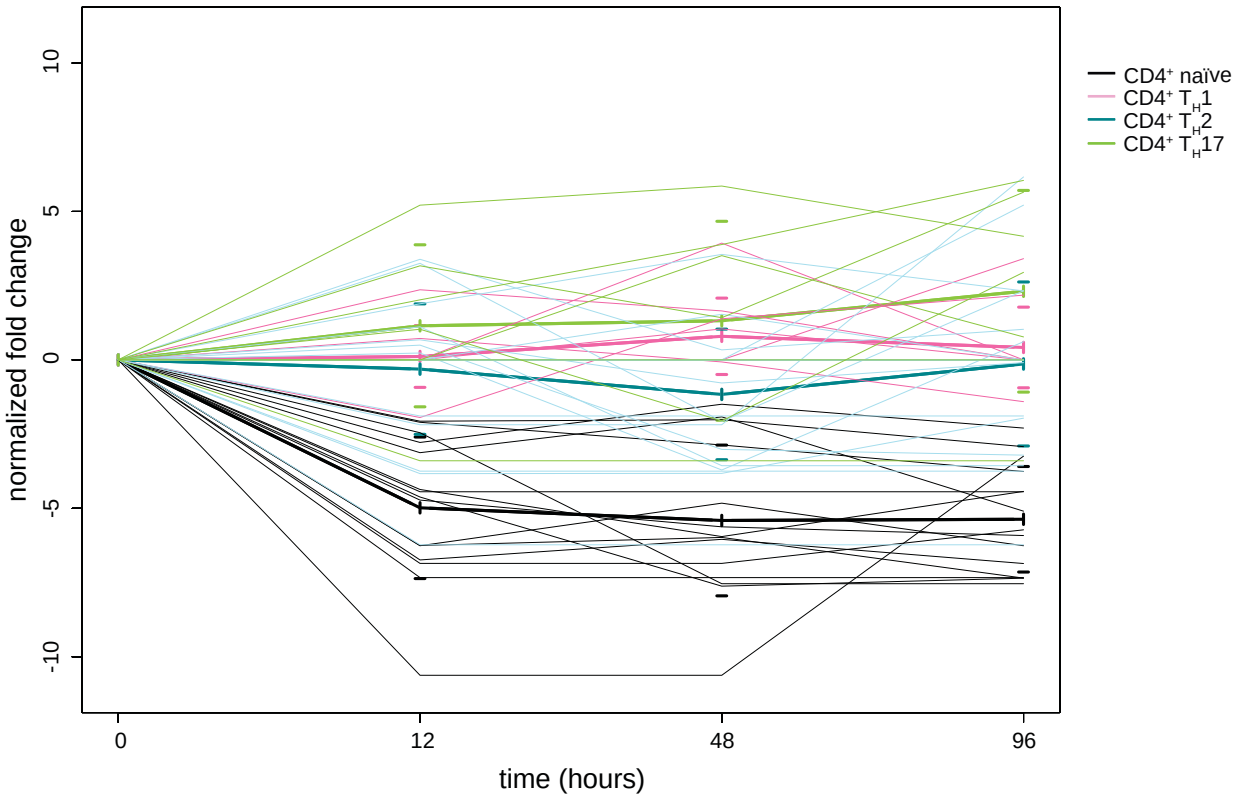


Supplementary Figure 1. LincRNAs distribution and specificity in primary human lymphocytes subsets

(a) Bar plot of expressed genes across a panel of 13 lymphocyte subsets. Average expression (± sdev) of the samples for each subset is reported.
(b) Stacked barplots 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 newly described (striped) and already annotated (black) lincRNAs in all human chromosomes.
(d) Distribution of expressed newly described (striped) and already 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).
(h) 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.
(i) 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 2

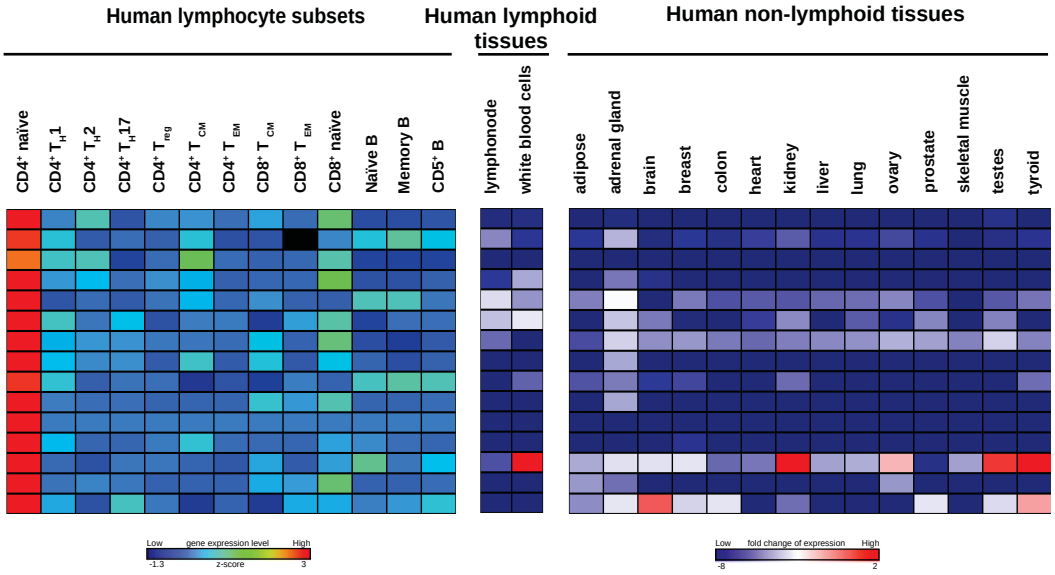
a



b

CD4⁺ naïve signature

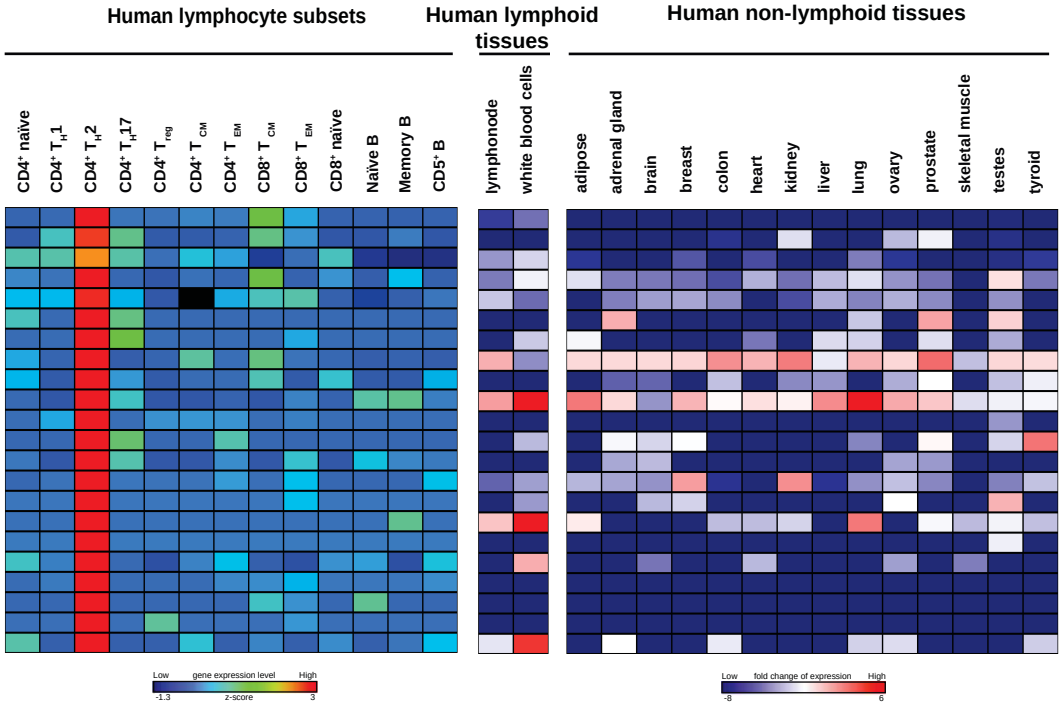
gene id	locus	strand	n° isoforms
INGMG_000614	12:53021754-53024658	-	2
ENSG00000262992	17:18932636-18935795	+	1
XLOC_009373	11:11173943-11177996	-	1
INGMG_001448	2:95740576-95742212	+	7
ENSG00000262292	17:19063665-19065046	+	1
ENSG00000254802	8:61878360-61880334	-	1
INGMG_002593	8:27447901-27450875	+	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



