# 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

## 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-$\gamma$ and to control susceptibility to Theiler's virus and Salmonella infection in mice through epigenetic regulation of the IFN-$\gamma$ 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$_H$1 specific lincRNA, correlates inversely with the transcription factor MAF and that its down-regulation skews CD4$^+$ T cell differentiation toward T$_H$2 phenotype.

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

## Results

### LincRNAs identify human lymphocyte subsets better than protein coding genes

To assess lincRNA expression in human primary lymphocytes, RNA was extracted from thirteen lymphocyte cell subsets (Table 1) purified from peripheral blood mononuclear cells (PBMCs) of five healthy donors[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$_H$1 cell subset; similar results were obtained for all the other subsets (Supplementary Fig. 2b-m).

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

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

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

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

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

We further characterized *MAF* transcriptional regulation by looking at H3K4 tri-methylation (H3K4me3) level and RNA polymerase II occupancy at *MAF* promoter region in T$_H$1 and T$_H$2 cells. Consistent with a higher active transcription of *MAF* in CD4$^+$ T$_H$2 cells, we found that H3K4me3 levels in T$_H$2 cells are greater than in T$_H$1 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$_H$1 cells. A common reverse primer mapping within the *MAF* promoter region, was used in combination with a set of primers spanning the locus, and interactions were analysed by PCR. Specific interactions between *MAF* promoter and 5' and 3' end regions of *linc-MAF-4* were detected (Fig. 4a,b and Supplementary Fig. 4a), indicating the existence of an *in cis* chromatin looping conformation that brings *linc-MAF-4* in close proximity to *MAF* promoter. Interestingly, the subcellular fractionation of *in vitro* differentiated CD4$^+$ T$_H$1 lymphocytes revealed a strong enrichment of linc-MAF-4 in the chromatin fraction (Fig. 4c). Because other chromatin-associated lincRNAs regulate neighbouring genes by recruiting specific chromatin remodellers, we tested in RNA immunoprecipitation (RIP) assays the interaction of linc-MAF-4 with different chromatin modifiers, including activators and repressors (data not shown), and found a specific enrichment of linc-MAF-4 in the immunoprecipitates of two repressors, EZH2 and LSD1 (Fig. 4d and Supplementary Fig. 4b). In agreement with these findings, we found that linc-MAF-4 knock-down in activated CD4$^+$ naïve T cells reduces both EZH2 and LSD1 levels and correlates with the reduction of EZH2 enzymatic activity at *MAF* promoter as demonstrated by the H3K27me3 reduction at this locus (Fig. 4e). Remarkably, H3K27me3 levels were reduced neither at *MYOD1* promoter region (a known target of EZH2) nor at a region within the chromatin loop between *linc-MAF-4* and *MAF* marked by H3K27me3 (Supplementary Fig. 4c).

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

## Discussion

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

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

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

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

The lincRNAs role in differentiation has been described in different cell types[17, 20, 23, 46, 47]. In the mouse immune system it has been found that lincRNAs expression changes during naïve to memory CD8$^+$ T cell differentiation[24] and during naïve CD4$^+$ T cells differentiation into distinct helper T cell lineages[28]. We show in human primary lymphocytes that activation induced differentiation of CD4$^+$ naïve

T cells is associated with increased expression of lincRNAs belonging to the CD4$^+$ T$_H$1 signature suggesting that upregulation of T$_H$1 lincRNAs is part of the cell differentiation transcriptional program. Indeed, linc-MAF-4, one of the T$_H$1 signature lincRNA, is poorly expressed in T$_H$2 cells and its experimental downregulation skews differentiating T helper cells toward a T$_H$2 transcription profile. We have found that linc-MAF-4 regulates transcription exploiting a chromatin loop that brings its genomic region close to the promoter of *MAF* gene. We propose that the chromatin organization of this region allows linc-MAF-4 transcript to recruit both EZH2 and LSD1 and modulate the enzymatic activity of EZH2 negatively regulating *MAF* transcription with a mechanism of action similar to that shown for the lincRNAs HOTAIR[48] and MEG3 [49]. We therefore provide a mechanistic proof of concept that lincRNAs can be important regulators of CD4$^+$ T-cell differentiation. Given the number of specific lincRNAs expressed in the different lymphocytes subsets, it can be postulated that many other lincRNAs might contribute to cell differentiation and to the definition of cell identity in human lymphocytes.

