doi:10.1038/nature14154

Super-enhancers delineate disease-associated regulatory nodes in T cells

Golnaz Vahedi[1], Yuka Kanno[1], Yasuko Furumoto[2], Kan Jiang[1], Stephen C. J. Parker[3]†, Michael R. Erdos[3], Sean R. Davis[4], Rahul Roychoudhuri[4], Nicholas P. Restifo[4], Massimo Gadina[2], Zhonghui Tang[5], Yijun Ruan[5], Francis S. Collins[3], Vittorio Sartorelli[6] & John J. O'Shea[1]

Enhancers regulate spatiotemporal gene expression and impart cell-specific transcriptional outputs that drive cell identity[1]. Super-enhancers (SEs), also known as stretch-enhancers, are a subset of enhancers especially important for genes associated with cell identity and genetic risk of disease[2–6]. CD4$^+$ T cells are critical for host defence and autoimmunity. Here we analysed maps of mouse T-cell SEs as a non-biased means of identifying key regulatory nodes involved in cell specification. We found that cytokines and cytokine receptors were the dominant class of genes exhibiting SE architecture in T cells. Nonetheless, the locus encoding *Bach2*, a key negative regulator of effector differentiation, emerged as the most prominent T-cell SE, revealing a network in which SE-associated genes critical for T-cell biology are repressed by BACH2. Disease-associated single-nucleotide polymorphisms for immune-mediated disorders, including rheumatoid arthritis, were highly enriched for T-cell SEs versus typical enhancers or SEs in other cell lineages[7]. Intriguingly, treatment of T cells with the Janus kinase (JAK) inhibitor tofacitinib disproportionately altered the expression of rheumatoid arthritis risk genes with SE structures. Together, these results indicate that genes with SE architecture in T cells encompass a variety of cytokines and cytokine receptors but are controlled by a 'guardian' transcription factor, itself endowed with an SE. Thus, enumeration of SEs allows the unbiased determination of key regulatory nodes in T cells, which are preferentially modulated by pharmacological intervention.

Histone acetyltransferase p300 loading demarcates regions of the genome bearing SE architecture[2,8]. Using chromatin immunoprecipitation followed by sequencing (ChIP-seq) for the p300 protein, we constructed SE catalogues of murine CD4$^+$ T helper (T_H)1, T_H2 and T_H17 cells. As predicted[2], the p300 load is exponentially distributed throughout the genome (Fig. 1a and Extended Data Fig. 1a). Approximately 40% of the p300 signal was found in a small fraction of p300-loaded enhancers in each lineage. The distribution of SEs was lineage-specific even in these closely related cells (Fig. 1b and Extended Data Fig. 1b). Regulatory regions of lineage-specific master transcription factors were endowed with SEs only in the relevant lineage (Extended Data Fig. 1c). We addressed the relationship between SEs and transcriptional activity in T cells by assigning SEs to associated genes using proximity measures[4], bearing in mind that alternative methods can conclusively establish such associations[6,9]. We found that SE architecture conferred significantly higher transcriptional activity compared with typical enhancer (TE) architecture and that this transcriptional activity was lineage-specific (Fig. 1c, d).

Widespread transcription at SEs themselves has been reported in embryonic stem (ES) cells and myogenic cells[2,10]. We next explored the extent to which SE domains were transcribed in T cells by employing high-resolution temporal expression maps of intergenic noncoding RNAs (ncRNAs)[11]. One-third of the ncRNAs expressed in T cells (501/1,524) were transcribed from an SE[10] (Fig. 1e and Extended Data Fig. 1d). Controlling for differences in the size of SEs and TEs, we found 80 ncRNAs per 10 megabase pairs (Mb) of SEs compared with 51 transcripts within TEs. The presence of an SE structure also distinguished highly lineage-specific and dynamic ncRNAs from constitutively expressed ones (Fig. 1f–h).

To elucidate the potential role of SEs in T-cell biology, we used ChIP-seq data sets to catalogue binding profiles of 13 transcription factors with major roles in T_H-cell differentiation across the merged map of SEs[12–15] (Fig. 2a–c). As in ES cells[2], STATs prominently bound SEs in CD4$^+$ T cells (Fig. 2a, d). Similarly, BATF, IRF4 and BACH2 were enriched at these regions (Fig. 2b, d). Lineage-specific transcription factors such as T-BET, GATA3 and ROR-γt showed preferential binding at lineage-specific SEs (Extended Data Fig. 2a). CTCF, an essential genome organizer, appeared to preferentially demarcate SE boundaries[6] (Extended Data Fig. 2b). Comparison of the enrichment of transcription factors at SEs and TEs revealed selective binding of STAT3 at SEs whereas other transcription factors demonstrated comparable binding at SEs and TEs (Extended Data Fig. 2c).

We next compared the identity of SE-associated genes in T cells with those in other cell lineages. In ES cells, SE structures primarily encompass transcription factors (Fig. 2e and Extended Data Fig. 3a). In macrophages, chemokine and cytokine activity were the most prominent categories. In T lymphocytes, genes relevant to cytokine biology were preferentially linked to SEs. Moreover, cytokine-related genes were not linked to SEs in non-immune related cells such as myotubes (Extended Data Fig. 3b). Thus, SEs are preferentially associated with genes that have a central role in the biology of specific cell lineages rather than a given class of genes (that is, transcription factors). In the case of T cells, SEs form an interactive network that reflects the biology of lymphocytes, their products and their mode of sensing the inflammatory environment.

We next ranked T-cell SEs on the basis of their p300 occupancy (Fig. 3a). Again, SEs with the highest p300 occupancy were typically associated with genes encoding cytokines and their receptors. However, the greatest p300 enrichment was associated with the *Bach2* locus, regardless of lineage subset (Fig. 3a, b). This is of interest as BACH2 is a broad regulator of immune activation that acts by stabilizing immunoregulatory capacity and attenuating effector differentiation[13]. Notably, genetic variations within this locus are associated with numerous immune-mediated diseases including rheumatoid arthritis[16], Crohn's disease[17], multiple sclerosis[18], asthma[19] and type 1 diabetes[20]. These observations prompted us to investigate the effect of *Bach2* deletion on the expression of SE-associated genes in T cells. Transcriptional profiling revealed that *Bach2* deficiency significantly affected the expression of genes with SE architecture compared to those with TEs or no enhancer mark in T cells (Fig. 3c, d). These findings were confirmed when we used synthetic RNA standards 'spiked-in' to rigorously normalize transcriptome data in wild-type and *Bach2*-deficient cells[21] (Methods; Extended Data

[1] Lymphocyte Cell Biology Section, National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), National Institutes of Health (NIH), Bethesda, Maryland 20892, USA. [2] Translational Immunology Section, NIAMS, NIH, Bethesda, Maryland 20892, USA. [3] Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, Maryland 20892, USA. [4] Center for Cancer Research, National Cancer Institute, NIH, Bethesda, Maryland 20892, USA. [5] The Jackson Laboratory for Genomic Medicine and Department of Genetic and Development Biology, University of Connecticut, Farmington, Connecticut 06030, USA. [6] Laboratory of Muscle Stem Cells and Gene Regulation, NIAMS, NIH, Bethesda, Maryland 20892, USA. †Present address: Departments of Computational Medicine & Bioinformatics and Human Genetics, University of Michigan, Ann Arbor, Michigan 48109, USA.