c

CD4⁺ T_H2 signature

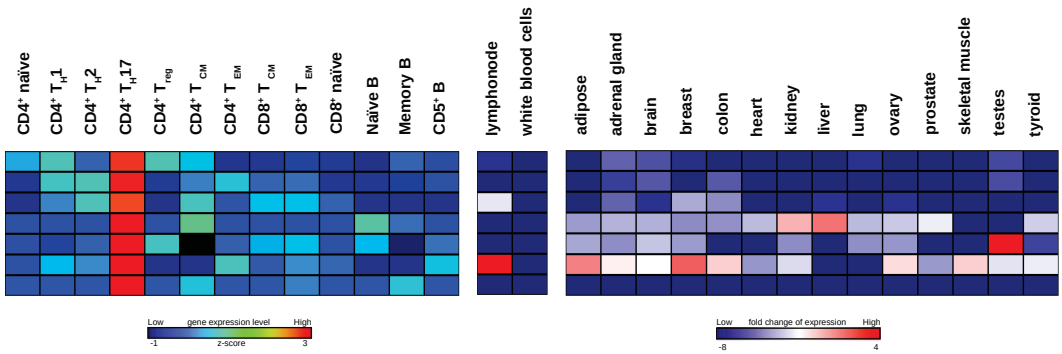
gene id	locus	strand	n° isoforms
INGMG_000354	10:9036884-9061427	-	1
XLOC_008357	10:3985204-4006403	+	2
XLOC_003738	4:153021905-153025384	+	1
XLOC_011681	16:27297150-27301839	+	2
ENSG00000260517	16:29150981-29228027	+	2
XLOC_009457	11:62178649-62179162	-	1
XLOC_009659	12:10393134-10412929	+	1
ENSG00000250786	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



d

CD4⁺ T_H17 signature

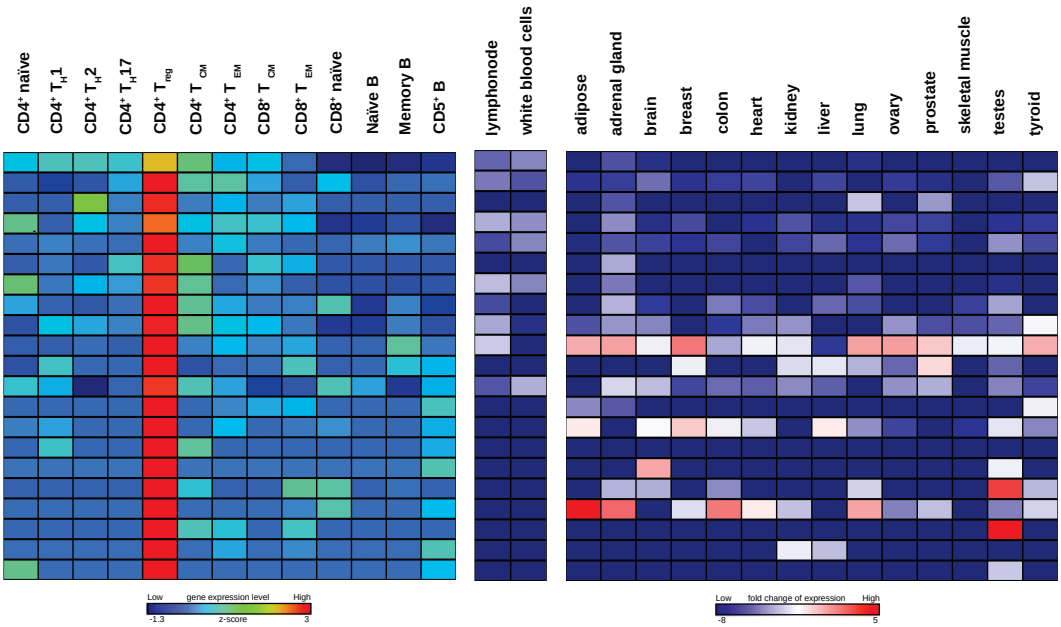
gene id	locus	strand	n° isoforms
INGMG_001733	21:47026508-47032208	+	3
INGMG_001408	2:7797732-7811547	+	10
INGMG_001410	2:7860237-7865579	+	2
XLOC_009027	11:2397410-2398419	+	2
XLOC_002630	3:44465601-44470995	+	5
XLOC_011112	14:95988348-95992377	-	1
ENSG00000260673	6:4599520-4602654	-	1



e

CD4⁺ T_{reg} signature

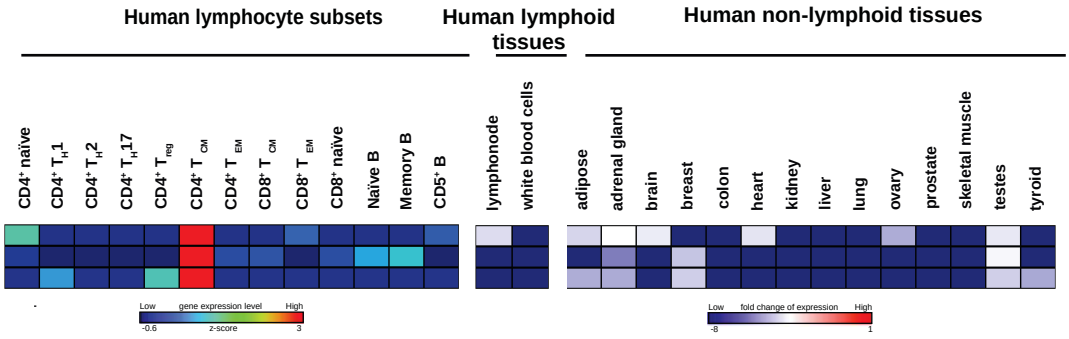
gene id	locus	strand	n° isoforms
INGMG_001638	20:21537986-21541942	+	3
INGMG_001237	19:5978323-5980738	+	1
ENSG00000236481	16:26596075-26606134	-	1
ENSG00000253522	5:159895274-159914433	+	1
XLOC_008164	X:49121663-49123331	-	1
INGMG_001500	2:204738800-204762117	+	2
ENSG00000235304	X:39164209-39186616	-	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



f

CD4⁺ T_{CM} signature

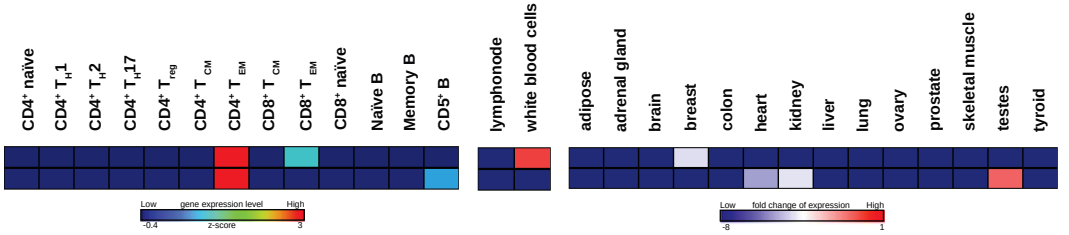
gene id	locus	strand	n° isoforms
ENSG00000254538	8:74582675-74645132	+	2
XLOC_013842	20:61775639-61783415	-	1
ENSG00000237899	1:41134760-41153260	-	1



