nature immunology

The long intergenic noncoding RNA landscape of human lymphocytes highlights the 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 J P 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]

Long noncoding RNAs are emerging as important regulators of cellular functions, but little is known of their role in the human immune system. Here we investigated long intergenic noncoding RNAs (lincRNAs) in 13 subsets of T lymphocytes and B lymphocytes by next-generation sequencing–based RNA sequencing (RNA-seq analysis) and *de novo* transcriptome reconstruction. We identified over 500 previously unknown lincRNAs and described lincRNA signatures. Expression of linc-MAF-4, a chromatin-associated lincRNA specific to the T_H1 subset of helper T cells, was inversely correlated with expression of MAF, a T_H2-associated transcription factor. Downregulation of linc-MAF-4 skewed T cell differentiation toward the T_H2 phenotype. We identified a long-distance interaction between the genomic regions of the gene encoding linc-MAF-4 and *MAF*, where linc-MAF-4 associated with the chromatin modifiers LSD1 and EZH2; this suggested that linc-MAF-4 regulated *MAF* transcription through the recruitment of chromatin modifiers. Our results demonstrate a key role for lincRNA in T lymphocyte differentiation.

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

In addition to the well-established role of transcription factors as instructive signals for cell differentiation toward a given lineage, other cues, such as epigenetic modifications, can regulate the maintenance of cellular states[4]. In this context, noncoding RNAs are emerging as a new regulatory layer that affects both the development of the immune system and its function[5,6]. Among the several classes of noncoding RNAs with a specific role in lymphocyte biology, microRNAs are the best characterized[7–11]. Although thousands of long intergenic noncoding RNAs (lincRNAs) have been identified in the mammalian genome by bioinformatics analyses of transcriptomic data[12–14], 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[16–18], whereas cytoplasmic lincRNAs can modulate translational control[19] and transcript stability[20] directly by base-pairing with specific targets or indirectly as competing endogenous RNAs[21–23]. A few examples of functional lincRNAs in the mouse immune system have been described. A broad analysis investigating naive and memory $CD8^+$ cells purified from mouse spleen with a custom array of lincRNAs has reported the identification of 96 lymphoid-specific lincRNAs and has suggested a role for lincRNAs in the differentiation and activation of lymphocytes[24]. The lincRNA NeST has been found to be down-regulated during lymphocyte activation in a manner reciprocal to the expression of interferon-γ (IFN-γ) and to control susceptibility to infection with Theiler's virus and salmonella in mice through epigenetic regulation of the *Ifng* locus[25,26]. Subsequently, mouse lincRNA-Cox2 was reported to be induced downstream of signaling via Toll-like receptors and to mediate the activation and repression of distinct sets of genes that are targets of the immune system that encode molecules involved in inflammatory responses[27]. Another study of mouse thymocytes and mature peripheral T cells has allowed the identification of lincRNAs with specific expression patterns

[1]Istituto Nazionale Genetica Molecolare 'Romeo ed Enrica Invernizzi', Milano, Italy. [2]IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy. [3]These authors contributed equally to this work. Correspondence should be addressed to M.P. (pagani@ingm.org) or S.A. (abrignani@ingm.org).

Received 25 February 2014; accepted 18 December 2014; published online 26 January 2015; doi:10.1038/ni.3093

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

Subset	Purity (%)	Sorting phenotype	Genes
$CD4^+$ naive	99.8 ± 0.1	$CD4^+CCR7^+CD45RA^+CD45RO^-$	20,061
$CD4^+$ T_H1	99.9 ± 0.05	$CD4^+CXCR3^+$	20,855
$CD4^+$ T_H2	99.7 ± 0.3	$CD4^+CRTH2^+CXCR3^-$	19,623
$CD4^+$ T_H17	99.1 ± 1	$CD4^+CCR6^+CD161^+CXCR3^-$	20,959
$CD4^+$ T_{reg}	99.0 ± 0.8	$CD4^+CD127^-CD25^+$	21,435
$CD4^+$ T_{CM}	98.4 ± 2.8	$CD4^+$ $CCR7^+$ $CD45RA^-$ $CD45RO^+$	20,600
$CD4^+$ T_{EM}	95.4 ± 5.5	$CD4^+$ $CCR7^-$ $CD45RA^-$ $CD45RO^+$	19,800
$CD8^+$ T_{CM}	98.3 ± 0.8	$CD8^+$ $CCR7^+$ $CD45RA^-$ $CD45RO^+$	20,901
$CD8^+$ T_{EM}	96.8 ± 0.9	$CD8^+$ $CCR7^-$ $CD45RA^-$ $CD45RO^+$	21,813
$CD8^+$ naive	99.3 ± 0.2	$CD8^+$ $CCR7^+$ $CD45RA^+$ $CD45RO^-$	20,611
Naive B	99.9 ± 0.1	$CD19^+$ $CD5^-$ $CD27^-$	21,692
Memory B	99.1 ± 0.8	$CD19^+$ $CD5^-$ $CD27^+$	21,239
$CD5^+$ B	99.1 ± 0.8	$CD19^+$ $CD5^+$	22,499

Purity achieved (middle left) by the sorting of 13 human lymphocyte subsets (isolated from peripheral blood lymphocytes of four to five different donors per subset) by various surface marker combinations (Sorting phenotype), as well as number of genes expressed (FPKM > 0.21) in at least one sample for each subset (far right). T_{reg}, regulatory T cells; T_{CM}, central memory T cells; T_{EM} effector memory T cells; B, B cells. Data are representative of at least four experiments (mean ± s.d. for purity).

during T cell differentiation and of LincR-Ccr2-5′AS, a lincRNA specific to $CD4^+$ T helper type 2 cells (T_H2 cells) that is involved in regulating the migration of $CD4^+$ T_H2 lymphocytes[28]. Although such studies highlight the relevance of lincRNAs in regulating immune responses, a thorough analysis of their expression profile and function in the human immune system is still lacking.

The present study was based on the analysis of 13 highly purified primary human lymphocytes subsets by high-throughput sequencing technologies for cDNA (RNA-seq analysis). We performed *de novo* transcriptome reconstruction (the creation of a transcriptome without the aid of a reference genome)[29] and discovered over 500 previously unknown long intergenic noncoding RNAs (lincRNAs). We identified several lymphocyte subset–specific lincRNA signatures and found that expression of linc-MAF-4, a chromatin-associated, $CD4^+$ T_H1 cell–specific lincRNA, correlated inversely with expression of the transcription factor c-Maf and that its downregulation skewed the differentiation of $CD4^+$ T cells toward the T_H2 phenotype. We provide the first comprehensive inventory, to our knowledge, of human lymphocyte lincRNAs and demonstrate that lincRNAs can be key to lymphocyte differentiation. This resource will probably help in providing a better definition of the role of lincRNAs in the differentiation, plasticity and effector functions of lymphocytes.

RESULTS

Discrimination of human lymphocyte subsets by lincRNAs

To assess lincRNA expression in human primary lymphocytes, we extracted RNA from 13 lymphocyte cell subsets (**Table 1**) purified from peripheral blood mononuclear cells from five healthy donors[11]. We then analyzed the polyadenylated RNA fraction by paired-end RNA sequencing and obtained about 1.7×10^9 mapped 'reads'. To enrich for transcripts derived from true active genes, we applied an expression threshold of 0.21 FPKM (fragments per kilobases of exons per million fragments mapped), defined through the integration of RNA-seq data and data on chromatin states from the ENCODE (Encyclopedia of DNA Elements) project[30]. We found a total of 31,902 expressed genes (including both protein-coding genes and noncoding genes) in the 13 subsets (**Table 1** and **Supplementary Fig. 1a**), of which 4,201 were lincRNAs annotated in public resources[12,31] (**Fig. 1**). To identify previously unknown lincRNAs expressed in primary human lymphocytes, we used three *de novo* transcriptome–reconstruction strategies based on the combination of two different sequence mappers, TopHat and Star[32,33], with two different tools for *de novo* transcript assembly, Cufflinks and Trinity[34,35]. We identified lincRNAs among the newly described transcripts by exploiting the following process. We selected transcripts that were longer than 200 nucleotides and multiexonic that did not overlap with protein-coding genes (and thus excluded unreliable single-exon fragments assembled by RNA-seq). We excluded transcripts with a conserved protein-coding region and those with open reading frames encoding protein domains catalogued in the Pfam database of protein families[36]. We used PhyloCSF, a comparative genomics method that assesses multispecies nucleotide-sequence alignment on the basis of a formal statistical comparison of phylogenetic codon models[37], which efficiently identifies noncoding RNAs, as demonstrated by ribosome-profiling experiments[38]. Finally, we defined a stringent new lincRNA set that included those genes for which at least one lincRNA isoform was reconstructed by two assemblers of three. Through this conservatively multilayered analysis we identified 563 previously unknown lincRNA-encoding genes, which increased by 11.8% the number of lincRNAs known to be expressed in human lymphocytes. The various classes of RNAs were evenly distributed among various lymphocyte subsets (**Supplementary Fig. 1b**), and the ratio of already annotated and newly identified