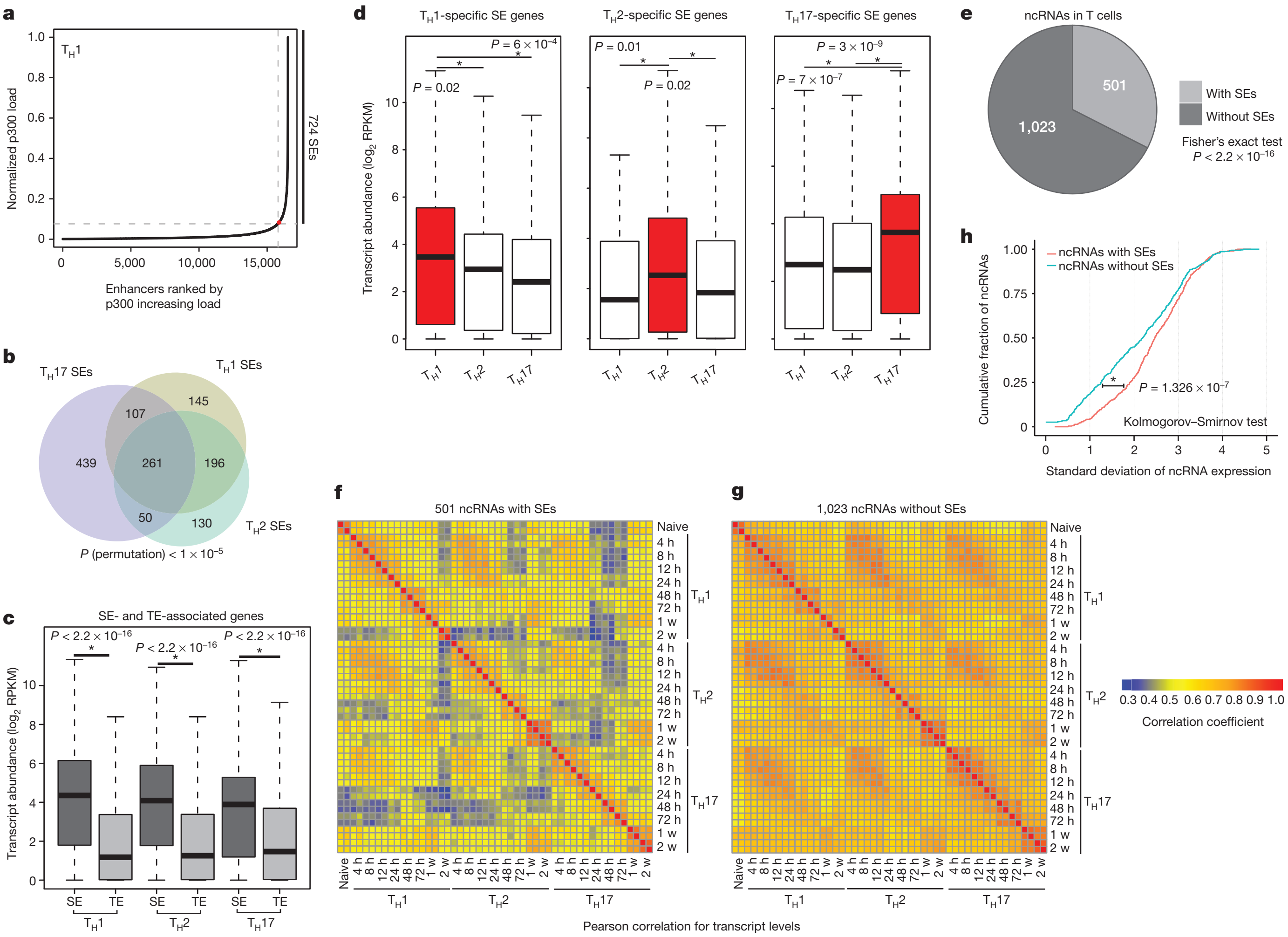


Figure 1 | SE structure predicts lineage- and stage-specific transcription. **a**, The histone acetyltransferase p300 is distributed asymmetrically across the genome in CD4$^+$ T cells with a subset of enhancers (SEs) that bind exceptionally high amounts of p300 (see Source Data). **b**, Closely related CD4$^+$ T-cell populations have distinctive SE landscapes. Venn diagram depicts shared and unique SE domains in T-cell subsets. **c**, SE-associated genes are highly transcribed compared with TE-associated genes. Proximity measures were used to assign SEs and TEs to their target genes (*P* values, Wilcoxon rank-sum test). RPKM, reads per kilobase of exon per million. **d**, Presence of lineage-specific SEs predicts cell-selective expression. Three groups of genes associated with unique SE structure in each lineage were defined as T_H1-, T_H2- and T_H17-specific SE genes. The expression of lineage-specific SE-associated genes was more significant in the corresponding lineage (*P* values, Wilcoxon rank-sum test). **e**, SE domains are themselves transcribed in CD4$^+$ T cells. The list of ncRNAs was derived from the map of intergenic transcripts in T-cell subsets[11]. One-third of ncRNAs in T cells (501/1,524) were transcribed from an SE. **f–h**, The SE structure differentiates highly lineage-specific and dynamic noncoding transcripts from constitutively expressed transcripts across T-cell lineages. **f**, **g**, Pearson correlation coefficients for transcription levels between each pair of differentiation stages were calculated for 501 ncRNAs with SEs (**f**) and 1,023 ncRNAs without SEs (**g**). **h**, ncRNA transcripts with SEs have a greater standard deviation across differentiation stages compared to those without SEs.

Fig. 3c, d). This transcriptional difference remained statistically significant when we controlled for higher levels of gene expression for SE-associated genes (Extended Data Fig. 3e). Furthermore, loss of BACH2 led to the largest difference between SEs and TEs in comparison with other transcription factors such as STATs, BATF and IRF4 (Extended Data Fig. 4a, b). In particular, 348 genes, 26% of those with SE structure in CD4$^+$ T cells, were repressed by BACH2 (Fig. 3e and Extended Data Fig. 4c–e). In addition to protein-coding genes, a subset of SE-linked ncRNAs (56) were also repressed by BACH2 (Fig. 3f). Transcriptional upregulation at some of these domains correlated with the upregulation of nearby genes in *Bach2*-deficient cells (Fig. 3g and Extended Data Fig. 4f, g). This previously unrecognized circuitry reveals that a subset of genes and noncoding transcripts endowed with SE architecture in CD4$^+$ T cells are tightly and negatively controlled by the 'guardian' transcription factor BACH2, which itself has a rich cassette of regulatory elements (Extended Data Fig. 4h).

It has been shown that single-nucleotide polymorphisms (SNPs) associated with diseases relevant to a particular cell type are more enriched in SEs compared with TEs[2,5]. CD4$^+$ T cells are important contributors to a wide variety of autoimmune diseases including rheumatoid arthritis. Thus, we explored the extent to which rheumatoid-arthritis-associated genetic variants were situated within SEs. We delineated SEs in human CD4$^+$ T-cell subsets and found that 26% of the SNPs highly associated with rheumatoid arthritis[7] (27/101) fell within SEs (Fig. 4a). In contrast, only 7% of rheumatoid arthritis SNPs overlapped with TEs (Fig. 4a). Controlling for differences in the size of genomic regions, we found that the number of SNPs per 10 Mb of SEs was significantly higher than that in TEs (Fig. 4a). Genetic variants associated with other autoimmune disorders such as inflammatory bowel disease, multiple sclerosis and type 1 diabetes also exhibited preferential enrichment in CD4$^+$ T-cell SEs compared to TEs (Fig. 4a). Such enrichment was also present when we considered variants in high linkage disequilibrium with disease-associated SNPs (Extended Data Fig. 5a). As a comparison, genetic variants associated with type 2 diabetes and cancer, diseases in which CD4$^+$ T cells are not thought to have major roles, were also assessed and found not to be significantly enriched within T-cell SEs (Fig. 4a).