g

CD4⁺ T_{EM} signature

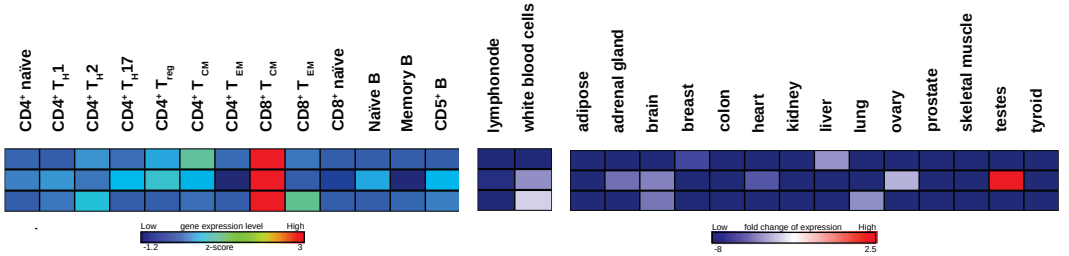
gene id	locus	strand	n° isoforms
XLOC_000627	1:235092977-235095736	+	1
XLOC_005870	6:148454944-148458540	-	1



h

CD8⁺ T_{CM} signature

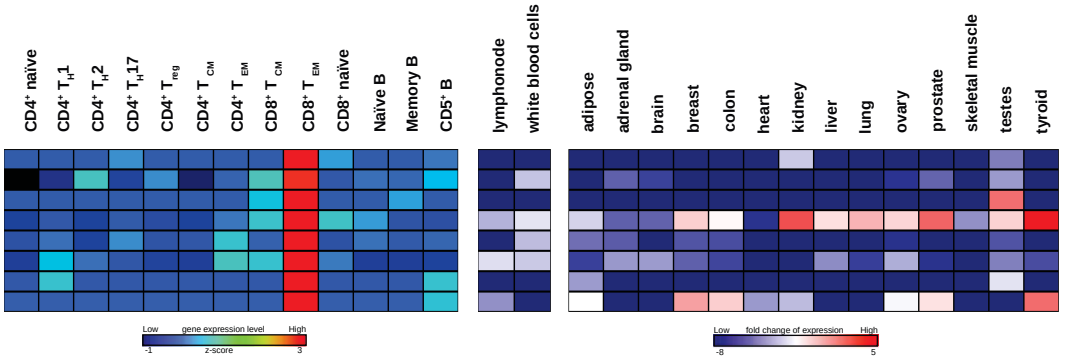
gene id	locus	strand	n° isoforms
INGMG_000280	10:8257084-8259668	+	1
XLOC_005737	6:45523579-45545334	-	11
ENSG00000255484	11:112404944-112426525	-	1



i

CD8⁺ T_{EM} signature

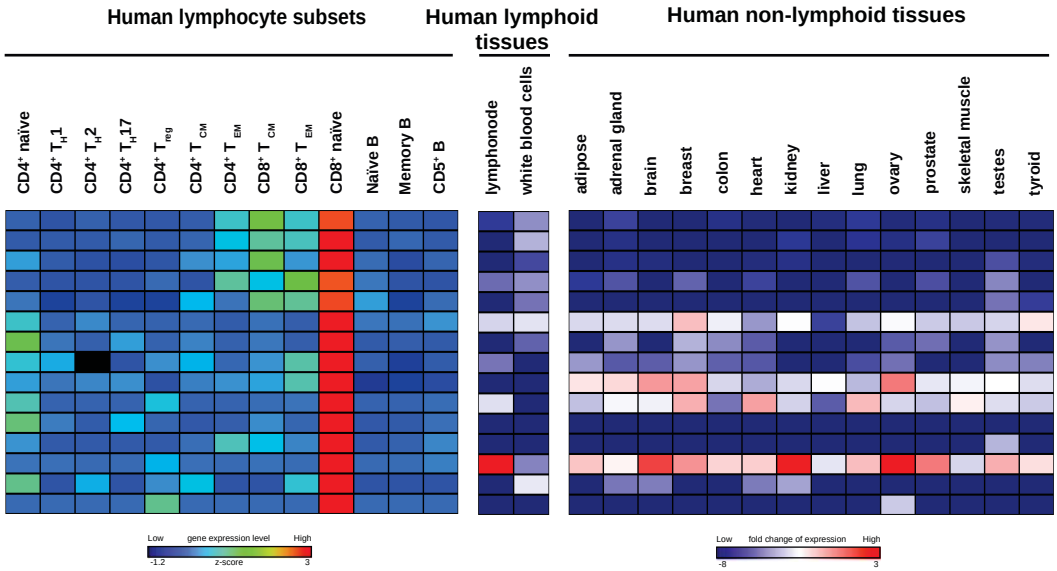
gene id	locus	strand	n° isoforms
XLOC_004238	5:524819-526709	+	1
XLOC_001288	1:244393072-244401962	-	1
XLOC_009505	11:75469515-75470461	-	1
XLOC_013703	20:24911283-24913619	-	1
INGMG_002670	8:128677375-128686846	-	3
INGMG_001017	17:34513843-34516804	+	3
ENSG00000254135	5:157912197-157961446	+	2
XLOC_009361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



j

CD8⁺ naïve signature

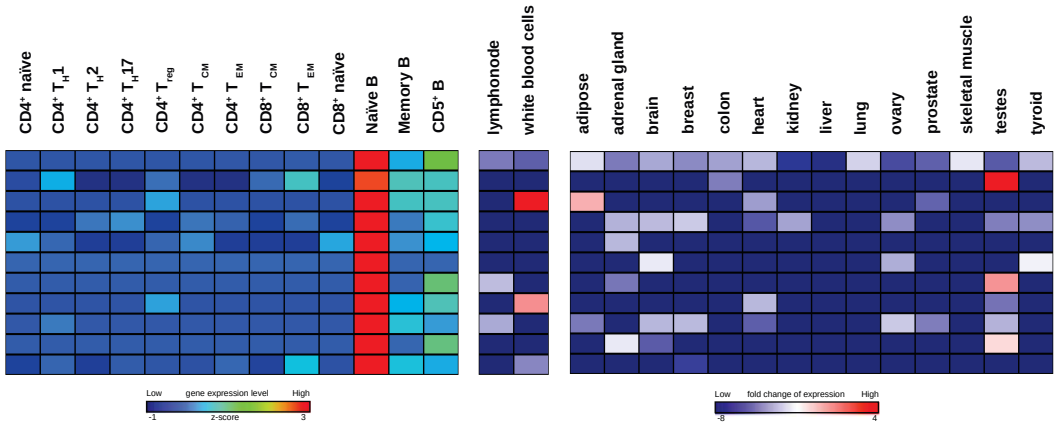
gene id	locus	strand	n° isoforms
XLOC_009661	12:10705978-10710816	+	1
XLOC_009662	12:10725616-10727581	+	1
INGMG_000685	13:114920047-114941975	+	5
INGMG_001014	17:34401948-34404160	+	1
XLOC_010517	13:114944062-114944563	+	1
ENSG00000100181	22:17082776-17179521	+	5
XLOC_010859	14:69446486-69448265	+	1
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