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

## Online Methods

### Purification of primary immunological cell subsets

Buffy-coated blood of healthy donors was obtained from the Ospedale Maggiore in Milan and peripheral blood mononuclear cells were isolated by Ficoll-hypaque density gradient centrifugation. The ethical committee of Istituto di Ricovero e Cura a Carattere Scientifico Policlinico Ospedale Maggiore approved the use of PBMCs from healthy donors for research purposes, and informed consent was obtained from subjects. Human blood primary lymphocyte subsets were purified >95% by cell sorting using different combinations of surface markers (Table 1). For *in vitro* differentiation experiments resting naïve $CD4^+$ T cells were purified >95% by negative selection with magnetic beads with the isolation kit for human $CD4^+$ Naïve T cells of Miltenyi and stimulated with Dynabeads Human T-Activator CD3/CD28 (Life Technologies). IL-2 was added at 20 IU/ml (Novartis). $T_H1$ polarization was initiated with 10 ng/ml IL12 (R&D Systems) and $T_H2$ neutralizing antibody anti-IL4 (2 μg/ml). $T_H2$ polarization was induced by activation with Phytohaemagglutinin, PHA (4μg/mL) in the presence of IL-4 (R&D Systems) (10 ng/ml), and neutralizing antibodies to IFN-γ (2 μg/ml) and anti-IL12 (2 μg/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-γ 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 $\mu$g 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 $\beta$-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 $\mu$g 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 $\mu$l 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 $\mu$g/$\mu$l RNase A (Sigma) for 2 hours at 37 °C and with 0.04 $\mu$g/$\mu$l 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 $\mu$l 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 $\mu$l o/n at 37°C. Then fragments were ligated with 400U T4 DNA ligase o/n at room temperature in 40 $\mu$l. 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".

## 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.

## 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$_H$1 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$_H$1 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).