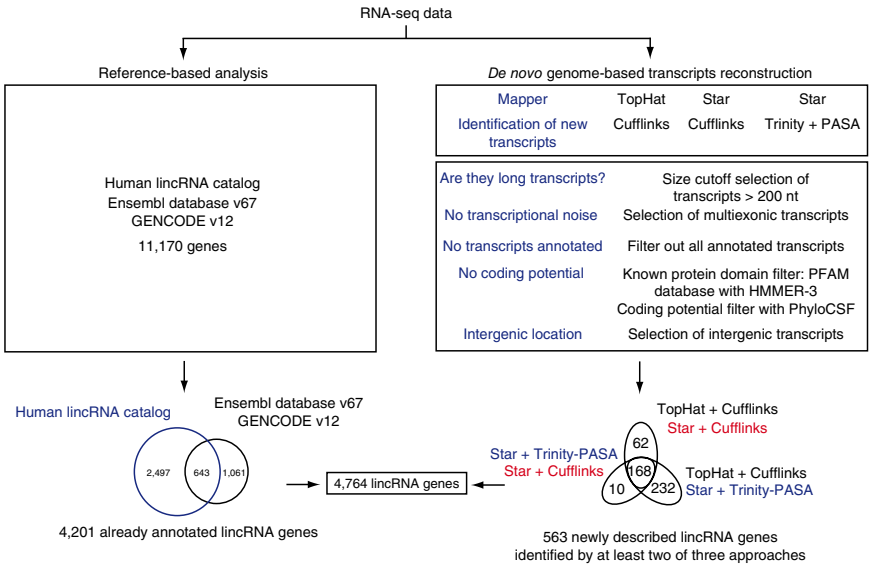

Figure 1 Identification of lincRNAs expressed in human lymphocyte subsets. For the identification of lincRNAs, RNA-seq data generated from 63 lymphocyte samples were processed by the quantification of lincRNAs already annotated in public resources (left) and by *de novo* genome-based transcripts reconstruction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

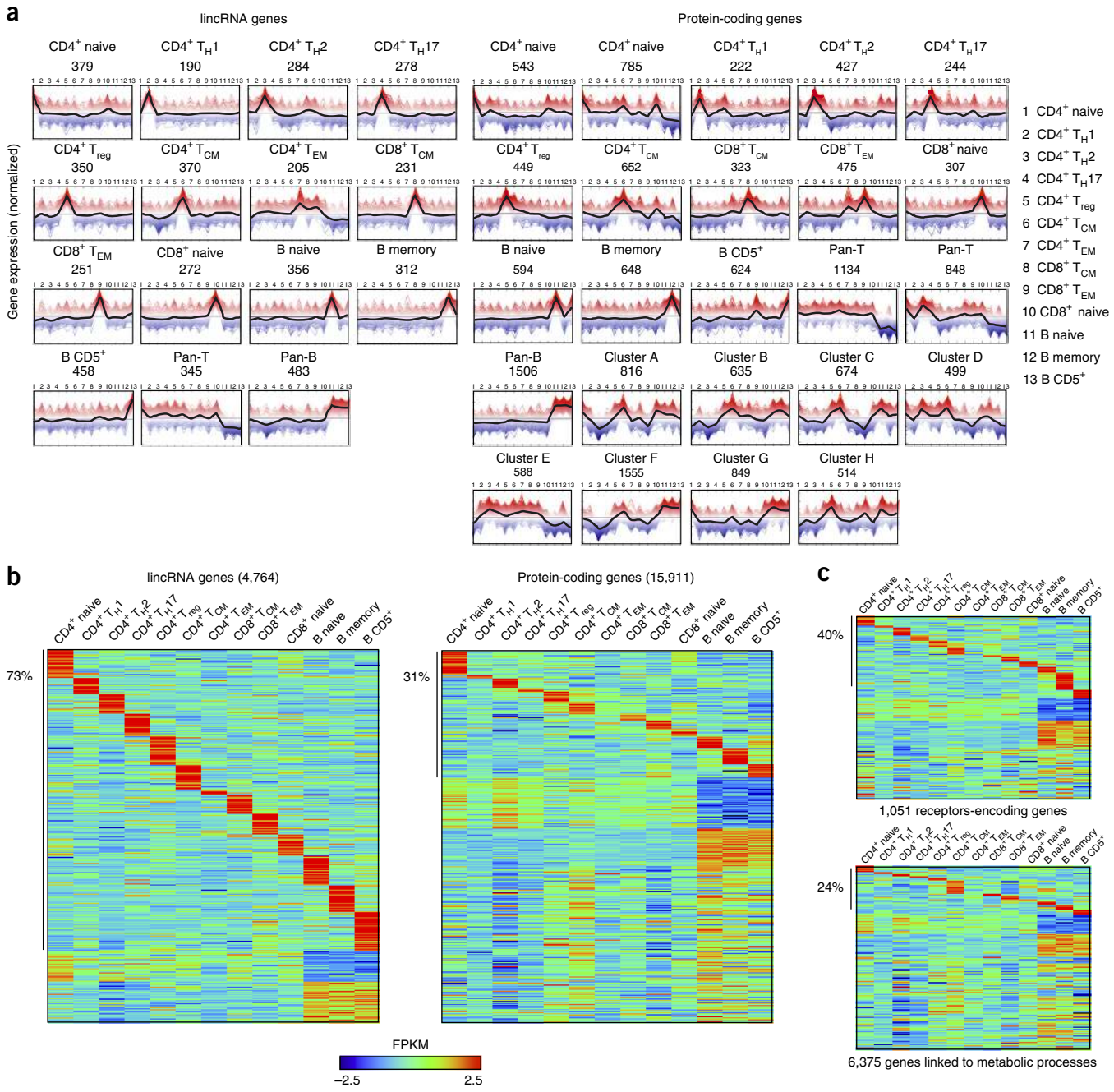

Figure 2 Definition of gene clusters in human lymphocytes. (**a**) Expression profiles of lincRNA and protein-coding genes across 13 human lymphocyte subsets (key at right; numbers above profile peaks correspond to key) according to K-means cluster definition. Black lines indicate mean expression of genes belonging to the same cluster. (**b**) Specificity of lincRNA-encoding genes (left) and protein-coding genes (right) across 13 human lymphocyte populations (above columns); order of rows and columns based on K-means clustering; color intensity (key) indicates *z*-score $\log_2$-normalized raw FPKM counts estimated by Cufflinks software; numbers at top left indicate percent assigned to specific clusters (additional information, **Supplementary Fig. 2a**). (**c**) Analysis as in **b** of genes encoding receptors (top) and molecules involved in metabolic processes (bottom).

lincRNAs was similar across different chromosomes (**Supplementary Fig. 1c**) and across various lymphocyte subsets (**Supplementary Fig. 1d**). As observed in various cell types[12,34], lincRNAs were also generally expressed at lower abundance than were protein-coding genes in human lymphocytes (**Supplementary Fig. 1e**). However, when we categorized transcripts on the basis of their cell-specific expression and non–cell-specific expression (**Supplementary Fig. 1f**), we found that cell-specific lincRNAs and cell-specific protein-coding genes displayed similar expression levels (**Supplementary Fig. 1e–g**).

Lymphocytes subsets display very different migratory abilities and effector functions, yet they are very closely related from the differentiation point of view. As lincRNAs are generally more tissue specific than are protein-coding genes[12,39], we assessed the lymphocyte cell–subset specificity of lincRNAs. We therefore classified genes according to their expression profiles by unsupervised K-means clustering and found that lincRNAs were defined by 15 clusters and protein-coding genes were defined by 24 clusters (**Fig. 2a** and **Supplementary Fig. 2a**). Notably, the frequency of genes assigned to the clusters specific