k

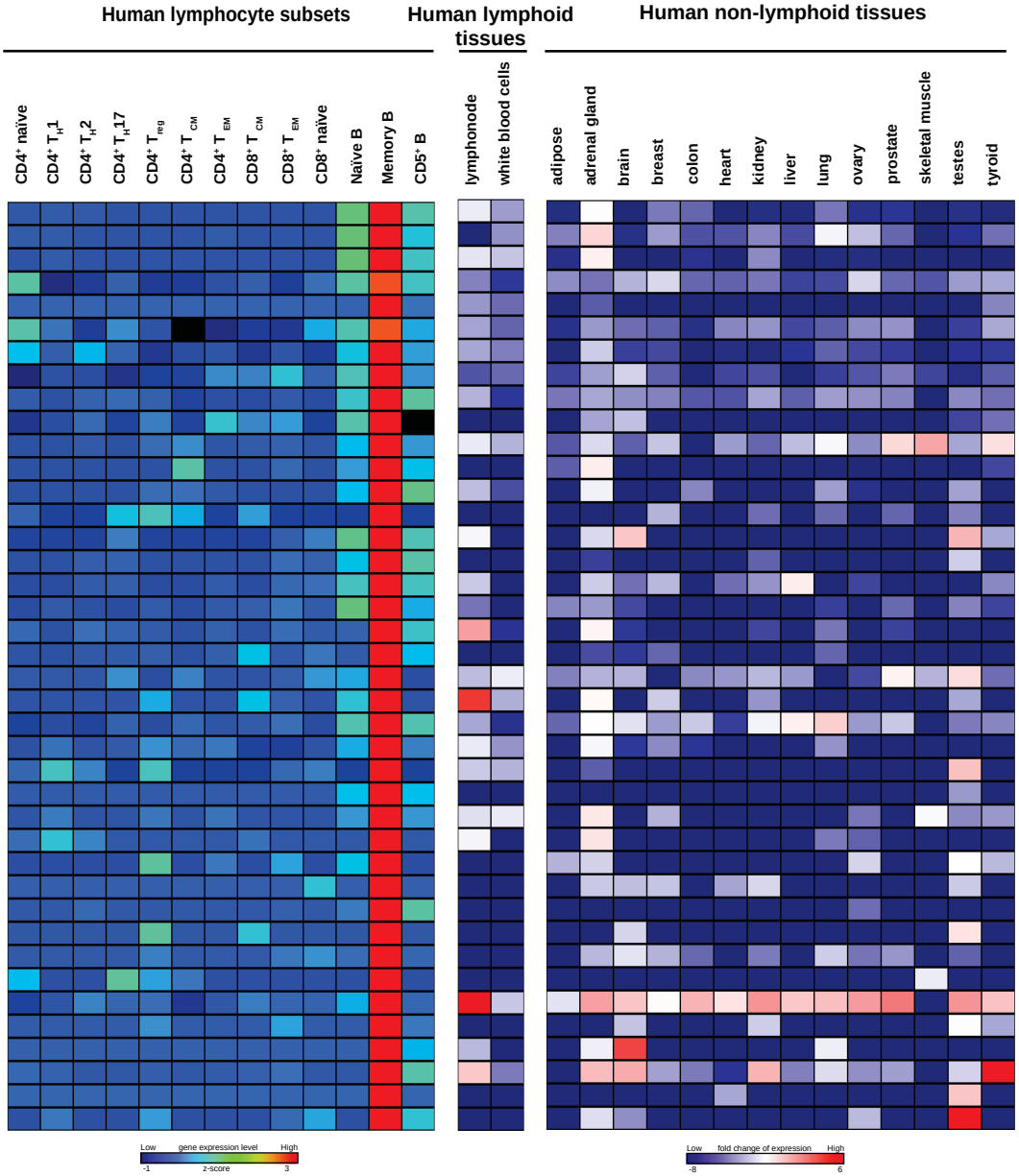
B Naïve signature

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

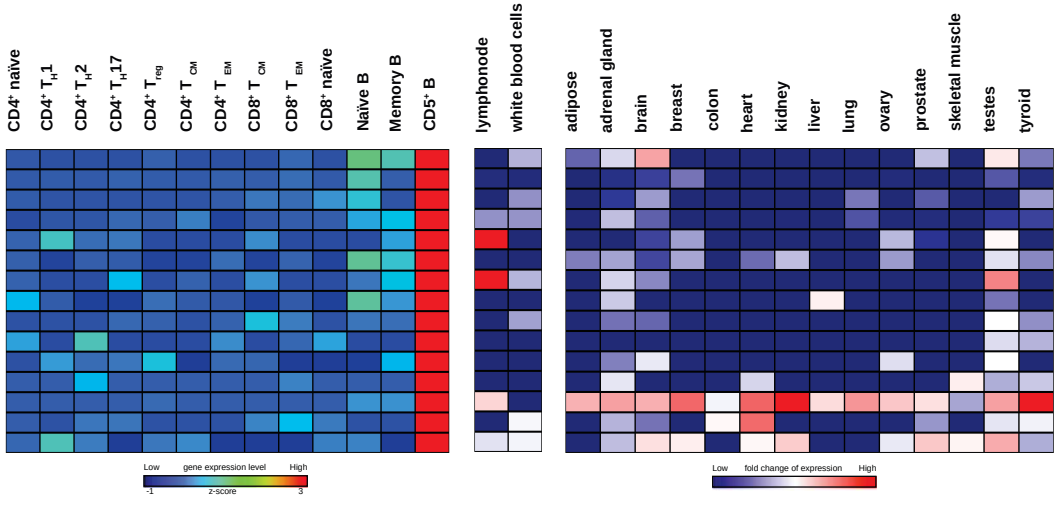
gene id	locus	strand	n° isoforms
ENSG00000253701	14:106170300-106170939	-	1
ENSG00000253364	14:106110832-106115394	-	2
XLOC_000268	1:81001439-81112834	+	4
ENSG00000237438	22:17517459-17539682	+	2
XLOC_002342	2:143628157-143628636	-	1
XLOC_001181	1:207978670-207980881	-	1
XLOC_006293	7:150130741-150145228	+	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



m

B CD5+ signature

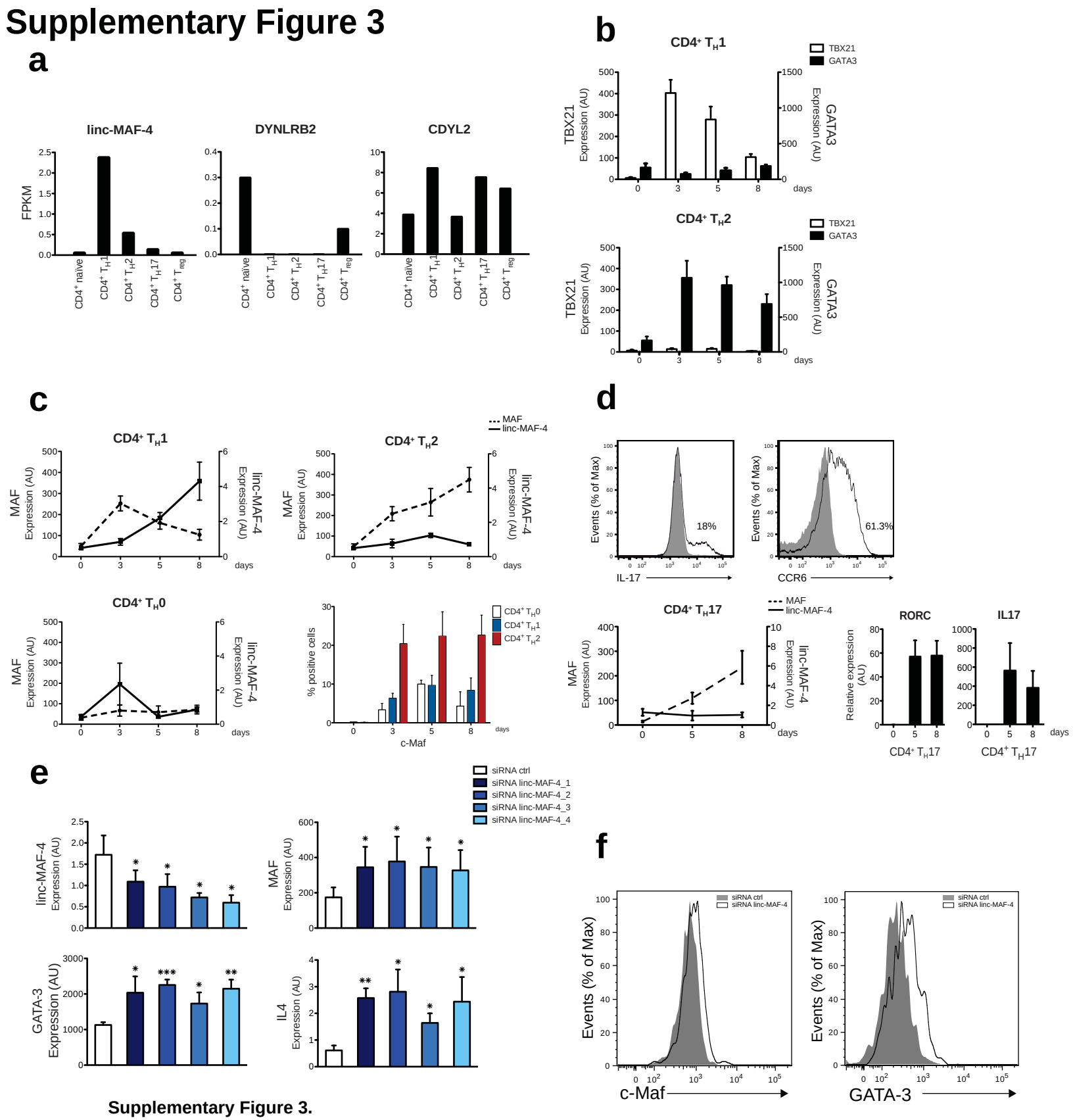
gene id	locus	strand	n° isoforms
XLOC_008554	10:91613272-91674712	+	5
INGMG_002637	8:239189-246382	-	1
INGMG_002250	6:868139-875388	+	2
XLOC_002130	2:65128973-65132271	-	1
XLOC_002613	3:34242414-34310303	+	2
INGMG_000383	10:96871114-96872423	-	1
XLOC_002612	3:34200825-34604551	+	9
CABG_006664	7:155061985-155069592	-	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