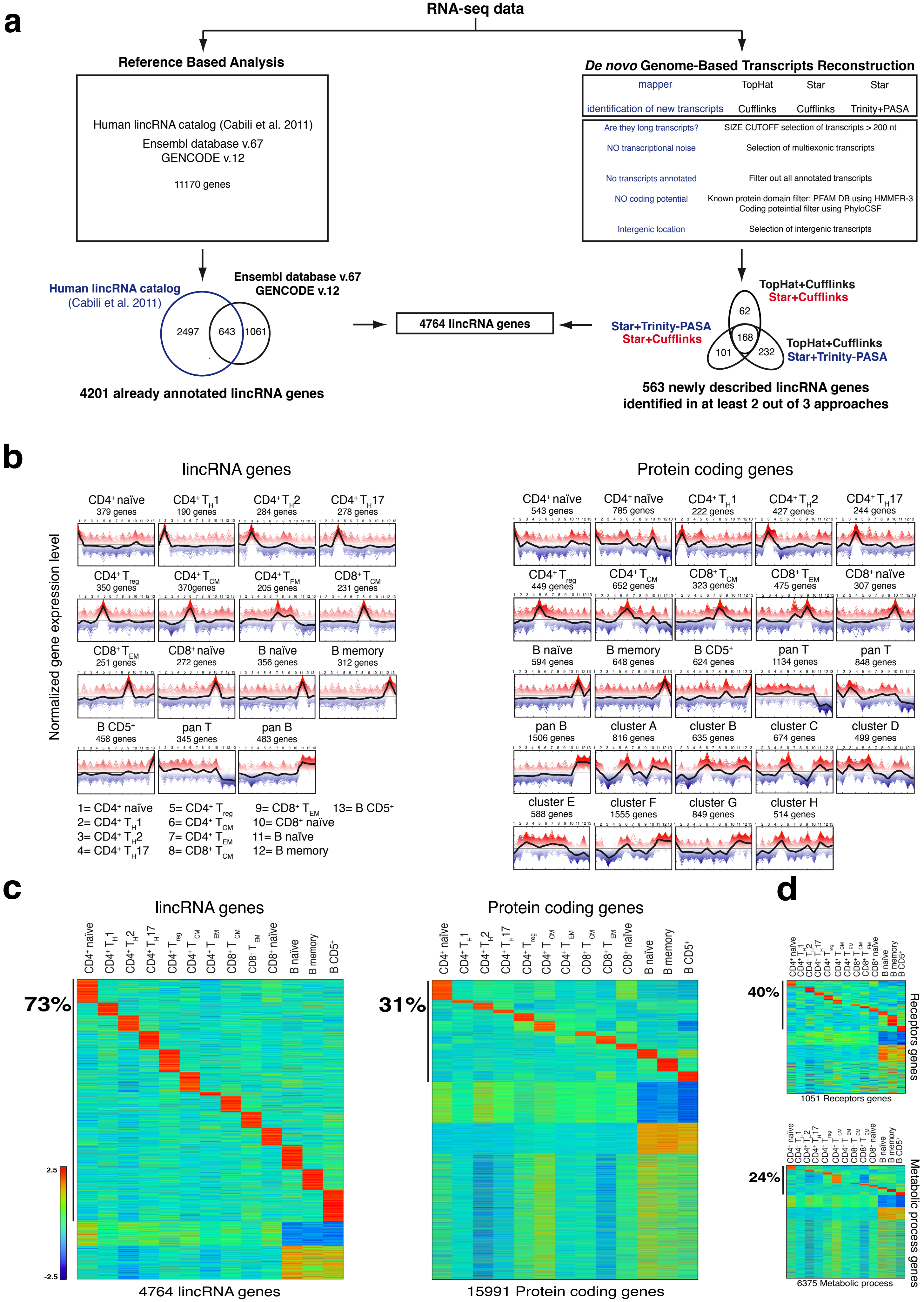
(d) RIP assay for LSD1 and EZH2 in CD4$^+$ naïve T cells differentiated in T$_H$1 polarizing conditions (day 8). The enrichment of linc-MAF-4 is relative to mock. $\beta$-actin, RNU2.1 and a region upstream the TSS of linc-MAF-4 were chosen as controls (average of six independent experiments ± SEM).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