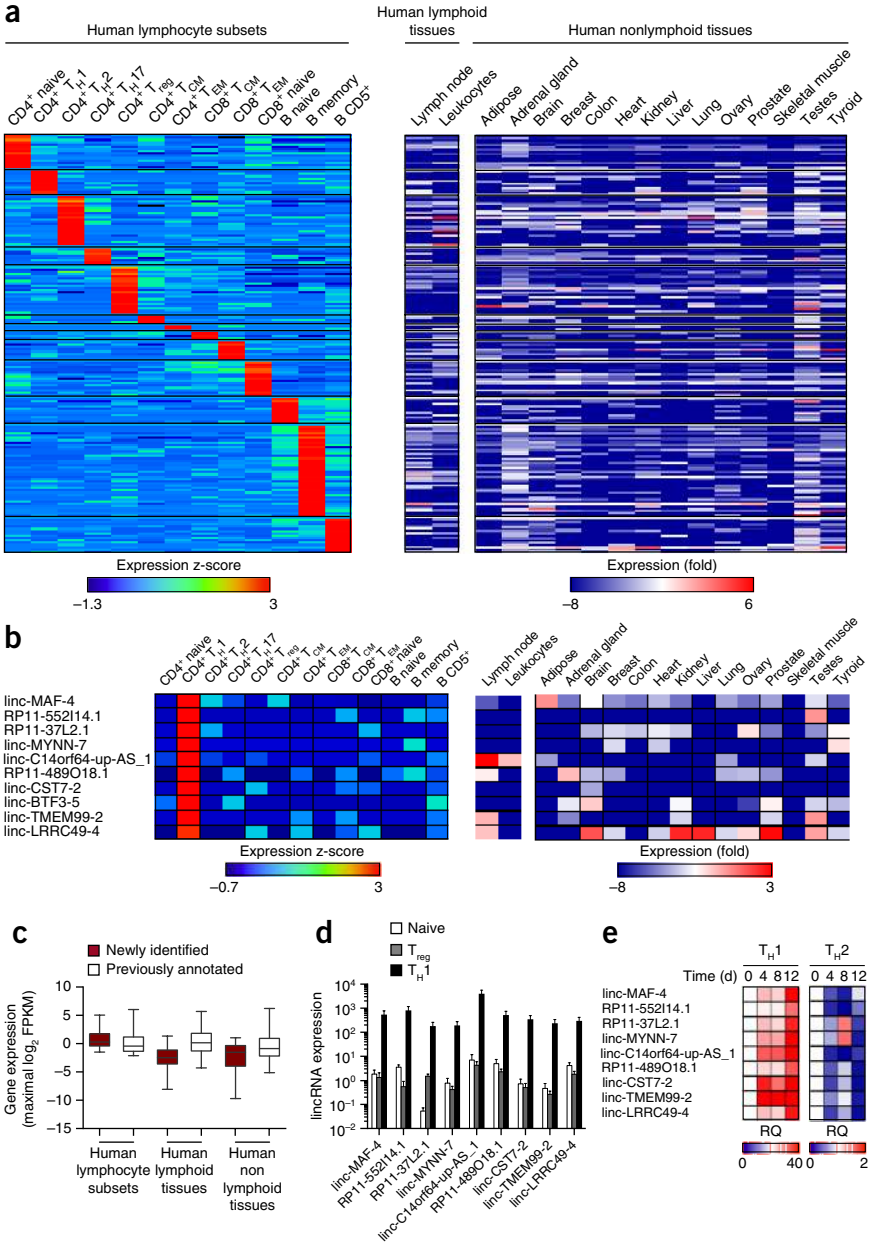

Figure 3 LincRNA signatures of human lymphocyte subsets. (**a**) Heat map of normalized expression values of lymphocyte signature lincRNAs selected on the basis of a difference in expression of >2.5-fold (relative to expression in all other subsets), intrapopulation consistency (expressed in at least three of five samples) and a *P* value of <0.05 (nonparametric Kruskal-Wallis test); normalized expression values were calculated as $\log_2$ ratio between expression in the lymphocyte subsets and in a panel of human lymphoid and nonlymphoid tissues of the Human BodyMap 2.0 project (additional information, **Supplementary Table 1**). (**b**) Expression of $CD4^+$ T_H1 cell signature lincRNAs (presented as in **a**). S, sense; AS, antisense. (**c**) Expression of newly identified and previously annotated lincRNAs (key) in human lymphocyte subsets and lymphoid or nonlymphoid human tissues (presented as the 2.5–97.5 percentile). (**d**) Quantitative RT-PCR analysis of the expression of T_H1 cell signature lincRNAs by primary $CD4^+$ naive cells (Naive), regulatory T cells (T_{reg}) and T_H1 cells (T_H1) sorted from the peripheral blood mononuclear cells of healthy donors. (**e**) Quantitative RT-PCR analysis of the expression of T_H1 cell signature lincRNAs over time in $CD4^+$ naive T cells differentiated in T_H1- or T_H2-polarizing conditions; results (average values) are presented as relative quantity (RQ) relative to expression at time zero. Data are from at least four experiments (**a**,**b**), one experiment with 63 independent samples (**c**), three independent experiments (**d**; average ± s.e.m.) or two independent experiments (**e**).

for the various lymphocyte subsets was higher for lincRNAs (71%) than for protein-coding genes (34%) (**Fig. 2b**). This superiority stood out even when we compared lincRNAs with genes encoding membrane receptors (40%) (**Fig. 2c**), which are generally considered the most accurate markers of various lymphocyte subsets. We obtained similar results with the heuristic expression threshold of FPKM > 1 (**Supplementary Fig. 2b**). Thus, by RNA-seq analyses of highly purified subsets of primary T lymphocytes and B lymphocytes, we were able to provide a comprehensive landscape of lincRNA expression in human lymphocytes. By exploiting *de novo* transcriptome reconstruction, we discovered 563 previously unknown lincRNAs and found that lincRNAs were effective in marking lymphocyte identity.

Identification of lincRNA signatures in lymphocytes

Next we investigated our data set for the presence of lincRNA signatures in the various lymphocyte subsets. We therefore looked for lincRNAs with a difference in expression of more than 2.5-fold in a given cell subset relative to their expression in all the other subsets ($P < 0.05$ (nonparametric Kruskal-Wallis test)) that were expressed in at least three of five donors and found 172 lincRNAs that met these criteria (**Fig. 3a** and **Supplementary Table 1**). We integrated the human transcriptome database with our newly identified transcripts and thus created a new reference with which to assess more thoroughly their expression in other human tissues. Assessing lincRNA signatures in a panel of 16 human tissues (from the Human BodyMap 2.0 project), we found that not only were lymphocyte signature lincRNAs expressed very poorly in nonlymphoid tissues but also most signature lincRNAs were not detectable even in lymphoid tissues (**Fig. 3a**,**b**). These findings emphasized the importance of assessing the expression of lincRNAs (as well as of any highly cell-specific transcripts) in purified primary cells rather than in total tissues in which a given cell subset–specific transcript is diluted by the transcripts of all the other cell types of the tissue. We note that the newly identified lincRNAs defined as signatures were more abundant (**Fig. 3c**) and more cell specific (**Supplementary Table 1**) than the already annotated lincRNAs defined as signatures. We present here data obtained from the $CD4^+$ T_H1 cell subset (**Fig. 2b**); we obtained similar results for all the other subsets (**Supplementary Table 1**).

Finally, to confirm and extend our signature data, we assessed expression of the signature lincRNAs of $CD4^+$ T_H1 cells (**Fig. 3b**) by

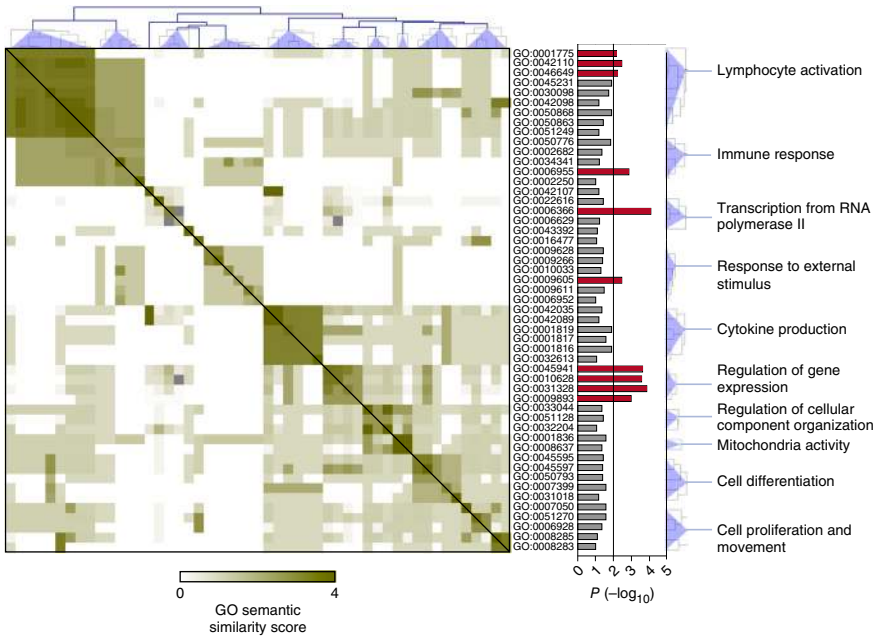

Figure 4 Gene-ontology semantic similarity matrix of 'protein-coding' genes proximal to lincRNA signatures. Semantic similarity scores for all gene-ontology (GO) term pairs clustered by a hierarchical clustering method (left), with adjusted *P* values for each GO term (middle), as well as common ancestors (right); red bars indicate gene-ontology terms with significant enrichment.