Supplementary Figure 2. LincRNA signatures in human lymphocyte subsets

(a) CD4+ naïve, T_H1, T_H2 and T_H17 signature lincRNAs trends in CD4+ naïve T cells differentiated in T_H0 conditions. RNA was collected at different time points during CD4+ naïve T cells differentiation and RNA-seq experiments were peformed. 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.
(b-m) Heatmaps of signature lincRNAs expression for each lymphocytes subset. The barcode on the left indicates already annotated lincRNAs (white) and newly described lincRNAs (brick red). 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).

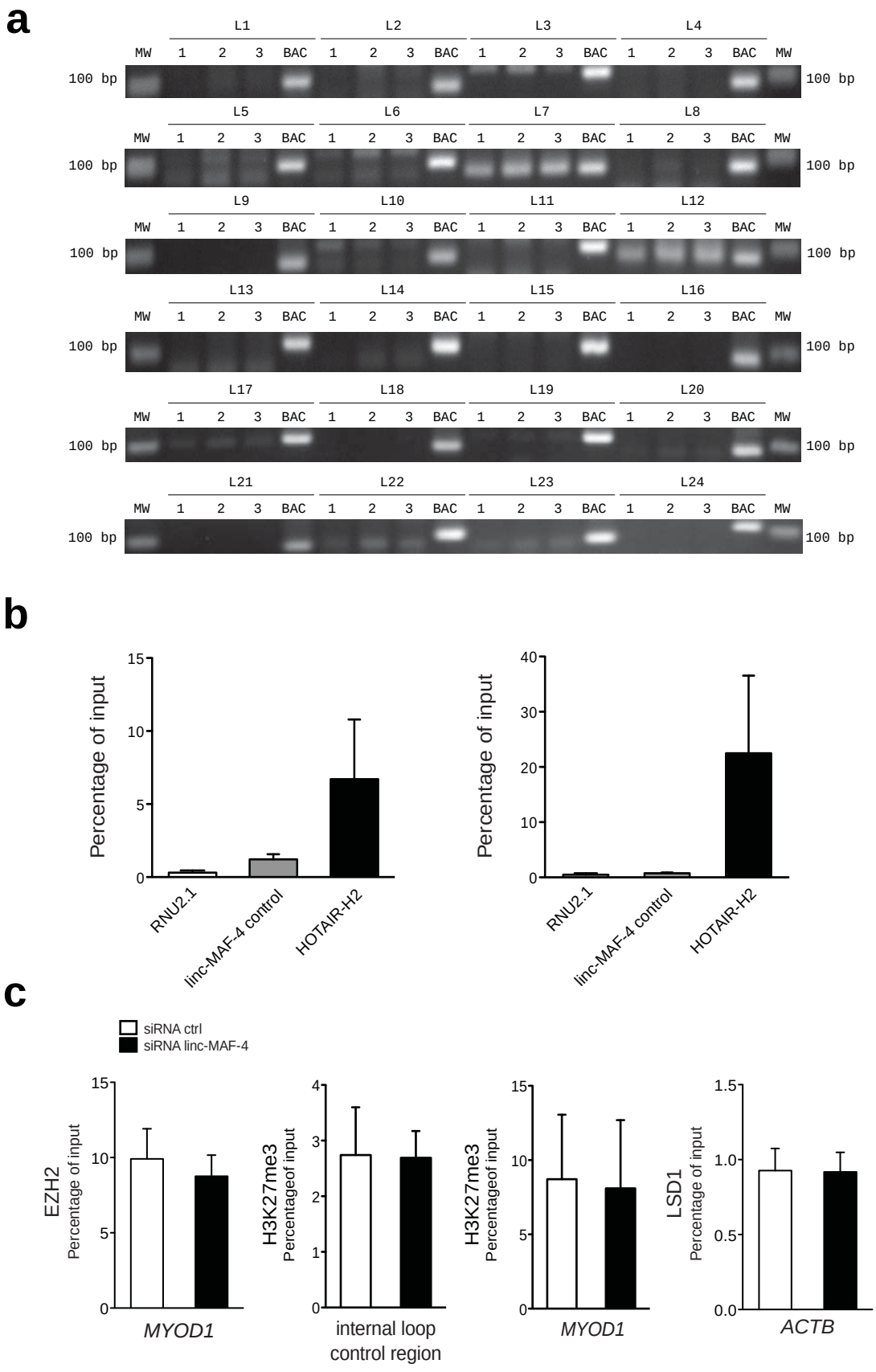
Supplementary Figure 3



Supplementary Figure 3.

(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. Barplot 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 contorl siRNA or linc-MAF-4 siRNA assessed at day 4 post-transfection and activation. Data are representative of five independent experiments.

Supplementary Figure 4



Supplementary Figure 4. 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) Validation of anti-LSD1 and EZH2 antibodies used in RIP assay. LSD1 and EZH2 immunoprecipitets 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

(c) 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)