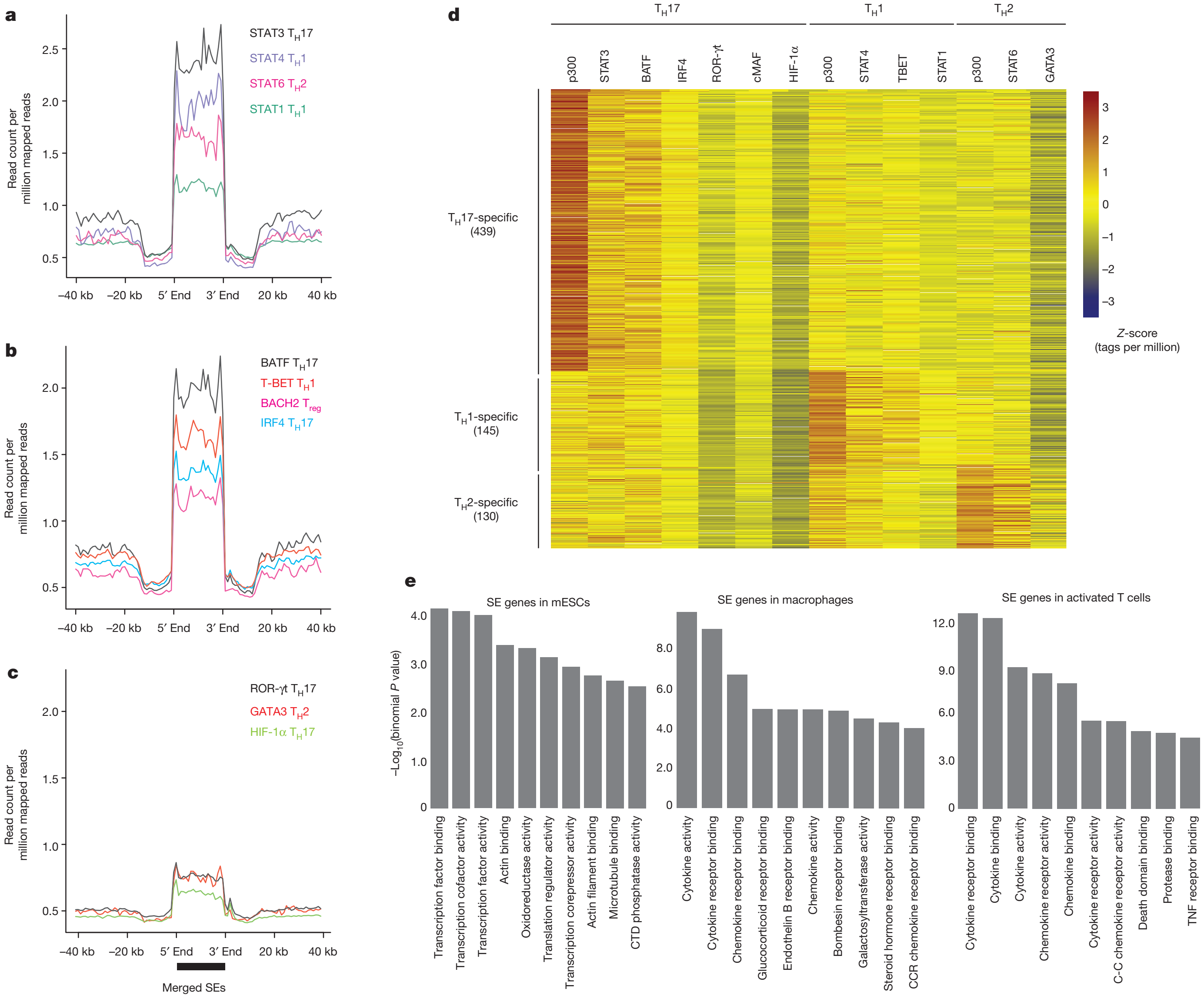


Figure 2 | Transcription factors with major roles in T_H-cell differentiation occupy SEs. a–c, Lineage-predicting transcription factors are enriched at SE domains. The catalogue of SEs in $CD4^+$ T cells was constructed by merging T_H1, T_H2 and T_H17 SEs. Binding patterns of STAT1, STAT3, STAT4 and STAT6 (**a**), BATF, T-BET, BACH2 and IRF4 (**b**), and HIF-1α, ROR-γt and GATA3 (**c**) are demonstrated at SEs. **d**, Binding of lineage-specific transcription factors correlates with the presence of lineage-specific SEs in T cells ($\log_2$ tags per million) (see Source Data). **e**, Gene ontology (GO) functional categories relevant to cytokines and cytokine receptors are enriched at SE-associated genes in T cells. GO analysis for SE regions was performed using GREAT[24]. mESCs, mouse ES cells.

We refined these observations by examining genes that were affected by rheumatoid-arthritis-associated genetic variants, focusing on 98 candidate genes associated with rheumatoid arthritis[7]. While SEs in muscle cells showed little association (Fig. 4b), rheumatoid arthritis risk genes were preferentially associated with SEs in cytotoxic natural killer cells ($CD56^+$) and monocytes ($CD14^+$). However, the strongest enrichment occurred in $CD4^+$ T cells, where half of the rheumatoid arthritis risk genes (53/98) were linked to $CD4^+$ T-cell SEs (Fig. 4b).

SE structures are thought to be particularly sensitive to perturbation owing to the cooperative and synergistic binding of numerous factors at these domains[3]. Given the enrichment of STATs at SEs and the prevalence of SEs at cytokines and their receptors, we measured the effect of tofacitinib, a JAK inhibitor recently approved by the US Food and Drug Administration for the treatment of rheumatoid arthritis, on T-cell transcriptomes. We found that tofacitinib treatment had a significantly greater impact on the transcription of genes with SEs than TEs (Extended Data Fig. 5b). Moreover, when genes were ranked on the basis of their transcript levels in T cells, the most highly expressed genes with SEs showed a larger change in their expression compared to those without SEs, emphasizing that tofacitinib discriminates genes with SE structure (Extended Data Fig. 5c). Although harbouring the strongest SE in T cells, BACH2 levels were not affected by acute tofacitinib treatment, suggesting that BACH2 is regulated in a JAK/STAT-independent manner. Finally, we related the effect of this rheumatoid arthritis drug to the genetics of the disease and found that tofacitinib treatment disproportionately affected the expression of rheumatoid arthritis risk genes with SE structures in $CD4^+$ T cells compared with those lacking this chromatin feature (Fig. 4c and Extended Data Fig. 5d). Furthermore, tofacitinib treatment selectively affected inflammatory bowel disease[22] and multiple sclerosis[23] risk genes with SEs (Extended Data Fig. 6).

We have defined the T_H-cell SE landscape in the hope of better defining key regulatory nodes in a non-biased fashion. We found that in T cells these nodes largely comprise cytokine and cytokine receptor genes. Thus, T-cell ‘identity’ relates largely to the precise regulation of these key effectors and sensors. However, a predominant SE-associated gene in all T-cell lineages was *Bach2*, which may represent the first example