quantitative RT-PCR of a new set of independent samples of primary human CD4$^+$ naive cells, regulatory T cells and T_H1 cells, as well as in naive CD4$^+$ T cells that were activated *in vitro* and induced to differentiate toward the T_H1 or T_H2 phenotype. We confirmed specific-subset expression for 90% of the CD4$^+$ T_H1 cell signature lincRNAs (**Fig. 3d**). Moreover, 90% of the CD4$^+$ T_H1 cell signature lincRNAs that were expressed in resting CD4$^+$ T_H1 cells purified *ex vivo* also had high expression in naive CD4$^+$ T cells differentiated under T_H1-polarizing conditions *in vitro*, whereas they had low expression in naive CD4$^+$ T cells differentiated toward the T_H2 phenotype *in vitro* (**Fig. 3e**). As a corollary to those findings, we observed by RNA-seq that the signature lincRNAs of CD4$^+$ naive cells were mostly downregulated during differentiation toward the T_H0 phenotype *in vitro*, whereas the signature lincRNAs of cells of the T_H1, T_H2 and T_H17 subsets of helper T cells were mostly upregulated (**Supplementary Fig. 3a**). Together our data demonstrated that lincRNAs provided signatures of human lymphocyte subsets and suggested that human CD4$^+$ T lymphocytes acquired most of their memory-specific lincRNA signatures during their activation-driven differentiation from naive cells to memory cells.

Downregulation of linc-MAF-4 skews CD4$^+$ T cells toward T_H2 cells

As lincRNAs have been reported to influence the expression of neighboring genes[25,26,28,40], we sought to determine whether protein-coding genes proximal to the signature lincRNAs of lymphocytes were involved in key cell functions. For this we used the FatiGO tool from the Babelomics suite for functional enrichment analysis[41] and found that

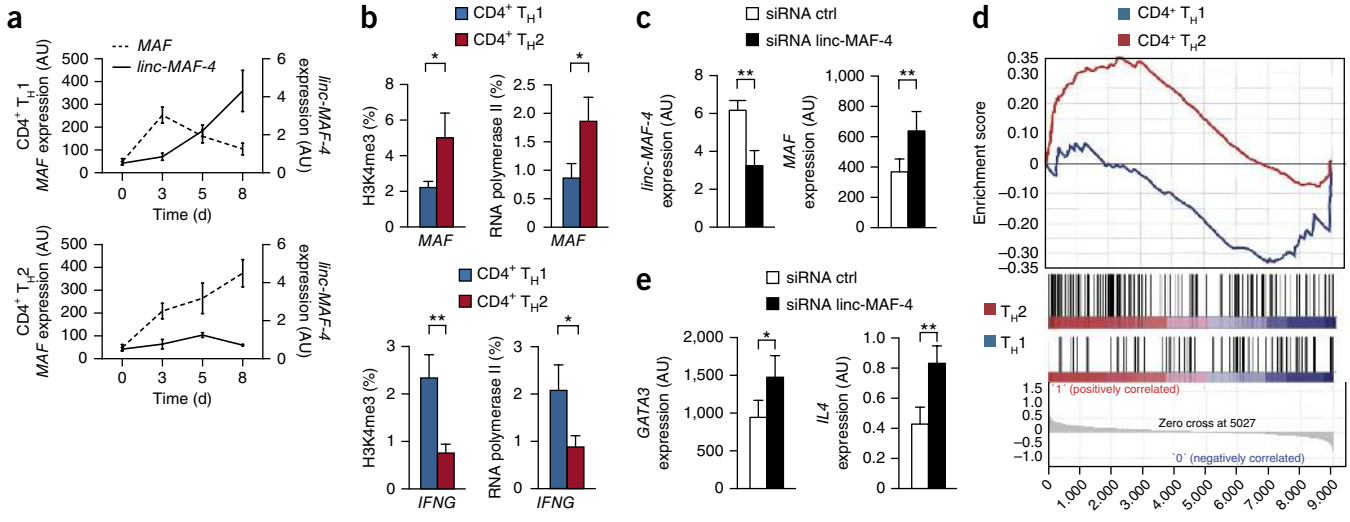

Figure 5 Linc-MAF-4 contributes to T_H1 differentiation. (**a**) Quantitative RT-PCR analysis of the expression of *linc-MAF-4* and *MAF* in activated CD4$^+$ naive T cells differentiated in T_H1- or T_H2-polarizing conditions (additional information, **Supplementary Fig. 4b**,**c**). AU, arbitrary units. (**b**) Occupancy of H3K4me3 and RNA polymerase II at the *MAF* locus (top) or the control *IFNG* locus (bottom) in CD4$^+$ naive T cells differentiated in T_H1- or T_H2-polarizing conditions at day 8 after activation, analyzed by chromatin immunoprecipitation followed by quantitative PCR and presented relative to input DNA. (**c**) Quantitative RT-PCR analysis of the expression of *linc-MAF-4* and *MAF* in activated CD4$^+$ naive T cells (in the absence of polarizing cytokines) 72 h after transfection of small interfering RNA (siRNA) targeting linc-MAF-4 or control (ctrl) siRNA. (**d**) Gene set–enrichment analysis, presented as enrichment score profiles for genes in activated CD4$^+$ naive T cells after transfection of siRNA targeting linc-MAF-4 or control siRNA compared with that of the CD4$^+$ T_H1 cell reference gene set or the T_H2 cell reference gene set, respectively. Nominal $P < 0.05$. (**e**) Quantitative RT-PCR analysis of the expression of *GATA3* and *IL4* transcripts in activated CD4$^+$ naive T cells transfected with siRNA as in **d**. $^*P < 0.05$ and $^{**}P < 0.01$ (one-tailed *t*-test). Data are representative of four independent experiments (**a**; average ± s.e.m.) or are from at least five (**b**, top) or ten (**b**, bottom) independent experiments (average and s.e.m.), six independent experiments (**c**,**e**; average and s.e.m.) or four independent experiments (**d**; average).