VALIDATION PRIMERS

PRIMER ID	TARGET	SEQUENCE (5'-> 3')
linc-MAF-4 F	XLOC_012017	GGCTACGTCTCCATTGTTT
linc-MAF-4 R	XLOC_012017	TGGTGTTTGGGATCATTTGT
T4F	ENSG00000241558	CTGTGGGTGACCAGCATCAT
T4R	ENSG00000241558	CAGCAGGCAGTGAGGACAG
T5F4	ENSG00000255094	CCAGATACAGAGAAGCCACAGATG
T5R	ENSG00000255094	GCAGTCATTCTTGAATTGCTGTTA
T6F	ENSG00000257860	TTTCATGGTGAGGGAGAATGG
T6R	ENSG00000257860	CTGGGTCTTGCCTCTTAATGT
T8F2	INGMG_000772	AGCCTGGGCTTTGGAGTC
T8R3	INGMG_000772	GGCTTTGCCAGGATCTCACA
T13F2	XLOC_008683	CGCACAGGAGAACACTCAA
T13R2	XLOC_008683	CACTGATGGCAATGCTCAC
T14F2	XLOC_013818	CCTTGGAAATGTTGCGGTAT
T14R2	XLOC_013818	AATTACCCTGGATGGCTTCA
T16F2	INGMG_000808	AACTGGATCTGAGGCAGATG
T16R3	INGMG_000808	GTAGCACAGGGACACAATTA
T17F2	INGMG_001099	AGTCTCCAGGTGGCTTCT
T17R	INGMG_001099	CTCCTTCTGCTGCCATGTAA
T18F2	INGMG_002122	CCACCATGCTCATTCTCCATT
T18R2	INGMG_002122	CTTGTCCCTCTTCCAGCATTT
T21F2	ENSG00000234535	GAAATGCCAATGAAGCAGAAAG
T21R2	ENSG00000234535	GTGCAAAGAATAGGAGGTTTGA
T24F1	XLOC_002906	GTTATCTGTTGCCAGTTGTT
T24R1	XLOC_002906	ACCTCTGCTTATTGCTGATT
T25F4	XLOC_004086	GAGAGTCTGGCTCTGTTGTC
T25R4	XLOC_004086	GCCTGTACTCCCAGCTATTC
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
T32F1	XLOC_009643	TCCAAGACACTGAGTGATTT
T32R1	XLOC_009643	GCAACAACGGATTTGTCAAG
T33F1	ENSG00000259849	ACCCTCCAGCATGTGTTC
T33R1	ENSG00000259849	CTCCCATTCTGGGCACTT
N1F	XLOC_010212	CTTGGCTGTGGAACCCAGAT
N1R	XLOC_010212	AGCCTCCGTTTACAAACTGGAA
N3F	ENSG00000226137	TGTCCCGACGCATTTACAGA
N3R	ENSG00000226137	AGGCAAGCAGTCAGGTTCC
N8F	INGMG_000894	ATAGGCGGGTAAATGTGGAC
N8R	INGMG_000894	TCTCAAAGGCCTAGGAATTGG
N10F	INGMG_002461	TGTGAACCTGTGGAGGATCT
N10R	INGMG_002461	CTTCAGGCAACATAGCCATTT

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

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.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

17. Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, *et al.* lincRNAs act in the circuitry controlling

pluripotency and differentiation. *Nature* 2011, **477**(7364)**:** 295-300.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

42. Ho IC, Lo D, Glimcher LH. c-maf promotes T helper cell type 2 (Th2) and attenuates Th1 differentiation by both interleukin 4-dependent and -independent mechanisms. *The Journal of experimental medicine* 1998, **188**(10)**:** 1859-1866.

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

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

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

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

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

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

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

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

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

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

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

Conclusions and perspectives in translational medicine

Long non-coding RNAs act as fine tuners of cellular functions throughout the human body. Here we provided a clear example of a lincRNA with a key fundamental function for the proper human immune system differentiation.

Given these observations, it is not surprising that lncRNAs altered expression has been linked to many different pathologies (table 5). Indeed association signals in complex diseases often derive from non-coding regions of the genome. Genome-wide association studies (GWAS) have revealed a large number of genetic variants in lncRNA genes related to diseases[1]. Studies showed that single nucleotide polymorphisms (SNPs) also serve as biomarkers for diagnosis and prognosis[2, 3, 4, 5].

The involvement of lncRNAs in diseases related to immune system remains more elusive, reflecting our poor knowledge of this field. Crohn's disease (CD) and ulcerative colitis were associated to the gene *LLRK2* in a GWAS study. This sequence is part of a complex that includes the lncRNA repressor of NFAT[6]. Similarly, lncRNA DQ786243 was found to be upregulated in the blood of patients with CD and to be related with the expression of CREB, a regulator of FoxP3. Therefore this lncRNA is likely to be involved in inflammation control and CD pathogenesis[7]. Recently a genome-wide association study showed that the expression profiles of lncRNAs associated with autoimmune and immune-related diseases (AIDs) predict cell type specificity better than AID protein-coding genes[8].

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

Table 5 - Characterized lncRNAs with potential role in human diseases

There are now many examples in literature of the beneficial effect of the therapeutical modulation of lncRNAs. ANRIL (also known as CDKN2B-AS1) is an antisense lncRNA overexpressed in prostate cancer. Loss of its expression is associated with a reduction in cellular lifespan and its downregulation increases the expression of the neighboring tumor suppressor genes INK4A and INK4B[9]. Similarly, the well-known lncRNA HOTAIR is found to be overexpressed in a

variety of primary and metastatic tumor and its expression levels are correlates with a worse prognosis[10, 11, 12]. Once again, its suppression promotes apoptosis under proapoptotic stimuli[13]. LncRNAs have also been related to normal and disease processes of the nervous system, such as Alzeheimer's[14, 15] and Huntington's diseases[16]. Intriguingly, microsatellite expansions, common in neurological disorders, generates antisense transcription[17, 18, 19]. Recently, the pathological expansion of the D4Z4 repeats in facioscapulohumeral muscular dystrophy proved to be the condition for the transcription of a chromatin-associated lncRNA that causes upregulation of gene transcription in the locus[20].

Long noncoding RNAs could also be useful novel biomarkers for diagnosis, prognosis and prediction of response to therapies. PCA3/DD3 was originally discovered in a differential display analysis comparing normal and tumor prostate cancer[21] and it is characterized by an increased and unique expression in tumors. Particularly, it is expressed in early-stage tumors and detectable in urine. In clinical trials it proved to be as powerful as classic prostate-specific antigen biomarkers[22, 23].

Targeting lncRNAs could be particularly useful when upregulation of gene expression is needed, for example of tumor suppressors, neuroprotective growth factors, proteins or transcription factors whose deficient or reduced expression is often related with Mendelian monogenic disorders. In these cases, traditional therapies such as peptides administration or enzyme replacement are not curative, but conversely require lifelong administration. Even recent viral vectors-mediated therapies have some drawback, including the

short-lived nature of the treatment[24], immune response activation[25], toxicity[26, 27] and insertional mutagenesis[28, 29, 30]. Targeting lncRNAs could expand the druggable portion of the genome and elicit more specific consequences in response to a less invasive treatment. Indeed, lncRNAs have higher cell specificity than protein coding genes; act on a restrict set of targets in a selected subpopulation of cells; exert a direct regulation on gene expression through the modulation of chromatin modifications and are less expressed than common protein-coding genes, therefore being easier to target and modulate. This last property should not be interpreted as a diminished efficacy of the therapy. As mentioned before, lncRNAs even if less expressed can have profound effects in cellular biochemistry. Indeed, oncogenic lncRNAs such as CCAT2 have demonstrated not only to promote oncogenic activity and induce chromosomal instability, but also to regulate the expression of key developmental genes such as *MYC*, involved in the WNT signaling network[31]. Targeting these lncRNAs is likely to have a broad effect on cancer-associated pathways.

LncRNAs can be targeted by traditional siRNA treatments, even though many of the described lncRNAs are enriched in the nucleus, where they may be less accessible. Also, an extensive secondary structure or repetitive nucleotide sequence could be unfavourable for an optimal siRNA design. Therefore antisense oligonucleotides have been introduced, having some advantages over siRNAs: they act independently to the RISC machinery, they are characterized by higher specificity and fewer off-target effects[32]. This strategy was used to target MALAT1 function thus blocking metastatic events in a lung cancer mouse model[33]. In 2005 an

approach was proposed to target natural antisense transcripts (NATs)[34] through antagoNATs[35, 36]: single-stranded oligonucleotides designed to block the interactions between NATs and their sense mRNA and or to degrade the antisense transcript. This class of transcripts is particularly interesting given that it has been estimated that approximately one-third of protein-coding genes are regulated by NATs[37]. Recently antagoNATs have been modified to improve their stability, for example through the introduction of locked nucleic acids (LNAs) within their structure[38]. This approach proved to be highly specific and capable of inducing locus-specific regulation without perturbing control genes, even in proximity to the target[35, 39, 40]. Ribozymes or deozyribozymes (hammerhead ribozymes) were also used for targeting shorter lncRNAs characterized by an extensive secondary structure. These molecules bind to a complementary target sequence catalyzing the cleavage of the RNA region flanking the pairing site[32]. Other approaches involve the use of synthetic hairpin-structured RNA molecules that mimic the target lncRNA, acting as competitors for its function. LncRNAs with these properties are already known: GAS5 inhibit the interaction of the glucocorticoid receptor with DNA promoters acting through a specific hairpin-structure[41].