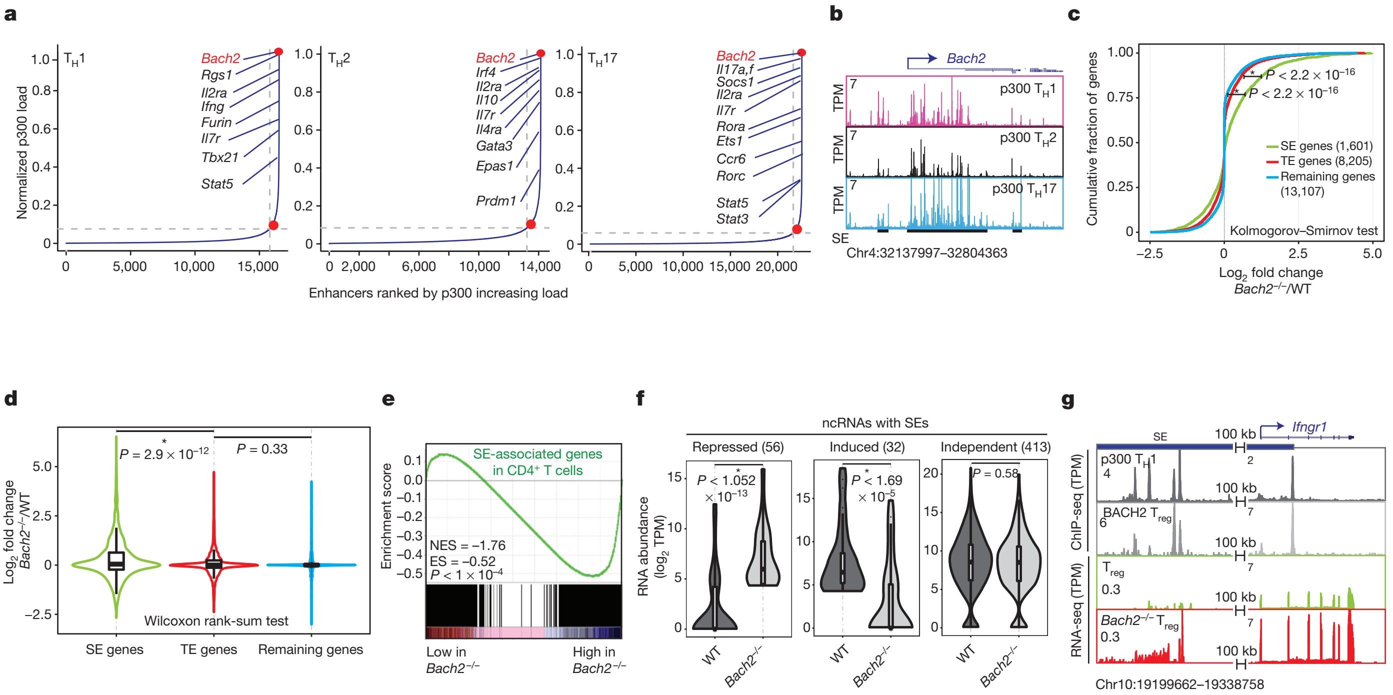


Figure 3 | *Bach2* is endowed with the highest p300-enriched SE in T cells. **a**, Ranked order of p300-loaded enhancers in T-cell subsets identifies *Bach2* as the strongest SE-associated gene in $CD4^+$ T cells. **b**, The *Bach2* locus, the top ranked SE, exhibits an exceptional amount of p300 binding. **c**, **d**, BACH2 preferentially represses SE genes. Wild-type (WT) and *Bach2*-deficient $CD4^+$ T cells were polarized to induced regulatory T cells (iT_{reg} cells) and were processed for total RNA extraction ($n = 3$). Normalized transcript abundance measured by RNA-seq (RPKM) was evaluated in wild-type and *Bach2*-deficient cells at SE- and TE-associated genes and compared to the remaining genes. Cumulative distribution (**c**) and violin plots (**d**) show the ($\log_2$) fold change in gene expression for wild-type versus *Bach2*-deficient cells (see Source Data). **e**, Gene set enrichment analysis (GSEA) of SE-associated genes reveals that SE genes are enriched in genes repressed by BACH2. ES, enrichment score; NES, normalized enrichment score. **f**, BACH2 affects a subset of noncoding transcripts at SE domains. Overall, 56 ncRNAs with SE structures are repressed while 32 transcripts are induced by BACH2 (see Source Data). *P* values, Wilcoxon rank-sum test. **g**, BACH2-associated repression of a noncoding transcript with an SE architecture correlates with the transcriptional repression of a nearby gene (*Ifngr1*). Direct BACH2 binding along with the transcript levels in wild-type and *Bach2*-deficient cells measured by RNA-seq were depicted in a 140 kb window accommodating the *Ifngr1* gene. TPM, tags per million.

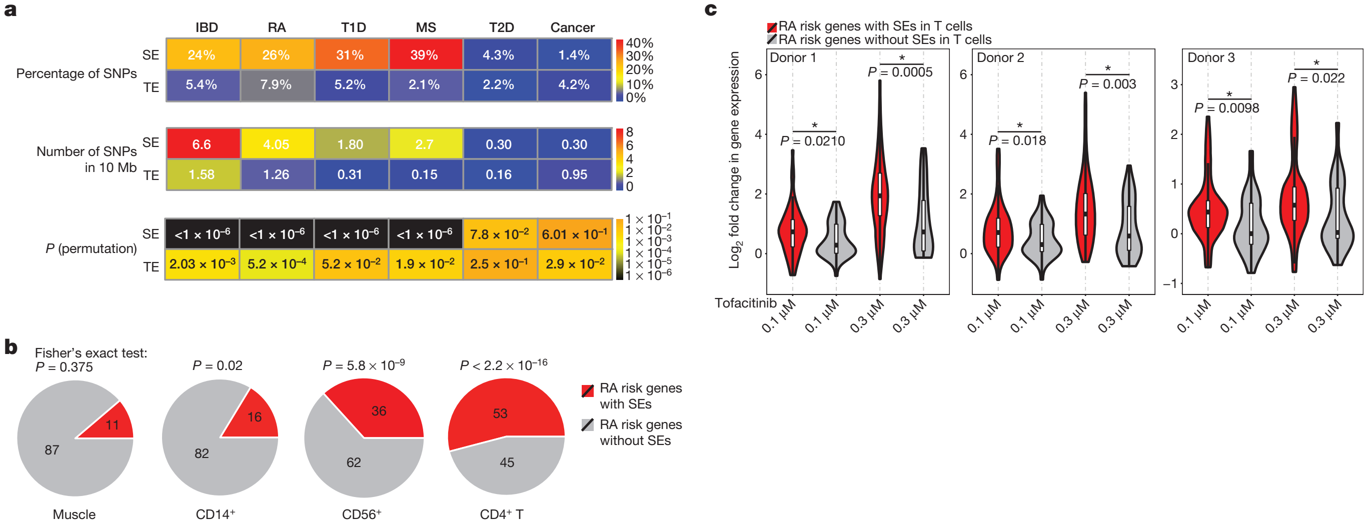


Figure 4 | Rheumatoid arthritis risk genes with SE structure are selectively targeted by the JAK inhibitor tofacitinib. **a**, SNPs associated with autoimmune diseases including rheumatoid arthritis (RA), inflammatory bowel disease (IBD), multiple sclerosis (MS), and type 1 diabetes (T1D) are preferentially enriched at the SE structure of human $CD4^+$ T cells. In contrast, SNPs associated with disorders in which $CD4^+$ T cells have limited roles, such as type 2 diabetes (T2D) and cancer, are not enriched in these genomic domains. A catalogue of 1,426 SEs in human T cells was constructed by aggregating SE predictions in human T_H1, T_H2 and T_H17 cells using H3K27ac data (see Source Data). We divided the number of SNPs enriched in SEs/TEs by the total size of SEs (66.5338 Mb) and TEs (63.12915 Mb) and reported the number of SNPs within every 10 Mb of the genome (*P* values, permutations test). **b**, Rheumatoid arthritis risk genes are linked to SEs in $CD4^+$ T cells. The 98 candidate genes associated with rheumatoid arthritis were from ref. 7. **c**, Rheumatoid arthritis risk genes with SEs are selectively targeted by a JAK inhibitor, tofacitinib. Violin plots depict the fold change in expression ($\log_2$) after tofacitinib treatment of human $CD4^+$ T cells at rheumatoid arthritis risk genes with or without SEs (three donors). To ensure accurate inference of the effect of tofacitinib on the transcriptome, spiked-in RNA standards were added and gene expression levels (RPKM) were renormalized based on the spiked-in standards (*P* values, Wilcoxon rank-sum test).

of a class of transcriptional regulators that broadly constrains transcription at SEs. Furthermore, SNPs associated with immune-related diseases were enriched at T-cell SEs, and a drug, which blocks cytokine signalling and is clinically efficacious in autoimmune disease, preferentially impacted SE-associated genes. Hence, our study provides a systematic approach by which the SE map of relevant cell types can be integrated with human genetics to discover drug target genes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 8 May; accepted 9 December 2014.

Published online 16 February 2015.

1. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).
2. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).
3. Lovén, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153,** 320–334 (2013).
4. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153,** 307–319 (2013).
5. Parker, S. C. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl Acad. Sci. USA* **110,** 17921–17926 (2013).
6. Dowen, J. M. *et al.* Control of cell identity genes occurs in insulated neighborhoods in Mammalian chromosomes. *Cell* **159,** 374–387 (2014).
7. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506,** 376–381 (2014).
8. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–283 (2011).
9. Kieffer-Kwon, K. R. *et al.* Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell* **155,** 1507–1520 (2013).
10. Mousavi, K. *et al.* eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Mol. Cell* **51,** 606–617 (2013).
11. Hu, G. *et al.* Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. *Nature Immunol.* **14,** 1190–1198 (2013).
12. Ciofani, M. *et al.* A validated regulatory network for Th17 cell specification. *Cell* **151,** 289–303 (2012).
13. Roychoudhuri, R. *et al.* BACH2 represses effector programs to stabilize T_{reg}-mediated immune homeostasis. *Nature* **498,** 506–510 (2013).
14. Wei, L. *et al.* Discrete roles of STAT4 and STAT6 transcription factors in tuning epigenetic modifications and transcription during T helper cell differentiation. *Immunity* **32,** 840–851 (2010).
15. Vahedi, G. *et al.* STATs shape the active enhancer landscape of T cell populations. *Cell* **151,** 981–993 (2012).
16. McAllister, K. *et al.* Identification of BACH2 and RAD51B as rheumatoid arthritis susceptibility loci in a meta-analysis of genome-wide data. *Arthritis Rheum.* **65,** 3058–3062 (2013).
17. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genet.* **42,** 1118–1125 (2010).
18. Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476,** 214–219 (2011).
19. Ferreira, M. A. *et al.* Identification of IL6R and chromosome 11q13.5 as risk loci for asthma. *Lancet* **378,** 1006–1014 (2011).
20. Cooper, J. D. *et al.* Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature Genet.* **40,** 1399–1401 (2008).
21. Lovén, J. *et al.* Revisiting global gene expression analysis. *Cell* **151,** 476–482 (2012).
22. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–124 (2012).
23. Beecham, A. H. *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genet.* **45,** 1353–1360 (2013).
24. McLean, C. Y. *et al.* GREAT improves functional interpretation of *cis*-regulatory regions. *Nature Biotechnol.* **28,** 495–501 (2010).