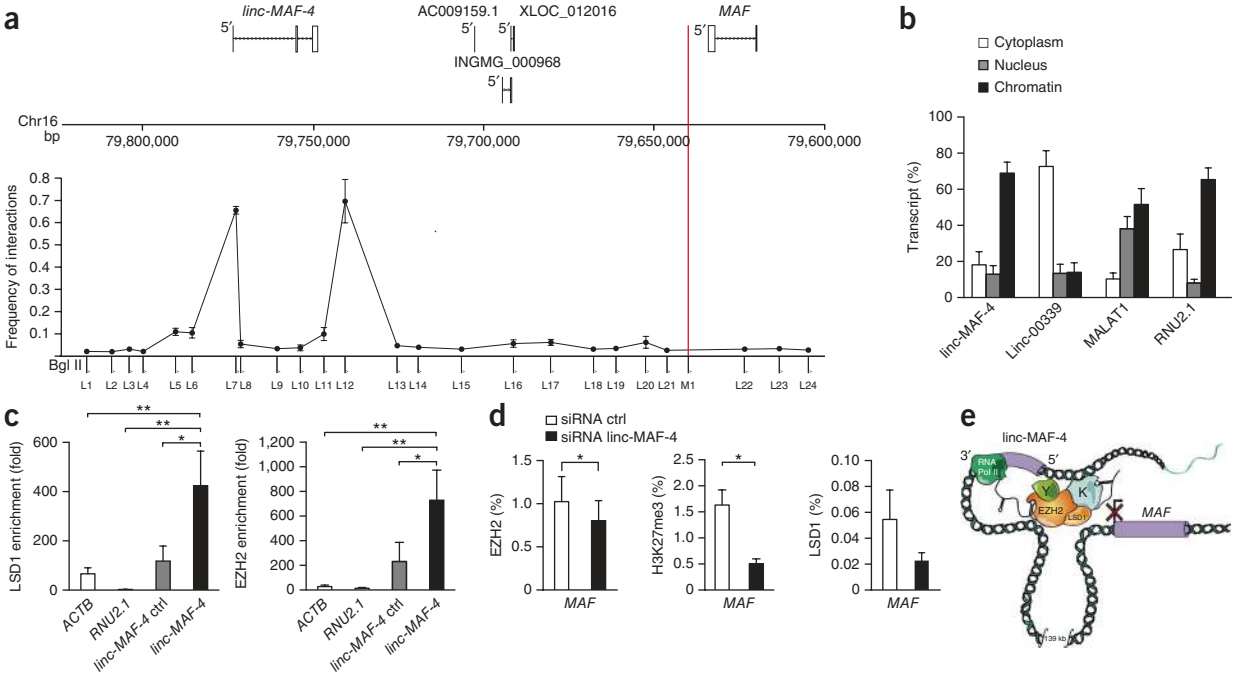

Figure 6 Epigenetic characterization of the *linc-MAF-4–MAF* genomic locus. (**a**) Chromosome-conformation capture analysis of the interactions between a 'bait' region M1 (red line) at the 5′ end of *MAF* and 24 'prey' regions spanning the *linc-MAF-4–MAF* genomic locus (L1–L24; horizontal axis) in CD4$^+$ naive T cells differentiated in T_H1-polarizing conditions at day 8 after activation. Top, organization of the genomic locus. (**b**) Abundance of linc-MAF-4 transcripts, as well as of linc-00339, MALAT1 and RNU2.1 (cytoplasmic, nuclear and chromatin-associated control transcripts, respectively), in the cytoplasm, nucleus and chromatin of CD4$^+$ naive T cells differentiated in T_H1-polarizing conditions at day 8 after activation. (**c**) RNA-immunoprecipitation assay of the interaction of LSD1 and EZH2 with *linc-MAF-4*, and with the controls *ACTB*, *RNU2.1* and a region upstream of the transcriptional start site of *linc-MAF-4* (*linc-MAF-4* ctrl), in CD4$^+$ naive T cells differentiated in T_H1-polarizing conditions at day 8 after activation; results are presented relative to control immunoprecipitation. $^*P < 0.05$ and $^{**}P < 0.01$ (analysis of variance and Dunnet post-hoc test). (**d**) Occupancy of EZH2, H3K27me3 and LSD1 at the *MAF* locus in activated CD4$^+$ naive T cells transfected with siRNA targeting linc-MAF-4 or control siRNA, analyzed by chromatin immunoprecipitation followed by quantitative PCR. $^*P < 0.05$ (one-tailed *t*-test). (**e**) Model for linc-MAF-4-mediated repression of *MAF* in T_H1 lymphocytes: when linc-MAF-4 is expressed, it recruits chromatin remodelers (i.e., LSD1 and EZH2) at the 5′ end of *MAF*, taking advantage of a DNA loop that brings the 5′ and 3′ ends of *linc-MAF-4* in close proximity to the 5′ end of *MAF*. Data are from three independent experiments (**a**,**b**; average and s.e.m.), six independent experiments (**c**; average and s.e.m.) or at least three independent experiments (**d**; average and s.e.m.).

protein-coding genes adjacent to signature lincRNAs showed enrichment for gene ontology terms correlated with the activation of lymphocyte T cells (**Fig. 4**), which indicated a possible role for signature lincRNAs in lymphocyte function. To obtain proof of concept of this hypothesis, we chose to characterize in depth linc-MAF-4 (lnc-MAF-2 in the LNCipedia database[42]), a signature lincRNA of T_H1 cells located 139.5 kilobases upstream of *MAF*. This gene encodes transcription factor c-Maf, which is involved in T_H2 differentiation[43] but is also required for the efficient development of T_H17 cells[44] and controls transcription of the gene encoding interleukin 4 in CD4$^+$ follicular helper T cells[45]. Our sequencing data showed that high expression of linc-MAF-4 correlated with a low abundance of *MAF* transcripts in CD4$^+$ T_H1 cells; conversely, T_H2 cells had low expression of linc-MAF-4 and abundant *MAF* transcripts (data not shown). The anti-correlation of expression between lincRNAs and their neighboring genes is not a common feature of all lincRNAs[12,16] and is probably restricted to a limited number of *cis*-acting lincRNAs. We also confirmed this observation in our data set (data not shown). Moreover, we observed no correlation between the expression of linc-MAF-4 and its proximal upstream protein-coding genes *CDYL2* and *DYNLRB2* (**Supplementary Fig. 4a**).

We observed a similar inverse relation between linc-MAF-4 and MAF when we differentiated naive CD4$^+$ T cells *in vitro* toward the T_H1 or T_H2 phenotype. In T lymphocytes differentiating toward the T_H1 phenotype, *MAF* transcripts increased up to day 3 and then decreased thereafter (**Fig. 5a**). Conversely, *linc-MAF-4* was poorly expressed for the first 3 d but then increased progressively (**Fig. 5a**). In CD4$^+$ T lymphocytes differentiating toward the T_H2 phenotype, the abundance of both *MAF* transcripts and c-Maf protein increased constantly up to day 8, while linc-MAF-4 remained constantly low (**Fig. 5a** and **Supplementary Fig. 4c**), similar to what we observed for CD4$^+$ T lymphocytes differentiating toward the T_H17 phenotype (**Supplementary Fig. 4d**).

We further characterized the transcriptional regulation of *MAF* by assessing the abundance of histone H3 trimethylated at Lys4 (H3K4me3) and occupancy by RNA polymerase II at the *MAF* promoter region in T_H1 and T_H2 cells. Consistent with the higher active transcription of *MAF* in CD4$^+$ T_H2 cells, we found enrichment for H3K4me3 in T_H2 cells relative to its abundance in T_H1 cells and that binding of RNA polymerase II at *MAF* promoter was higher in T_H2 than in T_H1 cells (**Fig. 5b**). Notably, knockdown of linc-MAF-4 in activated CD4$^+$ naive T cells led to increased *MAF* expression (**Fig. 5c** and **Supplementary Fig. 4e**). All the results presented above indicated that modulation of *MAF* transcription in T cells depended on tuning of its promoter setting, and suggested direct involvement of linc-MAF-4 in the regulation of *MAF* transcription.

We then assessed the overall effect of the knockdown of linc-MAF-4 on the differentiation of CD4$^+$ T cells by transcriptome profiling

and gene set–enrichment analysis. We defined as reference gene sets the groups of genes upregulated in $CD4^+$ naive T cells differentiated *in vitro* toward the T_H1 or T_H2 phenotype (**Supplementary Table 2**). We found that the $CD4^+$ T_H2 cell gene set showed enrichment for genes overexpressed in cells in which linc-MAF-4 was knocked down, whereas the $CD4^+$ T_H1 cell gene set showed depletion of those same genes (**Fig. 5d**). Concordant with those findings, the expression of *GATA3* and *IL4,* two genes characteristic of T_H2 cells, was increased after knockdown of linc-MAF-4 (**Fig. 5e** and **Supplementary Fig. 4f**). Together these results demonstrated that downregulation of linc-MAF-4 contributed to skewing of the differentiation of $CD4^+$ T cells toward the T_H2 phenotype.

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

Since the gene encoding linc-MAF-4 maps in relative proximity to *MAF* (within 139.5 kilobases), we sought to determine whether linc-MAF-4 was able to downregulate *MAF* transcription, and we investigated whether their genomic regions could physically interact. We exploited chromosome-conformation capture analysis to determine the relative crosslinking frequencies among regions of interest. We assessed the conformation of the genomic regions of the gene encoding linc-MAF-4 (called '*linc-MAF-4*' here) and *MAF* in differentiated $CD4^+$ T_H1 cells. We used common reverse-primer mapping of the *MAF* promoter region in combination with a set of primers spanning the locus and analyzed interactions by PCR. We detected specific interactions between the *MAF* promoter and the 5′ and 3′ end regions of *linc-MAF-4* (**Fig. 6a** and **Supplementary Fig. 5a**,**b**), which indicated the existence of an *in cis* chromatin-looping conformation that brought *linc-MAF-4* in close proximity to the *MAF* promoter. Notably, subcellular fractionation of $CD4^+$ T_H1 lymphocytes differentiated *in vitro* revealed considerable enrichment for linc-MAF-4 in the chromatin fraction (**Fig. 6b**). Because other chromatin-associated lincRNAs regulate neighboring genes by recruiting specific chromatin remodelers, we assessed by RNA-immunoprecipitation assay the interaction of linc-MAF-4 with various chromatin modifiers, including activators and repressors (data not shown), and found specific enrichment for linc-MAF-4 in the immunoprecipitates of two chromatin modifiers, EZH2 and LSD1 (**Fig. 6c** and **Supplementary Fig. 5c**). In agreement with those findings, we found that knockdown of linc-MAF-4 in activated $CD4^+$ naive T cells reduced the abundance of both EZH2 and LSD1 and correlated with lower enzymatic activity of EZH2 at the *MAF* promoter, as demonstrated by a lower abundance of H3K27me3 at this locus (**Fig. 6d**). Notably, the content of H3K27me3 was not diminished at either the *MYOD1* promoter region (a known target of EZH2) or at a region within the chromatin loop between *linc-MAF-4* and *MAF* marked by H3K27me3 (**Supplementary Fig. 5d**). Together these results demonstrated a long-distance interaction between the genomic regions of *linc-MAF-4* and *MAF*, through which linc-MAF-4 might act as a scaffold to recruit both EZH2 and LSD1 and modulate the enzymatic activity of EZH2 on the *MAF* promoter and thus regulate its transcription (**Fig. 6e**).