Targeting of different noncoding molecules can often be combined, as some of the effects mediated by microRNAs-mediated therapies can be attributed to the targeting of their downstream lncRNAs. For example, miR-155 was shown to directly target the transcribed ultraconserved region 160 (T-UCR 160)[42]. Again, competing endogenous RNAs (ceRNAs) can compete for microRNAs

binding therefore inhibiting their binding to targets[43]. In particular, the tumor suppressor gene *PTEN* and the lncRNA *PTENP1* are targeted by the same set of microRNAs. Therefore if *PTENP1* expression decreases, more microRNAs are available to target *PTEN*, that becomes downregualted, thus generating pro-tumorigenic effects[44]. Similarly, circular RNAs (circRNAs) have attracted much attention because they bind and sequester miRNAs, derepressing mRNA genes[45].

Today many companies in the USA are investigating these therapeutic approaches involving lncRNAs[37]. Of course, these technologies share some drawbacks with the already-mentioned classical therapeutic strategies, such as toxicity and pro-inflammatory properties that are intrinsic to oligonucleotides-mediated approaches[46, 47, 48]. Off-target effects could also be possible and it is therefore important to use stringent controls to evaluate alterations in the expression of genes other than the intended one[38, 49]. Safe and efficient *in vivo* delivery is another crucial hurdle that lncRNA targeting technologies have in common with oligonucleotide-based therapies. Systemic delivery by intravenous treatment was approved by the US Food and Drug Administration (FDA) for a NAT targeting the apolipoprotein B, but denied by the Europeaan Medicines Agency panel[50] due to potential adverse effect. Targeted or highly localized delivery was used to treat the central nervous system (CNS) with relative success[51, 52, 53, 54, 55], especially through the intrathecal route[56, 57]. Acting on chemical modifications[58] or using various delivery approaches (viral or non-viral vectors) can help to decrease immune activation and to achieve a proper dosing control of treatments[47].

Notably, a major advantage of antagoNATs is their ability to be administered systemically without requirement for any delivery vehicles[38, 49].

The immune system is particularly interesting for lncRNAs-mediated therapies. Indeed, lymphocytes are characterized by an extreme flexibility and plasticity. This property ensures a proper immune response to different external clues and challenges, in a way that is certainly advantageous in terms of host defense. Conversely though lymphocytes are also major players in mediating autoimmune and allergic diseases. The failure to generate a proper differentiation signaling cascade can indeed lead to diseases: dominant negative *STAT3* mutations characterize the hyperimmunoglobulin E syndrome (Job's syndrome), due to the failure to differentiate T lymphocytes into T_H17 cells[59, 60]. Indeed, IL-6 and IL-23 both signal through STAT3 driving T_H17 differentiation together with IL-1 and TGF-β. Similarly, gain-of-function mutation involving *STAT1* causes primary immunodeficiency such as the chronic mucocutaneous candidiasis again characterized by an impaired T_H17 generation, being STAT1 an important negative regulator of T_H17 differentiation[61]. Thus the proper balancing between the different lymphocytes subsets is of key importance to ensure protection against infections, hypo- or hyper-immune responses syndromes. We must not forget, though, that the picture is not so easy. The plasticity of T cell differentiation emerges also in case of pathologies. For example, it is now well appreciated that IgE is a central player in the pathophysiology of allergies and asthma[62, 63]. The generation of B cells producing this immunoglobulin is triggered by IL-4, but now is clear that also T_{FH} cells are

fundamental for providing B-cell help in germinal centers[64, 65]. The prototypical cytokine for T_{FH} cells is not IL-4, but IL-21. Nonetheless, these cells are able to produce also IFN-γ, IL-4, IL-17 and IL-10[65, 66, 67]. Thus, it is clear that lymphocytes provide host defense and generate immune-mediated diseases by attaining multiple distinct fates even after their initial differentiation, thanks to their intrinsic plasticity. In this context we can envision therapies to modulate the balance between effector lymphocytes reprogramming already differentiated cells, thanks to the exploitation of their peculiar plasticity through lncRNAs-mediated therapies. Indeed, given that lncRNAs act as fine-tuners of cell differentiation, we could modulate the differentiation network acting through lncRNAs. This strategy could have the already mentioned advantages and could act in a less invasive and more specific way. Indeed, given that lncRNAs are not major hubs within cell networks, their overexpression or downregulation would cause a more physiological cascade with minor perturbations if compared to the overexpression or downregulation of a key regulatory hub such as a master gene. Strategies like this could help us in the modulation of T_H1-T_H2 balance during allergic responses or could decrease Treg differentiation counteracting Treg mediated inhibition of immune responses at tumor sites, just to give an example.

Thus understanding the mechanisms of function of lncRNAs in driving the differentiation events in the human immune system, like in our case, is of central importance for the identification of novel and more specific therapeutic targets for immune-related diseases.

References

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

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

3. Liu Y, Pan S, Liu L, Zhai X, Liu J, Wen J*, et al.* A genetic variant in long non-coding RNA HULC contributes to risk of HBV-related hepatocellular carcinoma in a Chinese population. *PloS one* 2012, **7**(4)**:** e35145.

4. Zhu Z, Gao X, He Y, Zhao H, Yu Q, Jiang D*, et al.* An Insertion/Deletion Polymorphism within RERT-lncRNA Modulates Hepatocellular Carcinoma Risk. *Cancer Research* 2012, **72**(23)**:** 6163-6172.

5. Xue Y, Wang M, Kang M, Wang Q, Wu B, Chu H*, et al.* Association between lncrna PCGEM1 polymorphisms and prostate cancer risk. *Prostate Cancer Prostatic Dis* 2013, **16**(2)**:** 139-144.

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

7. Qiao Y, Huang M, Xu A, Zhao D, Ran Z, Shen J. LncRNA DQ786243 affects Treg related CREB and Foxp3 expression in Crohn's disease. *Journal of Biomedical Science* 2013, **20**(1)**:** 87.

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

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

10. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ*, et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 2010, **464**(7291)**:** 1071-1076.

11. Kogo R, Shimamura T, Mimori K, Kawahara K, Imoto S, Sudo T*, et al.* Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res* 2011, **71**(20)**:** 6320-6326.

12. Niinuma T, Suzuki H, Nojima M, Nosho K, Yamamoto H, Takamaru H*, et al.* Upregulation of miR-196a and HOTAIR Drive Malignant Character in Gastrointestinal Stromal Tumors. *Cancer Research* 2012, **72**(5)**:** 1126-1136.

13. Yang Z, Zhou L, Wu LM, Lai MC, Xie HY, Zhang F*, et al.* Overexpression of long non-coding RNA HOTAIR predicts tumor recurrence in hepatocellular carcinoma patients following liver transplantation. *Annals of surgical oncology* 2011, **18**(5)**:** 1243-1250.

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

15. Mus E, Hof PR, Tiedge H. Dendritic BC200 RNA in aging and in Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(25)**:** 10679-10684.

16. Johnson R. Long non-coding RNAs in Huntington's disease neurodegeneration. *Neurobiology of disease* 2012, **46**(2)**:** 245-254.

17. Mirkin SM. Expandable DNA repeats and human disease. *Nature* 2007, **447**(7147)**:** 932-940.

18. Moseley ML, Zu T, Ikeda Y, Gao W, Mosemiller AK, Daughters RS*, et al.* Bidirectional expression of CUG and CAG expansion transcripts and intranuclear polyglutamine inclusions in spinocerebellar ataxia type 8. *Nat Genet* 2006, **38**(7)**:** 758-769.

19. Cho DH, Thienes CP, Mahoney SE, Analau E, Filippova GN, Tapscott SJ. Antisense transcription and heterochromatin at the DM1 CTG repeats are constrained by CTCF. *Molecular cell* 2005, **20**(3)**:** 483-489.