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

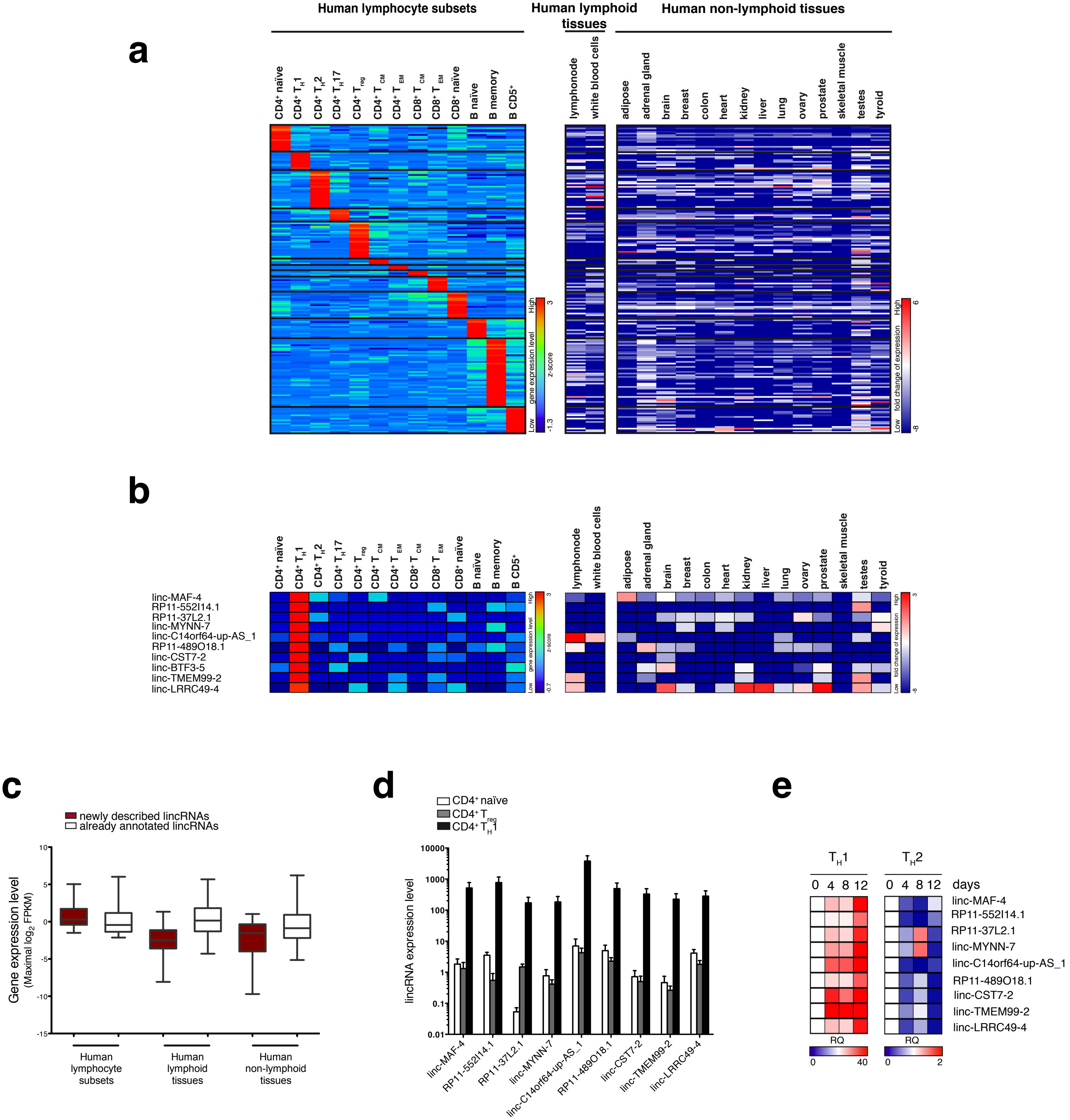
(f) Model for linc-MAF-4-mediated *MAF* repression in $T_H1$ lymphocytes. When linc-MAF-4 is expressed, it recruits chromatin remodelers (i.e. LSD1 and EZH2) at *MAF* 5'-end, taking advantage of a DNA loop that brings in close proximity *linc-MAF-4* 5'- and 3'- end and *MAF* 5'-end. This event causes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

| **Subset** | **Purity (%)** | **Sorting phenotype** | **Genes** |
|---|---|---|---|
| CD4$^+$ naïve | 99,8 ± 0,1 | CD4$^+$ CCR7$^+$ CD45RA$^+$ CD45RO$^-$ | 20061 |
| CD4$^+$ T$_H$1 | 99,9 ± 0,05 | CD4$^+$ CXCR3$^+$ | 20855 |
| CD4$^+$ T$_H$2 | 99,7 ± 0,3 | CD4$^+$ CRTH2$^+$ CXCR3$^-$ | 19623 |