DISCUSSION

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

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

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

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

The role of lincRNAs in differentiation has been described for various cell types[17,20,23,47,48]. In the mouse immune system, it has been found that lincRNA expression changes during the differentiation of naive $CD8^+$ T cells into memory $CD8^+$ T cells[24] and during the differentiation of naive $CD4^+$ T cells into distinct lineages of helper T cells[28]. We have shown for human primary lymphocytes that activation-induced differentiation of $CD4^+$ naive T cells was associated with increased expression of lincRNAs belonging to the $CD4^+$ T_H1 cell signature, which suggests that upregulation of T_H1 cell lincRNAs is part of the cell-differentiation transcriptional program. Indeed, linc-MAF-4, one of the T_H1 cell signature lincRNAs, had low expression in T_H2 cells, and its experimental downregulation skewed differentiating helper T cells toward a T_H2 transcription profile. We found that linc-MAF-4 regulated transcription by exploiting a chromosome loop that brought its genomic region close to the promoter of *MAF*. We propose that the chromosome organization of this region allows a linc-MAF-4 transcript to recruit both EZH2 and LSD1 and to modulate the enzymatic activity of EZH2 that negatively regulates *MAF* transcription via a mechanism of action similar to that shown for the lincRNAs HOTAIR[49] and MEG3 (ref. 50). We therefore have provided mechanistic proof of the concept that lincRNAs can be important regulators of $CD4^+$ T cell differentiation. Given the number of specific lincRNAs expressed in various lymphocyte subsets, it can be postulated that many other lincRNAs might contribute to cell differentiation and to the definition of identity in human

lymphocytes. These findings and the high cell specificity of lincRNAs suggest that lincRNAs might be highly specific molecular targets for the development of new therapies for diseases (such as autoimmunity, allergy and cancer) in which altered CD4$^+$ T cell functions have a pathogenic role.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. ArrayExpress: E-MTAB-2319.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank C. Cheroni for support in statistical analysis; M. Moro and M.C. Crosti for technical assistance with cell sorting; S. Biffo, D. Gabellini, P. Della Bona and A. Lanzavecchia for discussions and critical revision of the manuscript; B.J. Haas and A. Dobin for help with the integration of genome-guided Trinity with STAR aligner; the Istituto Nazionale Genetica Molecolare Bioinformatics Facility for support; and the Google Summer of Code Project for supporting C. Wheeler in the development of a plug-in used here for the open-source bioinformatics library BioRuby that adds support for the multiple-alignment format (https://github.com/csw/bioruby-maf). Supported by Il Consiglio Nazionale delle Ricerche–Il Ministero dell'Istuzione dell'Universita e della Ricerca (EPIGEN), Fondazione Cariplo (2013-0955), the Associazione Italiana per la Ricerca sul Cancro (IG2013-ID14596), the European Research Council (269022 to S.A.; 617978 to M.P.) and Fondazione Romeo ed Enrica Invernizzi.

AUTHOR CONTRIBUTIONS

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

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Zhu, J., Yamane, H. & Paul, W.E. Differentiation of effector CD4 T cell populations. *Annu. Rev. Immunol.* **28**, 445–489 (2010).
2. Zhou, L., Chong, M.M. & Littman, D.R. Plasticity of CD4$^+$ T cell lineage differentiation. *Immunity* **30**, 646–655 (2009).
3. O'Shea, J.J. & Paul, W.E. Mechanisms underlying lineage commitment and plasticity of helper CD4$^+$ T cells. *Science* **327**, 1098–1102 (2010).
4. Kanno, Y., Vahedi, G., Hirahara, K., Singleton, K. & O'Shea, J.J. Transcriptional and epigenetic control of T helper cell specification: molecular mechanisms underlying commitment and plasticity. *Annu. Rev. Immunol.* **30**, 707–731 (2012).
5. O'Connell, R.M., Rao, D.S., Chaudhuri, A.A. & Baltimore, D. Physiological and pathological roles for microRNAs in the immune system. *Nat. Rev. Immunol.* **10**, 111–122 (2010).
6. Pagani, M. *et al.* Role of microRNAs and long-non-coding RNAs in CD4$^+$ T-cell differentiation. *Immunol. Rev.* **253**, 82–96 (2013).
7. Cobb, B.S. *et al.* T cell lineage choice and differentiation in the absence of the RNase III enzyme Dicer. *J. Exp. Med.* **201**, 1367–1373 (2005).
8. Koralov, S.B. *et al.* Dicer ablation affects antibody diversity and cell survival in the B lymphocyte lineage. *Cell* **132**, 860–874 (2008).
9. O'Connell, R.M. *et al.* MicroRNA-155 promotes autoimmune inflammation by enhancing inflammatory T cell development. *Immunity* **33**, 607–619 (2010).
10. Rodriguez, A. *et al.* Requirement of bic/microRNA-155 for normal immune function. *Science* **316**, 608–611 (2007).
11. Rossi, R.L. *et al.* Distinct microRNA signatures in human lymphocyte subsets and enforcement of the naive state in CD4$^+$ T cells by the microRNA miR-125b. *Nat. Immunol.* **12**, 796–803 (2011).
12. Cabili, M.N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
13. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
14. Hrdlickova, B. *et al.* Expression profiles of long non-coding RNAs located in autoimmune disease-associated regions reveal immune cell-type specificity. *Genome Med* **6**, 88 (2014).
15. Fatica, A. & Bozzoni, I. Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* **15**, 7–21 (2014).
16. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
17. Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295–300 (2011).
18. Khalil, A.M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. USA* **106**, 11667–11672 (2009).
19. Yoon, J.H. *et al.* LincRNA-p21 suppresses target mRNA translation. *Mol. Cell* **47**, 648–655 (2012).
20. Kretz, M. *et al.* Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* **493**, 231–235 (2013).
21. Poliseno, L. *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033–1038 (2010).
22. Sumazin, P. *et al.* An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* **147**, 370–381 (2011).
23. Cesana, M. *et al.* A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* **147**, 358–369 (2011).
24. Pang, K.C. *et al.* Genome-wide identification of long noncoding RNAs in CD8$^+$ T cells. *J. Immunol.* **182**, 7738–7748 (2009).
25. Collier, S.P., Collins, P.L., Williams, C.L., Boothby, M.R. & Aune, T.M. Cutting edge: influence of Tmevpg1, a long intergenic noncoding RNA, on the expression of Ifng by Th1 cells. *J. Immunol.* **189**, 2084–2088 (2012).
26. Gomez, J.A. *et al.* The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-γ locus. *Cell* **152**, 743–754 (2013).
27. Carpenter, S. *et al.* A long noncoding RNA mediates both activation and repression of immune response genes. *Science* **341**, 789–792 (2013).
28. Hu, G. *et al.* Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. *Nat. Immunol.* **14**, 1190–1198 (2013).
29. Haas, B.J. *et al. De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
30. Hart, T., Komori, H.K., LaMere, S., Podshivalova, K. & Salomon, D.R. Finding the active genes in deep RNA-seq gene expression studies. *BMC Genomics* **14**, 778 (2013).
31. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41**, D48–D55 (2013).
32. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
33. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
34. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
35. Rhind, N. *et al.* Comparative functional genomics of the fission yeasts. *Science* **332**, 930–936 (2011).
36. Finn, R.D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–D222 (2010).
37. Lin, M.F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–i282 (2011).
38. Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S. & Lander, E.S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154**, 240–251 (2013).
39. Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F. & Mattick, J.S. Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl. Acad. Sci. USA* **105**, 716–721 (2008).
40. Ørom, U.A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46–58 (2010).
41. Al-Shahrour, F., Minguez, P., Vaquerizas, J.M., Conde, L. & Dopazo, J. BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res.* **33**, W472–W476 (2005).
42. Volders, P.J. *et al.* LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.* **41**, D246–D251 (2013).
43. Ho, I.C., Lo, D. & Glimcher, L.H. c-maf promotes T helper cell type 2 (Th2) and attenuates Th1 differentiation by both interleukin 4-dependent and -independent mechanisms. *J. Exp. Med.* **188**, 1859–1866 (1998).
44. Liu, X., Nurieva, R.I. & Dong, C. Transcriptional regulation of follicular T-helper (Tfh) cells. *Immunol. Rev.* **252**, 139–145 (2013).
45. Sato, K. *et al.* Marked induction of c-Maf protein during Th17 cell differentiation and its implication in memory Th cell development. *J. Biol. Chem.* **286**, 14963–14971 (2011).
46. Mattick, J.S. The genetic signatures of noncoding RNAs. *PLoS Genet.* **5**, e1000459 (2009).
47. Klattenhoff, C.A. *et al.* Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell* **152**, 570–583 (2013).
48. Cabianca, D.S. *et al.* A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. *Cell* **149**, 819–831 (2012).
49. Tsai, M.C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689–693 (2010).
50. Kaneko, S. *et al.* Interactions between JARID2 and noncoding RNAs regulate PRC2 recruitment to chromatin. *Mol. Cell* **53**, 290–300 (2014).