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

21. Bussemakers MJ, van Bokhoven A, Verhaegh GW, Smit FP, Karthaus HF, Schalken JA*, et al.* DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res* 1999, **59**(23)**:** 5975-5979.

22. Lee GL, Dobi A, Srivastava S. Prostate cancer: Diagnostic performance of the PCA3 urine test. *Nat Rev Urol* 2011, **8**(3)**:** 123-124.

23. Day JR, Jost M, Reynolds MA, Groskopf J, Rittenhouse H. PCA3: From basic molecular science to the clinical lab. *Cancer Letters*, **301**(1)**:** 1-6.

24. Ciesielska A, Hadaczek P, Mittermeyer G, Zhou S, Wright JF, Bankiewicz KS*, et al.* Cerebral infusion of AAV9 vector-encoding non-self proteins can elicit cell-mediated immune responses. *Molecular therapy : the journal of the American Society of Gene Therapy* 2013, **21**(1)**:** 158-166.

25. Unzu C, Hervas-Stubbs S, Sampedro A, Mauleon I, Mancheno U, Alfaro C*, et al.* Transient and intensive pharmacological immunosuppression fails to improve AAV-based liver gene transfer in non-human primates. *Journal of translational medicine* 2012, **10:** 122.

26. Waehler R, Russell SJ, Curiel DT. Engineering targeted viral vectors for gene therapy. *Nature reviews Genetics* 2007, **8**(8)**:** 573-587.

27. Mingozzi F, High KA. Therapeutic in vivo gene transfer for genetic disease using AAV: progress and challenges. *Nature reviews Genetics* 2011, **12**(5)**:** 341-355.

28. Thrasher AJ, Gaspar HB, Baum C, Modlich U, Schambach A, Candotti F*, et al.* Gene therapy: X-SCID transgene leukaemogenicity. *Nature* 2006, **443**(7109)**:** E5-E6.

29. Gaspar HB, Cooray S, Gilmour KC, Parsley KL, Adams S, Howe SJ*, et al.* Long-Term Persistence of a Polyclonal T Cell Repertoire After Gene Therapy for X-Linked Severe Combined Immunodeficiency. *Science Translational Medicine* 2011, **3**(97)**:** 97ra79.

30. Wood AJ, Lo T-W, Zeitler B, Pickle CS, Ralston EJ, Lee AH*, et al.* Targeted Genome Editing Across Species Using ZFNs and TALENs. *Science (New York, NY)* 2011, **333**(6040)**:** 307-307.

31. Ling H, Spizzo R, Atlasi Y, Nicoloso M, Shimizu M, Redis RS*, et al.* CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal

instability in colon cancer. *Genome research* 2013, **23**(9)**:** 1446-1461.

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

33. Gutschner T, Hammerle M, Eissmann M, Hsu J, Kim Y, Hung G*, et al.* The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res* 2013, **73**(3)**:** 1180-1189.

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

35. Modarresi F, Faghihi MA, Lopez-Toledano MA, Fatemi RP, Magistri M, Brothers SP*, et al.* Natural Antisense Inhibition Results in Transcriptional De-Repression and Gene Upregulation. *Nature biotechnology* 2012, **30**(5)**:** 453-459.

36. Wahlestedt C. Natural antisense and noncoding RNA transcripts as potential drug targets. *Drug discovery today* 2006, **11**(11-12)**:** 503-508.

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

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

39. Schwartz JC, Younger ST, Nguyen N-B, Hardy DB, Monia BP, Corey DR*, et al.* Antisense transcripts are targets for activating small RNAs. *Nat Struct Mol Biol* 2008, **15**(8)**:** 842-848.

40. Scheele C, Petrovic N, Faghihi M, Lassmann T, Fredriksson K, Rooyackers O*, et al.* The human PINK1 locus is regulated in vivo by a non-coding natural antisense RNA during modulation of mitochondrial function. *BMC Genomics* 2007, **8**(1)**:** 74.

41. Kino T, Hurt DE, Ichijo T, Nader N, Chrousos GP. Noncoding RNA Gas5 Is a Growth Arrest and Starvation-Associated Repressor of the Glucocorticoid Receptor. *Science signaling* 2010, **3**(107)**:** ra8.

42. Calin GA, Liu C-g, Ferracin M, Hyslop T, Spizzo R, Sevignani C*, et al.* Ultraconserved Regions Encoding ncRNAs Are Altered in Human Leukemias and Carcinomas. *Cancer Cell* 2007, **12**(3)**:** 215-229.

43. Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. *Nature* 2014, **505**(7483)**:** 344-352.

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

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

46. Krieg AM. Therapeutic potential of Toll-like receptor 9 activation. *Nat Rev Drug Discov* 2006, **5**(6)**:** 471-484.

47. Bennett CF, Swayze EE. RNA Targeting Therapeutics: Molecular Mechanisms of Antisense Oligonucleotides as a Therapeutic Platform. *Annual Review of Pharmacology and Toxicology* 2010, **50**(1)**:** 259-293.

48. Senn JJ, Burel S, Henry SP. Non-CpG-Containing Antisense 2′-Methoxyethyl Oligonucleotides Activate a Proinflammatory Response Independent of Toll-Like

Receptor 9 or Myeloid Differentiation Factor 88. *Journal of Pharmacology and Experimental Therapeutics* 2005, **314**(3)**:** 972-979.

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

50. Pollack. A. F.D.A. approves genetic drug to treat rare disease. *New York Times [online]* 2013.

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

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

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

54. Yee F, Ericson H, Reis D, Wahlestedt C. Cellular uptake of intracerebroventricularly administered biotin- or digoxigenin-Labeled antisense oligodeoxynucleotides in the rat. *Cell Mol Neurobiol* 1994, **14**(5)**:** 475-486.

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

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

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

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

59. Ma CS, Chew GYJ, Simpson N, Priyadarshi A, Wong M, Grimbacher B*, et al.* Deficiency of Th17 cells in hyper IgE syndrome due to mutations in STAT3. *The Journal of experimental medicine* 2008, **205**(7)**:** 1551-1557.

60. Milner JD, Brenchley JM, Laurence A, Freeman AF, Hill BJ, Elias KM*, et al.* Impaired T(H)17 cell differentiation in subjects with autosomal dominant hyper-IgE syndrome. *Nature* 2008, **452**(7188)**:** 773-776.

61. Liu L, Okada S, Kong XF, Kreins AY, Cypowyj S, Abhyankar A*, et al.* Gain-of-function human STAT1 mutations impair IL-17 immunity and underlie chronic mucocutaneous candidiasis. *The Journal of experimental medicine* 2011, **208**(8)**:** 1635-1648.

62. Strunk RC, Bloomberg GR. Omalizumab for Asthma. *New England Journal of Medicine* 2006, **354**(25)**:** 2689-2695.

63. Busse WW, Morgan WJ, Gergen PJ, Mitchell HE, Gern JE, Liu AH*, et al.* Randomized Trial of Omalizumab (Anti-IgE) for Asthma in Inner-City Children. *New England Journal of Medicine* 2011, **364**(11)**:** 1005-1015.

64. Victora GD, Nussenzweig MC. Germinal Centers. *Annual Review of Immunology* 2012, **30**(1)**:** 429-457.

65. Crotty S. Follicular Helper CD4 T Cells (TFH). *Annual Review of Immunology* 2011, **29**(1)**:** 621-663.

66. Reinhardt RL, Liang HE, Locksley RM. Cytokine-secreting follicular T cells shape the antibody repertoire. *Nature immunology* 2009, **10**(4)**:** 385-393.

67. Hsu HC, Yang P, Wang J, Wu Q, Myers R, Chen J*, et al.* Interleukin 17-producing T helper cells and interleukin 17 orchestrate autoreactive germinal center development in autoimmune BXD2 mice. *Nature immunology* 2008, **9**(2)**:** 166-175.