| CD4$^+$ T$_H$17 | 99,1 ± 1 | CD4$^+$ CCR6$^+$ CD161$^+$ CXCR3$^-$ | 20959 |
| CD4$^+$ T$_{reg}$ | 99,0 ± 0,8 | CD4$^+$ CD127$^-$ CD25$^+$ | 21435 |
| CD4$^+$ T$_{CM}$ | 98,4 ± 2,8 | CD4$^+$ CCR7$^+$ CD45RA$^-$ CD45RO$^+$ | 20600 |
| CD4$^+$ T$_{EM}$ | 95,4 ± 5,5 | CD4$^+$ CCR7$^-$ CD45RA$^-$ CD45RO$^+$ | 19800 |
| CD8$^+$ T$_{CM}$ | 98,3 ± 0,8 | CD8$^+$ CCR7$^+$ CD45RA$^-$ CD45RO$^+$ | 20901 |
| CD8$^+$ T$_{EM}$ | 96,8 ± 0,9 | CD8$^+$ CCR7$^-$ CD45RA$^-$ CD45RO$^+$ | 21813 |
| CD8$^+$ naïve | 99,3 ± 0,2 | CD8$^+$ CCR7$^+$ CD45RA$^+$ CD45RO$^-$ | 20611 |
| B naïve | 99,9 ± 0,1 | CD19$^+$ CD5$^-$ CD27$^-$ | 21692 |
| B memory | 99,1 ± 0,8 | CD19$^+$ CD5$^-$ CD27$^+$ | 21239 |
| B CD5$^+$ | 99,1 ± 0,8 | CD19$^+$ CD5$^+$ | 22499 |