ONLINE METHODS

Purification of primary immunological cell subsets. Blood buffy coat cells of healthy donors were obtained from Fondazione Istituto di Ricovero e Cura a Carattere Scientifico Ca'Granda Ospedale Maggiore Policlinico in Milan, and peripheral blood mononuclear cells were isolated by ficoll-hypaque density-gradient centrifugation. The ethical committee of Fondazione Istituto di Ricovero e Cura a Carattere Scientifico Ca'Granda Ospedale Maggiore Policlinico approved the use of peripheral blood mononuclear cells from healthy donors for research purposes, and informed consent was obtained from subjects. Human blood primary lymphocyte subsets were purified to a purity of >95% by cell sorting through the use of various combinations of surface markers (**Table 1**). For *in vitro* differentiation experiments, resting naive $CD4^+$ T cells were purified to a purity of >95% by negative selection with magnetic beads with an isolation kit for human $CD4^+$ Naive T cells (Miltenyi) and were stimulated with Dynabeads Human T-Activator CD3/CD28 (Life Technologies). Interleukin 2 (IL-2) was added at 20 IU/ml (202-IL; R&D Systems). T_H1 polarization was initiated with 10 ng/ml IL-12 (219-IL; R&D Systems) and T_H2-neutralizing antibody anti-IL-4 (2 µg/ml; MAB3007; R&D Systems). T_H2 polarization was induced by activation with phytohemagglutinin (4 µg/ml; L2769; Sigma) in the presence of IL-4 (10 ng/ml; 204-IL; R&D Systems), and neutralizing anti-IFN-γ (2 µg/ml; MAB 285; R&D Systems) and anti-IL-12 (2 µg/ml; MAB219; R&D Systems). For intracellular staining of GATA-3 and c-Maf, cells were harvested and then were fixed for 30 min at 4 °C in Fixation/Permeabilization Buffer (eBioscience). Cells were stained for 30 min at 4 °C with anti-GATA-3 (TWAJ; eBioscience) and anti-c-Maf (sym0F1; eBioscience) in washing buffer. Cells were then washed two times, resuspended in autoMACS buffer (Miltenyi) and analyzed by flow cytometry.

RNA isolation and RNA sequencing. Total RNA was isolated with an mirVana Isolation Kit. Libraries for Illumina sequencing were constructed from 100 ng of total RNA with the Illumina TruSeq RNA Sample Preparation Kit v2 (Set A). The libraries generated were loaded on to the cBot automated clonal amplification system (Illumina) for clustering on a HiSeq Flow Cell v3. The libriaries clustered on a HiSeq Flow Cell v3 were then sequenced with a HiScanSQ optical imaging system (Illumina). A paired-end run (with a read length of 101 bases) was performed with an SBS Kit v3 DNA sequencing kit (Illumina). Real-time analysis and base calling was performed with HiSeq Control Software Version 1.5 (Illumina).

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

Reference annotation. An initial custom reference annotation of unique, non-redundant transcripts was built by integration of the Ensembl database (version 67 from May 2012) with the lincRNAs identified by another group[13] through the use of the Cuffcompare tool (version 2.1.1) of the Cufflinks suite[34]. The annotated human lincRNAs were extracted from Ensembl through the use of the BioMart software suite (version 67) and were categorized by gene biotype 'lincRNA' (5,804 genes). Other classes of genes were integrated in the annotation: the list of protein-coding genes (21,976 genes), the collection of receptor-encoding genes defined in BioMart under GO term GO:000487 (2,043 genes encoding molecules with receptor activity function) and the class of genes encoding molecules involved in metabolic processes corresponding to GO term GO:0008152 (7,756 genes). Hence, the complete reference annotation consisted of 195,392 transcripts that referred to 62,641 genes, 11,170 of which were nonredundant lincRNA-encoding genes.

***De novo* genome-based transcripts reconstruction.** A comprehensive catalog of lincRNAs specifically expressed in human lymphocyte subsets was generated with a *de novo* genome-based transcripts reconstruction procedure by three different approaches. Two aligners were used: TopHat (version 1.4.1) and STAR (version 2.2.0). The *de novo* transcriptome assembly was performed on the aligned sequences (samples of the same population were concatenated into one 'population alignment') generated by STAR and TopHat using Cufflinks (version 2.1.1) with reference annotation to guide the assembly (-g option) coupled with multi-read (-u option) and fragment-bias correction (-b option) to improve the accuracy with which transcript abundance was estimated. By this method, about 3×10^4 to 5×10^4 previously unknown transcripts were identified in each lymphocyte population. The third approach used genome-guided Trinity software (additional information available at http://pasa.sourceforge.net/#A_ComprehensiveTranscriptome), which generates novel transcripts by local assembly on previously mapped reads from specific location. STAR was used instead of the Trinity default aligner[29]. Each candidate transcript was then processed via the PASA 'pipeline' (Program to Assemble Spliced Alignments; a genome annotation tool), which reconstructs the complete transcript and gene structures, resolving incongruences derived from transcript misalignments and alternatively splices events, refining the reference annotation when there was enough evidence and proposing new transcripts and genes in case no previous annotation was able to explain the new data (**Supplementary Note**).

Identification of previously unknown lincRNA-encoding genes. Annotated transcripts and previously unknown isoforms of known genes were discarded, and only previously unknown genes and their isoforms located in intergenic positions were retained. To filter out artifactual transcripts due to transcriptional noise or low polymerase fidelity, only multi-exonic transcripts longer than 200 bases were retained. Then, the HMMER3 algorithm[36] was run for each transcript to identify occurrences of any protein family domain documented in the Pfam database (release 26; both PfamA and PfamB were used). All six possible frames were considered for the analysis, and the matching transcripts were excluded from the final catalog.

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

***De novo* transcriptome data integration.** Duplicates among the transcripts identified with the same *de novo* method were resolved through the use of Cuffcompare (version 2.1.1). In the same way, the resulting three data sets were further merged to generate a nonredundant atlas of lincRNAs in human lymphocytes and only those genes identified by at least two of the three software programs used were considered. A unique name was given to each newly identified lincRNA gene composed by the prefix 'linc-' followed by the Ensembl gene name of the nearest protein-coding gene (irrespective of the strand). The additional designation 'up' or 'down' defines the location of the lincRNA relative to the sense of transcription of the nearest protein-coding gene. In addition, either 'sense' or 'antisense' was added to describe the concordance of transcription between the lincRNA and its nearest coding gene. A numerical counter only of newly identified lincRNAs related to the same protein-coding gene is added as suffix (such as 'linc-geneX-(up|down)-(sense|antisense)_#n'). This final nonredundant

doi:10.1038/ni.3093

catalog of newly identified lincRNAs includes 4,666 previously unknown transcripts referencing 3,005 previously unknown genes.

Definition of lincRNA signatures. Analysis of differences in expression among the 13 cell subsets profiled was performed with the Cuffdiff program of Cufflinks (version 2.1.1). This analysis was run using multi-read correction (-u option) and upper-quartile normalization (–library-norm-method quartile) to improve robustness of 'calls' for differences in expression for less-abundant genes and transcripts. Only genes expressed at an FPKM value over 0.21 (ref. 29) were considered in the downstream analysis to filter out genes that are merely byproducts of 'leaky' gene expression, sequencing errors, and/or off-target read mapping. After a pseudo-count of 1 was added to the raw FPKM value for each gene, with the application of $\log_2$ transformation and *z*-score normalization, K-means clustering with Euclidean metric was performed on lincRNA expression values with the MultiExperiment Viewer tool (version 4.6) (**Supplementary Note**). The same procedure was then applied to the expression values of genes encoding proteins, products involved in metabolic processes and receptors. The Silhouette function[52] was used to select an appropriate K value (number of clusters). K values ranging from 13 to 60 were tested, and the value associated with the highest Silhouette score for each class of genes was selected. The number of clusters that maximized the Silhouette score was 15 for lincRNA (**Supplementary Fig. 2a**), 24 for protein-coding genes, 23 genes encoding receptors and 36 for genes encoding products involved in metabolic processes. The centroid expression profile of each cluster was then evaluated to associate each cluster to a single cellular population (**Fig. 2**).