Acknowledgements The authors thank B. Afzali, A. Nussenzweig, A. Poholek, S. Canna, A. Richard and E. Mathe for critically reading this manuscript. We are grateful to R. Faryabi, H.-Y. Shih, W. Resch, M. Ombrello, Z. Deng and E. Remmers for their contributions to experimental and analytical components of this study. We also thank H. Sun, G. Gutierrez-Cruz, J. Simone, J. Lay and K. Tinsley for their excellent technical support. This study used the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health. R.R. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (grant number 105663/Z/14/Z). This work was supported by the Intramural Research Program of NIAMS and by NCI grant R01 CA186714.

Author Contributions G.V., V.S., F.S.C. and J.J.O'S. participated in the study design. Y.K., Y.F. and K.J. performed sequencing experiments. Z.T. and Y.R. supervised and performed sequencing experiments. Y.F. and M.G. performed tofacitinib-related experiments. G.V. performed computational analysis. S.C.J.P., M.R.E. and S.R.D. participated in statistical analysis relevant to human genetics. R.R. and N.P.R. supervised and performed experiments involving *Bach2*-deficient cells. Y.K., Y.F. and S.R.D. participated in writing of the methodology. G.V., V.S. and J.J.O'S. wrote the manuscript and all authors reviewed it. J.J.O'S. supervised the project.

Author Information All ChIP- and RNA-sequencing data sets have been deposited in the Gene Expression Omnibus under accession number GSE60482. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.V. (vahedig@mail.nih.gov) or J.J.O'S. (osheaj@arb.niams.nih.gov).

METHODS

Antibodies and reagents. The following antibodies and reagents were obtained from eBioscience: CD4-PerCPCy5.5, CD45RA-PE, CD45RO-eFluor450, CD28-purified. Anti-CD3 antibody was obtained from BioXcell. CP-690,550 (tofacitinib) was prepared by the National Institutes of Health (NIH) Chemical Genomics Center and dissolved in dimethylsulphoxide (DMSO). Source Data associated with Fig. 1 summarizes ChIP-seq data sets generated or used for this data set along with relevant antibodies.

Cell culture and stimulation for tofacitinib-treated human T cells. Whole blood from healthy donors was provided from the NIH blood bank and informed consent was obtained from subjects. To obtain lymphocyte population, heparinized whole blood from healthy donors was separated by Ficoll Paque PLUS (Sigma). Naive $CD4^+$ $CD45RA^+$ $CD45RO^-$ T-cell population was sorted on a FACS Aria III (BD Bioscience). Cells were activated by plate-bound anti-CD3/anti-CD28 (10 μg ml^{-1}) in supplemented RPMI 1640 medium containing 10% FCS, 2 mM glutamine, 100 IU ml^{-1} penicillin, 0.1 mg ml^{-1} streptomycin, 20 mM HEPES buffer (all from Invitrogen) for 3 days and cultured in the presence of IL-2 for 1 day. During T-cell activation, cells were treated with the indicated concentrations of CP-690,550 (tofacitinib).

RNA-seq preparation. Total RNA was prepared from approximately 1 million cells by using mirVana miRNA Isolation Kit (AM1560, ABI). Two-hundred nanograms to 1 μg of total RNA was subsequently used to prepare RNA-seq libraries by using TruSeq SR RNA sample prep kit (FC-122-1001, Illumina) or by a combination of NEBNext RNA library prep kit (New England BioLabs) and Ovation SP Ultralow DR Multiplex system (Nugen) by following the manufacturer's protocol. The libraries were sequenced for 50 cycles (single read) with HiSeq 2000 (Illumina). Where indicated, ERCC RNA spike-in mix 1 (Invitrogen) was added to the samples based on the cell counts (1 μl of 1/10 dilution to 1 million cells).

RNA-seq analysis. RNA-seq libraries made by Illumina TruSeq were first trimmed using 'cutadapt' with TruSeq Indexed Adapters. An error rate of 0.1 was chosen for 'cutadapt'. Overall, the percentage of trimmed reads was lower than 3% of the total reads across different libraries. Trimmed fastq files were then aligned to mm9 or hg19 reference genomes using tophat with bowtie2 indexes derived based on UCSC annotations. The normalization of RNA-seq libraries shown on the genome browser was carried out using 'bedtools genomecov' to 'scale' the bam files to tags-per-million values. 'HT-seq' was used to find the read counts across the UCSC reference genome and DEseq was further employed to characterize differentially regulated genes where repeats were available (*Bach2*-deficient RNA-seq).

RNA-seq analysis of *Bach2*-deficient cells. Wild-type and *Bach2*-deficient naive ($CD44^-$ $CD62L^+$ $CD25^-$) $CD4^+$ cells were isolated to >95% purity from C57BL/6 mice reconstituted with mixtures of wild-type and knockout OT-II TCR-transgenic bone marrow. Cells were stimulated at 1×10^5 cells per 96-well plate coated in 5 μg ml^{-1} anti-CD3 in the presence of soluble anti-CD28 (5 μg ml^{-1}), 100 IU recombinant human IL-2 and 5 ng ml^{-1} recombinant human TGF-β for 3 days. Cells were counted using a haematocytometer, or analysed by FACS for cell size or intracellular Foxp3 content. Cells were harvested and subjected to total RNA extraction (Qiagen RNeasy Plus kit with column-based DNA removal).

RNA-seq with spiked-in standards. ERCC RNA spike-in mix 1 (Invitrogen) was added to samples based on the cell counts (1 μl of 1/10 dilution to 1 million cells). The ERCC RNA Spike-In Control Mixes used here comprise a set of 92 polyadenylated transcripts that mimic natural eukaryotic mRNAs. On the basis of page 12 of the ERCC manual, we calculated the concentrations of the RNA molecules added to total RNA (that is, number of copies of spiked-in molecules per million cell) (Source Data associated with Fig. 3). It is clear that the standards cover a wide range of copy numbers.

Spiked-in RNA-seq analysis. The spiked-in RNA-seq libraries were subsequently sequenced on Illumina HiSeq 2000 and then trimmed using 'cutadapt' with TruSeq Indexed Adapters. The sequences of the ERCC synthetic spiked-in RNAs (http://tools.invitrogen.com/downloads/ERCC92.fa) were then added to both mouse and human genomes (genome.fa). The exon reference (http://tools.invitrogen.com/downloads/ERCC92.gtf) has also been added to the UCSC exon reference.

New bowtie indexes were then built and reads were aligned to the newly built genomes using tophat. The RPKM (reads per kilobase of exon per million) was then computed for each gene and synthetic spiked-in RNA using cufflinks. To renormalize the RNA-seq data using spiked-in control, we followed the same procedure as previously recommended[21]. We used a loess regression to renormalize the RPKM values by using only the spiked-in values to fit the loess. The *affy* package in R provides a function, *loess.normalize*, which will perform loess regression on a matrix of values (defined by using the parameter *mat*) and allows for the user to specify which subset of data to use when fitting the loess (defined by using the parameter *subset*). For this application the parameters mat and subset were set as a matrix of all RPKM values and the row indices of the ERCC spiked-ins, respectively. The default settings for all other parameters were used. The result of this was a matrix of RPKM values normalized to the control ERCC spiked-ins. Source Data associated with Fig. 3 quantitates the fraction of spiked-in tag counts in each RNA-seq library when tag counts were generated using "htseq-count–mode=intersection-nonempty–stranded=no".