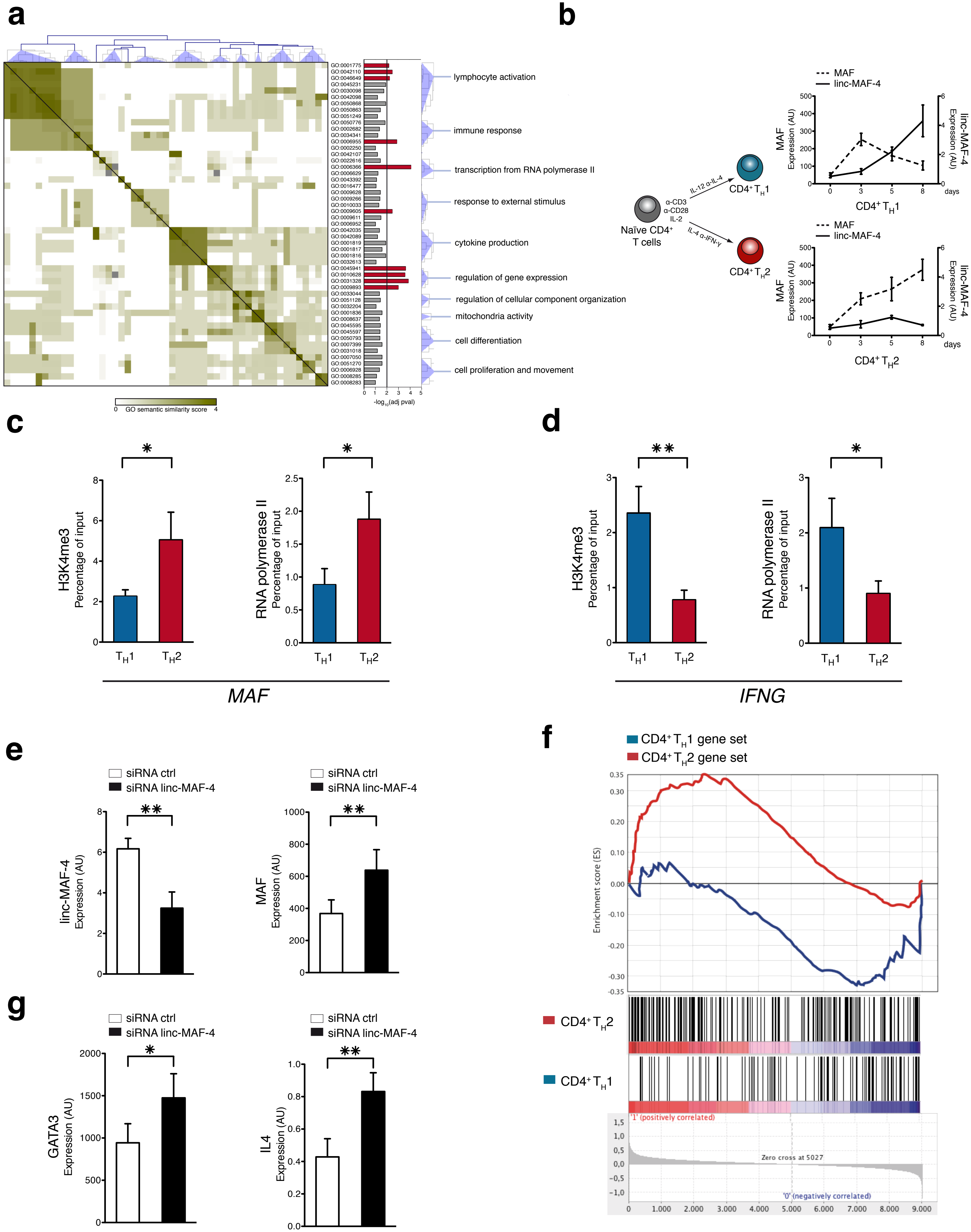
# Figure 1


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
1 2 3 4 5 6 7 8 9 10 11 12 13
c
lincRNA genes
Protein coding genes
CD4+ naïve
CD4+ TH1
CD4+ TH2
CD4+ TH17
CD4+ Treg
CD4+ TCM
CD4+ TEM
CD8+ TCM
CD8+ TEM
CD8+ naïve
B naïve
B memory
B CD5+
73%
31%
2.5
-2.5
4764 lincRNA genes
15991 Protein coding genes
d
40%
Receptors genes
1051 Receptors genes
24%
Metabolic process genes
6375 Metabolic process