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

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

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

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

Gene set–enrichment analysis (GSEA). GSEA is a statistical methodology used to evaluate whether a given gene set shows significant enrichment for a list of gene markers (ranked by their correlation with a phenotype of interest). To evaluate this degree of 'enrichment', the software calculates an enrichment score (ES) by moving down the ranked list; i.e., it increases the value of the sum if the marker is included in the gene set and decreases this value if the marker is not in the gene set. The value of the increase depends on the gene-phenotype correlation. GSEA was performed by comparison of gene-expression data obtained from activated $CD4^+$ naive T cells transfected with siRNA specific for linc-MAF-4 or control siRNA. The experimentally generated data set from cells differentiated *in vitro* (in T_H1- or T_H2-polarizing conditions) from $CD4^+$ naive T cells of the same donors in which linc-MAF-4 was downregulated were used to construct reference gene sets for T_H1 and T_H2 cells. RNA for analysis of gene expression in differentiating T_H1 and T_H2 cells was collected 72 h after activation (i.e., the same time point of RNA collection in the linc-MAF-4-downregulation experiments), but a fraction of cells was further differentiated up to day 8 to assess the production of IFN-γ and IL-13 by T_H1 and T_H2 cells. The T_H1 and T_H2 data sets were ranked as $\log_2$ ratios of the expression values for each gene in the two conditions (T_H1/T_H2), and the genes with the greatest upregulation or downregulation (with $\log_2$ ratios ranging from $|3|$ to $|0.6|$) were assigned to the T_H1 or T_H2 reference sets, respectively.

Genes from the T_H1 gene list that were downregulated in a comparison of T_H1 cells versus cells transfected with control siRNA and genes from the T_H2 gene list that were downregulated in a comparison of T_H2 cells versus cells transfected with control siRNA were filtered out, which resulted in a T_H1 cell–specific gene set (74 genes) and a T_H2 cell–specific gene set (141 genes) (**Supplementary Table 2**). GSEA was then performed on the data set for the comparison of cells transfected linc-MAF-4-specific siRNA versus cells transfected with control siRNA. The metric used for the analysis is the $\log_2$ ratio of classes, with 1,000 gene set permutations for testing of significance.

Quantitative RT-PCR analysis. For reverse transcription, equal amounts of DNA-free RNA (500 ng) were reverse-transcribed with SuperScript III in the conditions suggested by the manufacturer (LifeTechnologies). Diluted cDNA was then used as input for quantitative RT-PCR to assess the expression of *MAF* (Hs00193519_m1), *IL4* (Hs00174122_m1), *GATA3* (Hs01651755_m1), *TBX21* (Hs00203436_m1), *RORC* (Hs01076119_m1), *IL17* (Hs00174383_m1), *Linc00339* (Hs04331223_m1), *MALAT1* (Hs01910177_s1), *RNU2.1* (Hs03023892_g1) and *GAPDH* (Hs02758991_g1) with Inventoried TaqMan Gene Expression assays (LifeTechnologies). For assessment of linc-MAF-4 and confirmation of $CD4^+$ T_H1 cell signature lincRNAs, specific primers were designed, and 2.5 μg RNA from $CD4^+$ T_H1 cells, regulatory T cells or naive cells was used for reverse transcription with SuperScript III (LifeTechnologies). Quantitative RT-PCR was performed on diluted cDNA with PowerSyberGreen (LifeTechnologies), and the specificity of each amplified product was monitored through the use of melting curves at the end of each amplification reaction. The primers used in quantitative PCR are listed in **Supplementary Table 3**.

Cell fractionation. T_H1 cells differentiated *in vitro* were resuspended for 10 min on ice in RLN1 buffer (50 mM Tris-HCl pH 8, 140 mM NaCl, 1.5 mM $MgCl_2$, 0.5% NP-40) supplemented with SUPERaseIn (Ambion). After a centrifugation at 300*g* for 2 min, the supernatant was collected as the cytoplasmic fraction. The pellet was resuspended for 10 min on ice in RLN2

buffer (50 mM Tris-HCl, pH 8, 500 mM NaCl, 1.5 mM $MgCl_2$ and 0.5% NP-40) supplemented with RNase inhibitors. Chromatin was pelletted at maximum speed for 3 min. The supernatant represented the nuclear fraction. All fractions were resuspended in TRIzol (Ambion) to a volume of 1 ml, and RNA was extracted following a standard protocol.

RNA immunoprecipitation. T_H1 cells differentiated *in vitro* underwent crosslinking by ultraviolet irradiation at 400 mJ/cm^2 in ice-cold Dulbecco's-PBS and then were pelleted at 1,350*g* for 5 min. The pellet was resuspended in ice-cold lysis buffer (25 mM Tris-HCl pH 7.5, 150 mM NaCl and 0.5% NP-40) supplemented with 0.5 mM β-mercaptoethanol, Protease Inhibitor Cocktail Tablets cOmplete, EDTA-free (Roche) and SUPERaseIn (Ambion) and was incubated with rocking at 4 °C until lysis was complete. The debris were centrifuged at 13,000*g* for 10 min. The lysate was precleared for 30 min at 4 °C with Dynabeads Protein G (Novex) and then was incubated for 2 h with 7 µg anti-EZH2 (39875; Active Motif) or anti-LSD1 (ab17721; Abcam), or with anti-HA (sc7392; Santa Cruz) as mock control. The lysate was coupled for 1 h at 4 °C to Dynabeads Protein G (Novex). Immunoprecipitates were washed for five times with lysis buffer. RNA was then extracted according to the protocol of the mirVana miRNA Isolation Kit (Ambion). The abundance of RNA transcripts encoding linc-MAF-4 or the negative controls β-actin, RNU2.1 and a region upstream the TSS of linc-MAF-4 (linc-MAF-4 control) was assessed by quantitative RT-PCR.

Chromatin immunoprecipitation. T_H1 and T_H2 cells differentiated *in vitro* were crosslinked for 12 min in their medium with 1:10 dilution of fresh formaldehyde solution (50 mM HEPES-KOH, pH 7.5, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA and 11% formaldehyde). Subsequently, they were treated for 5 min with 1:10 dilution of 1.25 M glycine and were centrifuged at 1,350*g* for 5 min at 4 °C. Cells were lysed at 4 °C in LB1 (50 mM HEPES-KOH, pH 7.5, 10 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40 and 0.25% Triton X-100) supplemented with Protease Inhibitor Cocktail Tablets cOmplete, EDTA-free (Roche) and phenylmethanesulfonyl fluoride (Sigma). Nuclei were pelleted at 1,350*g* for 5 min at 4 °C and were washed in LB2 (10 mM Tris-HCl, pH 8.0, 200 mM NaCl, 1 mM EDTA and 0.5 mM EGTA) supplemented protease inhibitors. Nuclei were again pelleted at 1,350*g* for 5 min at 4 °C and then were resuspended with a syringe in 200 µl LB3 (10 mM Tris-HCl, pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Na-deoxycholate and 0.5% N-lauroylsarcosine) supplemented with protease inhibitors. Cell debris were pelleted at 20,000*g* for 10 min at 4 °C, followed by chromatin immunoprecipitation overnight at 4 °C in LB3 supplemented with 1% Triton X-100 and protease inhibitors, with anti-H3K4me3 (07-473; Millipore), anti-H3K27me3 (07-449; Millipore), antibody to the RNA polymerase II STD repeat YSPTSPS (ab5408; Abcam), LSD1 (ab17721; Abcam), anti-EZH2 (39875; Active Motif) or no antibody (as negative control) . The next day, Dynabeads Protein G (Novex) were added, followed by incubation for 2 h at 4 °C with rocking. Then, the beads were washed twice with low-salt wash buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.1% SDS, 2 mM EDTA and 1% Triton X-100) and with a high-salt wash buffer (20 mM Tris-HCl, pH 8.0, 500 mM NaCl, 0.1% SDS, 2 mM EDTA and 1% Triton X-100). Samples obtained by immunoprecipitation with antibodies to histones (identified above) were also washed with a LiCl solution (10 mM Tris-HCl, pH 8.0, 250 mM LiCl, 1% NP-40 and 1 mM EDTA). All samples were finally washed with 50 mM NaCl in 1× Tris-EDTA buffer. Elution was performed overnight at 65 °C in 50 mM Tris-HCl pH 8.0, 10 mM EDTA and 1% SDS. Samples were treated for 2 h at 37 °C with 0.02 µg/µl RNase A (Sigma) and for 2 h at 55 °C with 0.04 µg/µl proteinase K (Sigma). DNA was purified with phenol-chloroform extraction.

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

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

51. Bonnal, R.J. *et al.* Biogem: an effective tool-based approach for scaling up open source software development in bioinformatics. *Bioinformatics* **28**, 1035–1037 (2012).
52. Rousseeuw, P.J. & Leroy, A.M. John Wiley & Sons. in *Wiley Series in Probability and Mathematical Statistics Applied Probability and Statistics* (Wiley, New York, 1987).
53. Bodega, B. *et al.* Remodeling of the chromatin structure of the facioscapulohumeral muscular dystrophy (FSHD) locus and upregulation of FSHD-related gene 1 (FRG1) expression during human myogenic differentiation. *BMC Biol.* **7**, 41 (2009).

doi:10.1038/ni.3093