ChIP-seq. For p300, we chemically crosslinked and sonicated cells to generate fractionated genomic DNA. ChIP was performed by using anti-p300 (sc-585, Santa Cruz Biotechnology). The DNA fragments were blunt-end ligated to the Illumina adaptors, amplified, and sequenced by using the Illumina Genome Analyzer II (Illumina). Sequence reads of 25 or 36 bp were obtained by using the Illumina Analysis Pipeline. Publically available ChIP-seq data sets are listed in Source Data associated with Fig. 1 and were obtained from several published studies[12–15,25–28].

ChIP-seq analysis. ChIP libraries were sequenced for 36 or 50 cycles on an Illumina Genome Analyzer II or HiSeq 2000, respectively, according to the manufacturer's instructions. ChIP libraries were aligned to mm9 or hg19 reference genomes using bowtie2 with bowtie indexes derived based on UCSC annotations and Phred+33 selected for qualities. Source Data associated with Fig. 1 summarizes ChIP-seq data sets generated or used for this study along with relevant antibodies. Peak calling for all transcription factors and p300 binding was performed by macs14 (ref. 29) using P value $= 1 \times 10^{-7}$. The control library for all peak-calling libraries was the input DNA performed under T_H0 condition. Peaks with false discovery rate (FDR) values more than 30% were further excluded. Peak intensities ('tags' column) were normalized as tags-per-million reads in the original library. Peak calling for H3K27ac libraries was performed using SICER[30] where the window size = 200 bp, gap size = 200 bp and E value = 200. To visualize and normalize ChIP-seq libraries on the UCSC genome browser, we used 'bedtools genomecov' to 'scale' the bam files to tags-per-million values. Furthermore, 'wigToBigWig' was used to generate bigwig files. y-Axis in all gene tracks is in tags per million (TPM).

Delineation of SEs and typical enhancers TEs. To accurately delineate SE domains, we followed the same approach that was proposed earlier[2–4]. We first merged genomic regions within 12.5 kb of one another (using mergeBed in bedtools). We then ranked all regions in a cell type by increasing total ChIP-seq occupancy of p300 or H3K27 acetylation, scaled the data such that the x and y axes were from 0–1 by normalizing to the largest value, and plotted the intensity of ChIP-seq (Fig. 1a). These plots revealed a clear point in the distribution of enhancers where the occupancy signal began increasing rapidly. To geometrically define this point, we found the x-axis point for which a line with a slope of 1 was tangent to the curve. As suggested by Young and colleagues, we defined genomic regions above this point to be SEs. All genomic regions below that point that did not harbour promoters (±5 kb of RefSeq transcription start sites) were then referred as TEs. The single map of SEs in $CD4^+$ T cells was constructed by merging maps of T_H1, T_H2 and T_H17 SEs (unionBedGraphs). Similarly, TEs in each lineage were delineated as described and then merged in different lineages to build one map for TEs. Since SEs in one lineage can be TEs in other lineages, SE coordinates were then excluded from the final TE map for $CD4^+$ T cells. Source Data associated with Figs 1 and 4 summarize the coordinates of SEs in both human and mouse in our study.

Delineation of cell-type specific SEs. To define cell-type-specific and shared SE domains, we started from the merged map of SEs in T_H1, T_H2 and T_H17 cells (Source Data associated with Fig. 1). We then used 'bedtools intersect' with $-f$ 0.1, 0.3, 0.5 or 0.7 with $-a$ being the coordinate of merged map and $-b$ being the SE coordinates in the corresponding condition and reporting $-c$ in the output (for each entry in A, report the number of overlaps with B and reporting 0 for A entries that have no overlap with B) (Fig. 1b and Extended Data Fig. 1b). We used the pheatmap function to demonstrate the shared and unique SEs based on the outputs of 'bedtools intersect' for the three cell types. Figure 1b corresponds to $f = 0.1$.

Characterizing SE- and TE-associated genes. SE- and TE-associated genes were defined based on the closest genes to these genomic regions (bedtools closest) using RefSeq coordinates of genes. As described in this package, closestBed first searches for features in B (gene coordinates) that overlap a feature in A (SE coordinates). If overlaps are found, gene coordinates that overlap the highest fraction of SE regions are reported. Then in the case of multiple genes overlapping SEs, the gene with the highest fraction of overlap is reported. If no overlaps are found, closestBed looks for the feature in B that is closest (that is, least genomic distance to the start or end of A) to A.

Transcription at T_H-specific SE genes. We delineated SE genes as described earlier for T_H1, T_H2 and T_H17 p300 binding (Fig. 1d). We defined a gene to be specific to a lineage if that gene was not present in SE-associated genes in the other two lineages. We then showed the $\log_2$ RPKM values for this list of genes across three different lineages. P values were calculated using the Wilcoxon rank-sum test.

Characterization of long ncRNAs associated with SEs. The list of transcribed ncRNAs in T cells was compiled from Hu *et al.*[11]. Hu *et al.*[11] performed the following steps for the identification of ncRNA clusters: (1) call RNA-seq read enriched islands from intergenic regions using SICER (window = 100 bp, gap = 200 bp, E value = 100); (2) keep islands shared by duplicates; (3) pool islands from all samples, independently done for data sets from total RNA-seq and from PolyA$^+$ RNA-seq; (4) cluster neighbouring islands based on similarity in expression profiles across different samples ($r > 0.8$). Transcribed regions that overlapped SEs were identified using the countOverlaps function in the GenomicRanges package in R (Fig. 1e) . To quantitate the correlation levels in transcripts across different T-cell lineages and time points, we used the 'cor' function in R with 'pearson' as 'method' (no $\log_2$ transformation was performed prior to the calculation of correlation) (Fig. 1f, g). Transcript levels for polyA RNAs used for this analysis were extracted from the supplementary table provided in ref. 11. Genomic coordinates of these two groups of ncRNAs are provided in Source Data associated with Fig. 1.

Cumulative distribution of ncRNAs with and without SEs. We used the 'rowSds' function from the library 'matrixStats' in R to calculate the standard deviation in each row for expression levels of ncRNAs with and without SEs. We used ggplot and stat_ecdf() to plot the cumulative distribution of standard deviation in these two groups of ncRNAs. Cumulative distribution in Fig. 1h shows quantitative shift in standard deviation of transcript levels for ncRNAs with SEs relative to those without SEs (P value $= 1.326 \times 10^{-7}$, Kolmogorov–Smirnov test).

Profile of transcription factor binding at SE genomic regions. To plot the normalized tags-per-million transcription factor binding at SEs and their flanking 40 kb regions, we used the 'ngs.plot.r' package[31] (for example, Fig. 2a). To generate the enrichment of transcription factors at T_H-preferred SEs, we started by counting all tags in .bed files for each transcription factor binding using "bedtools coverage –counts" across the one map of SEs in T cells ($T_H1/T_H2/T_H17$ merged). Furthermore, in Fig. 2d we selected the T_H (1, 2, 17)-preferred SEs as genomic regions identified based on overlapping fraction = 0.1 identified in Fig. 1b. Extended Data Fig. 2a was generated by using ngs.plot on the same set of cell-type-specific coordinates. The normalization has been done as described previously[31]: the coverage data were normalized in two steps. In the first step, the coverage vectors were normalized to have equal length using spline fit. In this case, a cubic spline is fit through all data points and values are taken at equal intervals. This first step of length normalization leads to regions of variable sizes to have equal lengths and is particularly useful for custom regions. The purpose of the second step is to normalize vectors against the corresponding library size—that is, the total read count.

Profile of transcription-factor binding at constituent elements of SEs. We first recovered the original peak regions for p300 binding (constituent enhancers) within SEs from outputs of the peak-calling method (MACS) overlapping SEs/TEs. We then used the HOMER 'annotatePeaks.pl' function to plot the enrichment of transcription factor binding at constituent enhancers in SEs and TEs (Extended Data Fig. 2c).