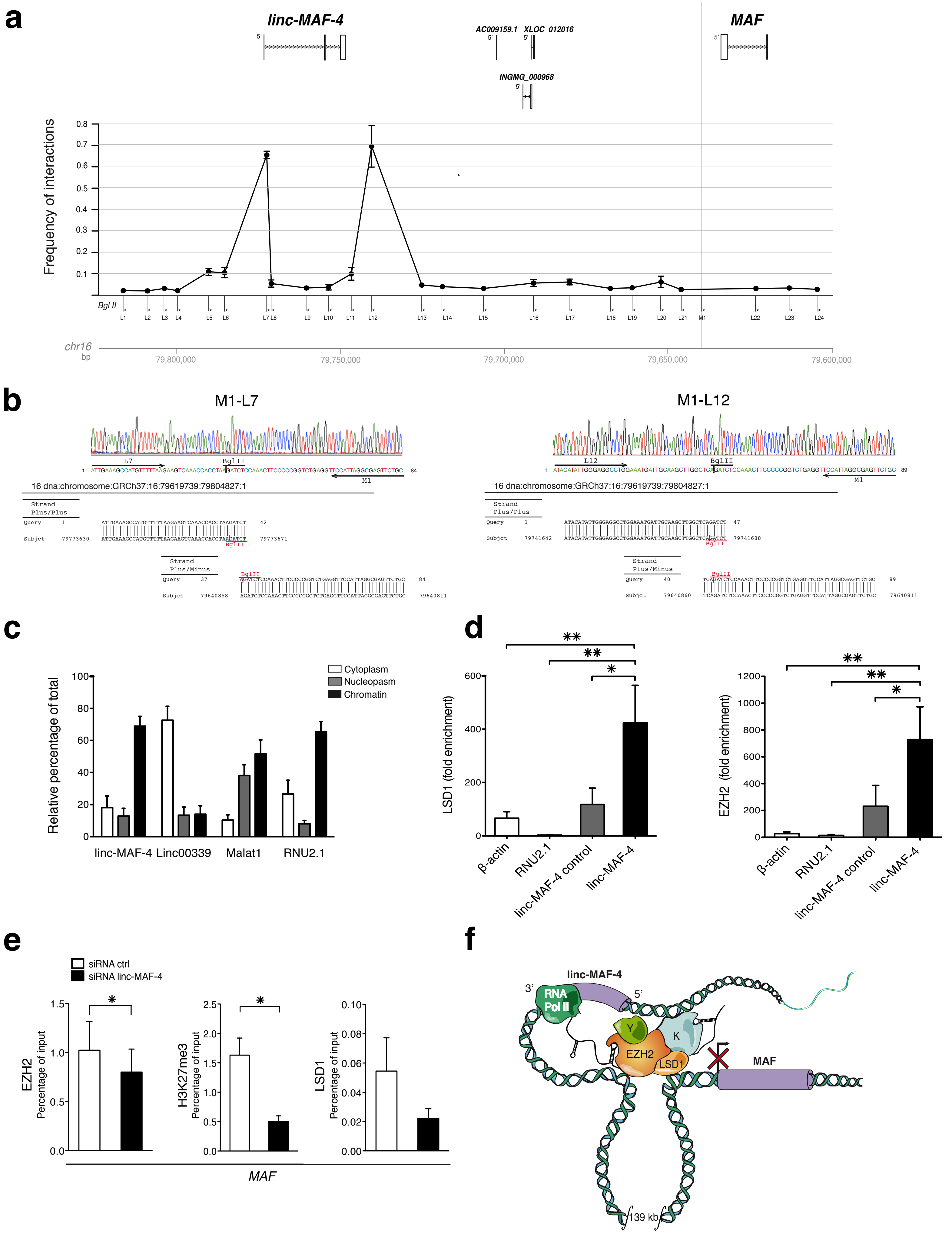
# Figure 2


Human lymphocyte subsets
Human lymphoid tissues
Human non-lymphoid tissues
a
CD4⁺ naïve
CD4⁺ T_H1
CD4⁺ T_H2
CD4⁺ T_H17
CD4⁺ T_reg
CD4⁺ T CM
CD4⁺ T EM
CD8⁺ T CM
CD8⁺ T EM
CD8⁺ naïve
B naïve
B memory
B CD5⁺
lymphonode
white blood cells
adipose
adrenal gland
brain
breast
colon
heart
kidney
liver
lung
ovary
prostate
skeletal muscle
testes
tyroid
High
Low
gene expression level
z-score
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
3
-0.7
3
-8
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⁺ T_reg
CD4⁺ T_H1
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
T_H1
T_H2
0 4 8 12
0 4 8 12
days
linc-MAF-4
RP11-552I14.1
RP11-37L2.1
linc-MYNN-7
linc-C14orf64-up-AS_1
RP11-489O18.1
linc-CST7-2
linc-TMEM99-2
linc-LRRC49-4
RQ
0
40
RQ
0
2