GO analysis for SE-associated genes. In Fig. 2e, GO enrichment for SE genomic coordinates was carried out using GREAT[24] with default parameters. The top ten terms based on binomial P values were selected in Fig. 2e. In a completely different approach, we characterized the closest genes to SEs. The top GO molecular functions in terms of GSEA "Investigate Gene Sets" were then selected. To calculate the statistical significance of these enrichments, we randomly moved the SE regions around the genome 10^5 times, delineated the closest gene sets to the random genomic domains, and assessed the relative proportion of a gene set that is captured in the actual data versus the shifted SEs. P values for this permutation test are reported in Extended Data Fig. 3a.

GO functional category relevant to cytokine binding is enriched at SE-associated genes in T cells and to a lesser extent in macrophages but not in mouse ES cells and myotubes (Extended Data Fig. 3b). To explore whether 'cytokine binding' is specific to the SE structure in CD4$^+$ T cells, we explored its association within the SE structures of other cell types. The GO molecular function associated with cytokine binding (GO:0019955) was chosen. SE-associated genes in myotubes were used from Whyte *et al.*[4]. SE regions in mouse ES cells and macrophages were chosen based on data sets reported in Source Data associated with Fig. 1. To calculate the statistical significance of this gene category, we shuffled the SE regions of mouse ES cells, macrophages, myotubes and CD4$^+$ T cells around the genome 10^5 times, delineating the gene sets in proximity to the random genomic domains associated with each cell type. We then assessed the relative proportion of the gene set captured in the actual data versus the shifted SEs. P values for this permutation test are reported in the bar graph in Extended Data Fig. 3b.

Analysis of RNA-seq data from *Bach2*-deficient cells. The $\log_2$ fold change of average RPKM values in wild-type and knockout repeats were calculated for SE genes and an equal number of randomly selected TE and other genes in the violin plots (Fig. 3d and Extended Data Fig. 3d). In Extended Data Fig. 3c, d, the RPKM values for the spiked-in measurements were renormalized based on the spiked-in standards. We used ggplot and geom_violin(scale = "area") to plot the impact of loss of BACH2 on gene expression. All genes in SEs, TEs, or the rest of genes were used for the cumulative distribution plots (Fig. 3c and Extended Data Fig. 3c). In Extended Data Fig. 3e, we focused on the top 500 highly expressed genes and explored the effect of BACH2 on three categories among them: genes with SEs (77), with TEs (125), and without either SEs or TEs (298). Expression levels among these three categories of genes were comparable (Wilcoxon rank-sum test P value = 0.644). However, BACH2 selectively affected highly expressed SE genes in contrast to those with TEs or no enhancers (Kolmogorov–Smirnov test P value $= 9.813 \times 10^{-7}$ and 4.669×10^{-8}).

GSEA plot. The 'gene-set' for GSEA was generated based on genes closest to SEs with minimum 1 RPKM value in any of the three lineages (T_H1, T_H2, T_H17 cells). Three repeats for wild-type and *Bach2*-knockout RNA-seq data were used in the GSEA analysis with default settings (Fig. 3e). The P value for the enrichment was calculated as 0 although -nperm = 10000 was used (with command-line usage of GSEA). In the case of spiked-in GSEA analysis (Extended Data Fig. 4c), two repeats for wild-type and knockouts of renormalized spiked-in data were used.

Pie chart demonstrating *Bach2*-dependent SE genes. BACH2 up- or downregulated genes (Source Data associated with Fig. 3) were delineated by the 'DEseq' package in R with FDR < 0.05 and fold change > 1.5. Tag counts were calculated using "htseq-count–mode=intersection-nonempty–stranded=no" (Extended Data Fig. 4d). Three repeats of RNA-seq data for wild-type and knockout samples (no spiked-in) were used (Source Data associated with Fig. 1). Direct targets of BACH2 were identified based on BACH2 ChIP-seq data at these two groups of genes. A list of SE genes with at least 1 RPKM expression in T_H1, T_H2 or T_H17 cells was used for this analysis.

Characterization of BACH2-dependent noncoding RNAs. We used "bedtools coverage –counts" to quantitate the enrichment of RNA-seq reads at 501 ncRNAs with SE structure in wild-type and *Bach2*-knockout cells (Source Data associated with Fig. 3). Transcript levels were further normalized to the size of each library (tags per million) and the average of enrichment in three repeats were calculated. Next, we selected ncRNAs with SE structure that were up- or downregulated by BACH2 (>4 fold-change) (Fig. 3f).

Impact of transcription factors on SE- and TE-associated genes. The fold change in RPKM values between wild-type and knockout samples was calculated for SE genes and an equal number of randomly selected TE genes (Extended Data Fig. 4a). For each transcription factor, the difference between SEs and TEs was quantitated using Kullback–Leibler distance between the two distributions for fold changes in the two groups of genes using the KL.dist function in the FNN library in R (Extended Data Fig. 4b). The largest difference between SEs and TEs generated because of loss of BACH2, STAT4 and STAT6 suggests the more selective impact of these transcription factors on SEs.

Pruning SNPs. To ensure that the SNPs associated with disease are in physically independent segments of the genome, we pruned our lists of SNPs (Fig. 4a). Data from the 1000 Genomes (release 20110521) were downloaded from the 1000 Genomes open ftp site. SNPs that were present in each of the six disease conditions were extracted. For each disease, the all-versus-all pairwise r^2 values were calculated. Finally, all variants were greedily pruned until no pair had an r^2 value greater than the threshold (0.5). The number of SNPs pruned for each disease and their genomic coordinates can be found in Source Data associated with Fig. 4.

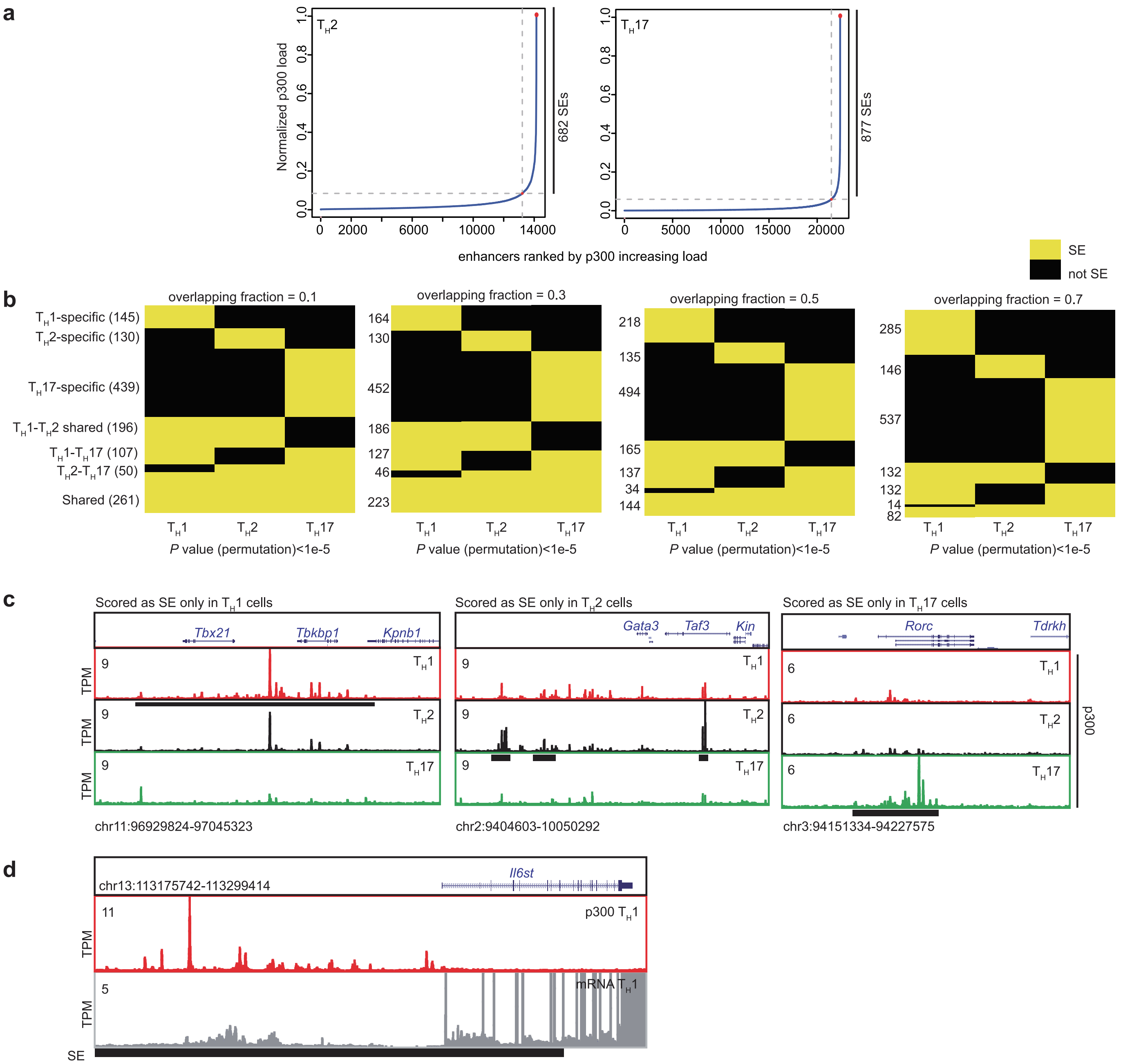
T-cell SEs in human and enrichment of SNPs. Human SEs in T-cell subsets were characterized based on H3K27ac data in T_H1, T_H2 and T_H17 cells (Source Data associated with Fig. 4). The methodology for the delineation of SEs for human T cells was the same as the one described for the mouse data. We referred to the merged map of the T_H1, T_H2 and T_H17 SEs as the single map of SEs in CD4$^+$ T cells (Source Data associated with Fig. 4). The lists of tag SNPs for all traits except RA were extracted from the GWAS catalogue (December 2013) and only those with P values less than 1×10^{-8} were selected. The list of 101 RA SNPs were chosen from the recent meta-analysis of RA GWASs[7]. The percentages of SNPs within SEs/TEs were calculated based on the number of SNPs falling into the genomic domains labelled as SEs/TEs. To account for the size of the genome that these two types of enhancers span, we divided the number of SNPs enriched in SEs/TEs by the total size of SEs (66.5338 Mb) and TEs (63.12915 Mb) and reported the number of SNPs in every 10 Mb of the genome in Fig. 4a. The permutation test for the enrichment P value was calculated by generating 10^6 permutations of SEs and TEs in the genome (excluding unmappable regions in each permutation) and considering the number of iterations where the number of overlapping SNPs with random SEs/TEs exceeded the observed ones in CD4$^+$ T cells. SNPs in linkage disequilibrium with the list of tag SNPs were determined from the 1000 Genomes Project using $r^2 = 0.9$ and distance limit = 500 using SNAP toolbox (Extended Data Fig. 5a).