# Figure 3


a
GO:0001775
GO:0042110
GO:0046649
GO:0045231
GO:0030098
GO:0042098
GO:0050868
GO:0050863
GO:0051249
GO:0050776
GO:0002682
GO:0034341
GO:0006955
GO:0002250
GO:0042107
GO:0022616
GO:0006366
GO:0006629
GO:0043392
GO:0016477
GO:0009628
GO:0009266
GO:0010033
GO:0009605
GO:0009611
GO:0006952
GO:0042035
GO:0042089
GO:0001819
GO:0001817
GO:0001816
GO:0032613
GO:0045941
GO:0010628
GO:0031328
GO:0009893
GO:0033044
GO:0051128
GO:0032204
GO:0001836
GO:0008637
GO:0045595
GO:0045597
GO:0050793
GO:0007399
GO:0031018
GO:0007050
GO:0051270
GO:0006928
GO:0008285
GO:0008283
0 1 2 3 4 5
-log$_{10}$(adj pval)
0 GO semantic similarity score 4
lymphocyte activation
immune response
transcription from RNA polymerase II
response to external stimulus
cytokine production
regulation of gene expression
regulation of cellular component organization
mitochondria activity
cell differentiation
cell proliferation and movement
b
Naïve CD4$^+$ T cells
α-CD3
α-CD28
IL-2
IL-12 α-IL-4
IL-4 α-IFN-γ
CD4$^+$ T$_H$1
CD4$^+$ T$_H$2
MAF
linc-MAF-4
MAF Expression (AU)
linc-MAF-4 Expression (AU)
0 100 200 300 400 500
0 2 4 6
0 3 5 8
days
CD4$^+$ T$_H$1
MAF
linc-MAF-4
MAF Expression (AU)
linc-MAF-4 Expression (AU)
0 100 200 300 400 500
0 2 4 6
0 3 5 8
days
CD4$^+$ T$_H$2
c
*
H3K4me3 Percentage of input
0 2 4 6 8
T$_H$1 T$_H$2
*
RNA polymerase II Percentage of input
0.0 0.5 1.0 1.5 2.0 2.5
T$_H$1 T$_H$2
*MAF*
d
**
H3K4me3 Percentage of input
0 1 2 3
T$_H$1 T$_H$2
*
RNA polymerase II Percentage of input
0 1 2 3
T$_H$1 T$_H$2
*IFNG*
e
siRNA ctrl
siRNA linc-MAF-4
**
linc-MAF-4 Expression (AU)
0 2 4 6 8
siRNA ctrl
siRNA linc-MAF-4
**
MAF Expression (AU)
0 200 400 600 800 1000
f
CD4$^+$ T$_H$1 gene set
CD4$^+$ T$_H$2 gene set
Enrichment score (ES)
0.35 0.30 0.20 0.10 0.00 -0.10 -0.20 -0.30 -0.35
CD4$^+$ T$_H$2
CD4$^+$ T$_H$1
'1' (positively correlated)
1,5 1,0 0,5 0,0 -0,5 -1,0
Zero cross at 5027
'0' (negatively correlated)
0 1.000 2.000 3.000 4.000 5.000 6.000 7.000 8.000 9.000
g
siRNA ctrl
siRNA linc-MAF-4
*
GATA3 Expression (AU)
0 500 1000 1500 2000
siRNA ctrl
siRNA linc-MAF-4
**
IL4 Expression (AU)
0.0 0.2 0.4 0.6 0.8 1.0

# Figure 4



M1-L7

```
1 ATTGAAAGCCATGTTTTTAAGAAGTCAAACCACCTAAGATCTCCAAACTTCCCCCGGTCTGAGGTTCCATTAGGCGAGTTCTGC 84
```

16 dna:chromosome:GRCh37:16:79619739:79804827:1

```
 Strand
 Plus/Plus
Query     1          ATTGAAAGCCATGTTTTTAAGAAGTCAAACCACCTAAGATCT    42
                     ||||||||||||||||||||||||||||||||||||||||||
Subjct    79773630   ATTGAAAGCCATGTTTTTAAGAAGTCAAACCACCTAAGATCT    79773671
                                                         BglII

                          Strand
                          Plus/Minus
                                            BglII
                          Query     37      AGATCTCCAAACTTCCCCCGGTCTGAGGTTCCATTAGGCGAGTTCTGC    84
                                            ||||||||||||||||||||||||||||||||||||||||||||||||
                          Subjct    79640858 AGATCTCCAAACTTCCCCCGGTCTGAGGTTCCATTAGGCGAGTTCTGC   79640811
```

M1-L12

```
1 ATACATATTGGGAGGCCTGGAAATGATTGCAAGCTTGGCTCAGATCTCCAAACTTCCCCCGGTCTGAGGTTCCATTAGGCGAGTTCTGC 89
```

16 dna:chromosome:GRCh37:16:79619739:79804827:1

```
 Strand
 Plus/Plus
Query     1          ATACATATTGGGAGGCCTGGAAATGATTGCAAGCTTGGCTCAGATCT  47
                     |||||||||||||||||||||||||||||||||||||||||||||||
Subjct    79741642   ATACATATTGGGAGGCCTGGAAATGATTGCAAGCTTGGCTCAGATCT  79741688
                                                              BglII

                          Strand
                          Plus/Minus
                                               BglII
                          Query     40      TCAGATCTCCAAACTTCCCCCGGTCTGAGGTTCCATTAGGCGAGTTCTGC    89
                                            ||||||||||||||||||||||||||||||||||||||||||||||||||
                          Subjct    79640860 TCAGATCTCCAAACTTCCCCCGGTCTGAGGTTCCATTAGGCGAGTTCTGC   79640811
```