RA risk genes and SEs. The list of 98 rheumatoid arthritis (RA) risk genes was extracted from the study of Plenge and colleagues[7] (Fig. 4b). H3K27ac data for muscle, CD14$^+$ and CD56$^+$ cells are summarized in Source Data associated with Fig. 4.

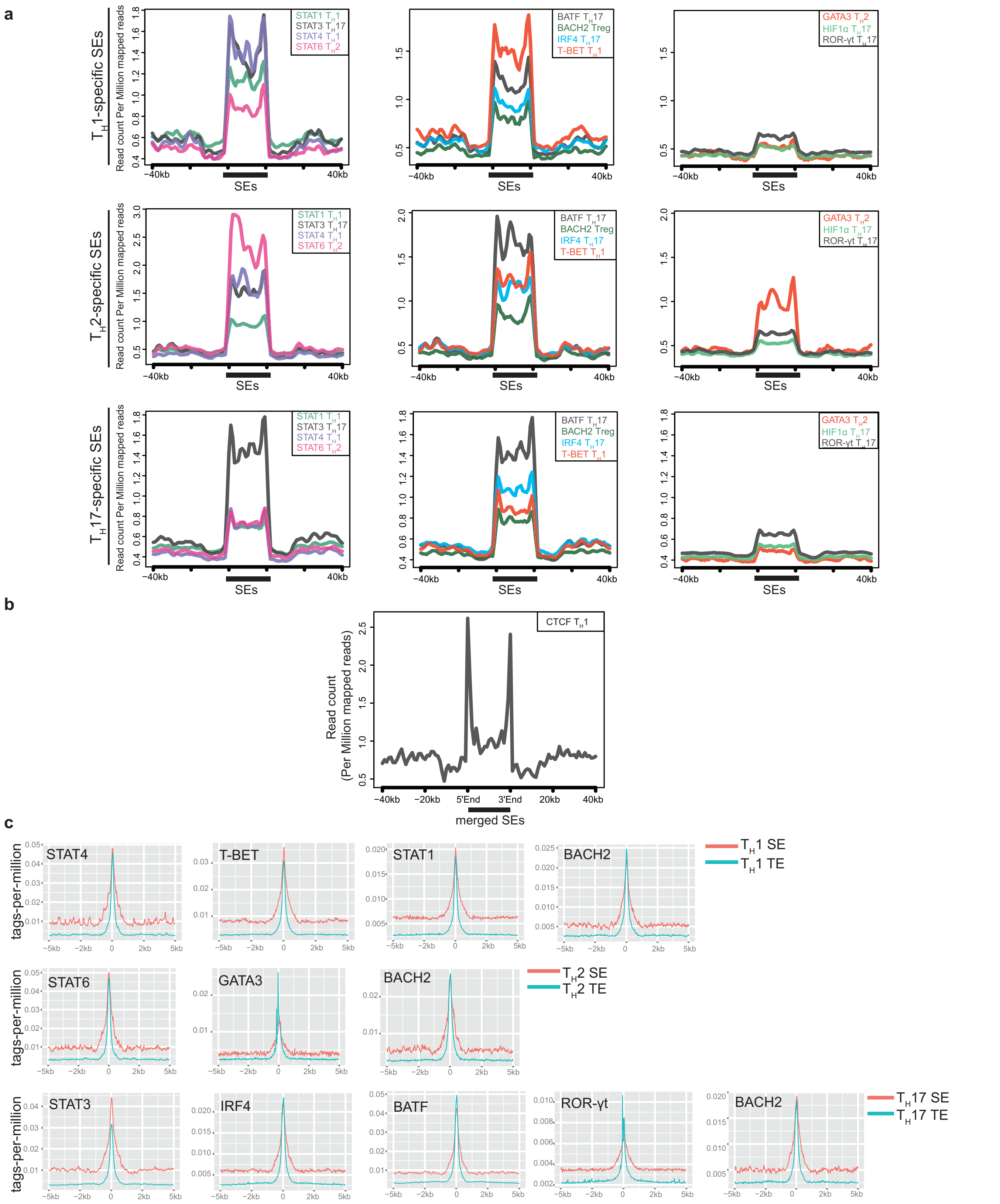
Quantitating the effect of tofacitinib on different groups of genes. For each donor (except donor 4), the RPKM values with spiked-in were renormalized and the fold changes at SE/TE genes were reported (Fig. 4c). No spiked-in was used for the RNA-seq analysis of donor 4. The *P* values were calculated based on Wilcoxon signed-rank test (wilcox.test function in R) for violin and box plots. The violin plots used 'scale = 'area''. In Extended Data Fig. 4c, for each donor, the top 100 highly expressed genes in non-treated RNA-seq data were selected and categorized as having SEs or not.

IBD, MS and T2D risk genes and SEs. The candidate genes associated with RA[7], inflammatory bowel disease (IBD)[22], multiple sclerosis (MS)[23] and type 2 diabetes (T2D)[32] were chosen based on a recent meta-analysis of GWAS data. More than half of RA risk genes (53/98) accommodated SEs in CD4$^+$ T cells. In line with the enrichment of SNPs associated with IBD and MS in T-cell SEs (Fig. 4a), around half of IBD (91/216) and MS risk genes (36/87) were associated with SEs in T cells. In contrast, T2D risk genes showed little association with SEs (4/65) (Fisher's exact test, *P* value = 0.4). RA and IBD risk genes with SEs are selectively targeted by a JAK inhibitor, tofacitinib. Cumulative plots depict the fold change in expression ($\log_2$) after 0.3 μM tofacitinib treatment of human CD4$^+$ T cells at RA (Extended Data Fig. 6b), IBD (Extended Data Fig. 6c) and MS (Extended Data Fig. 6d) risk genes with or without SEs.

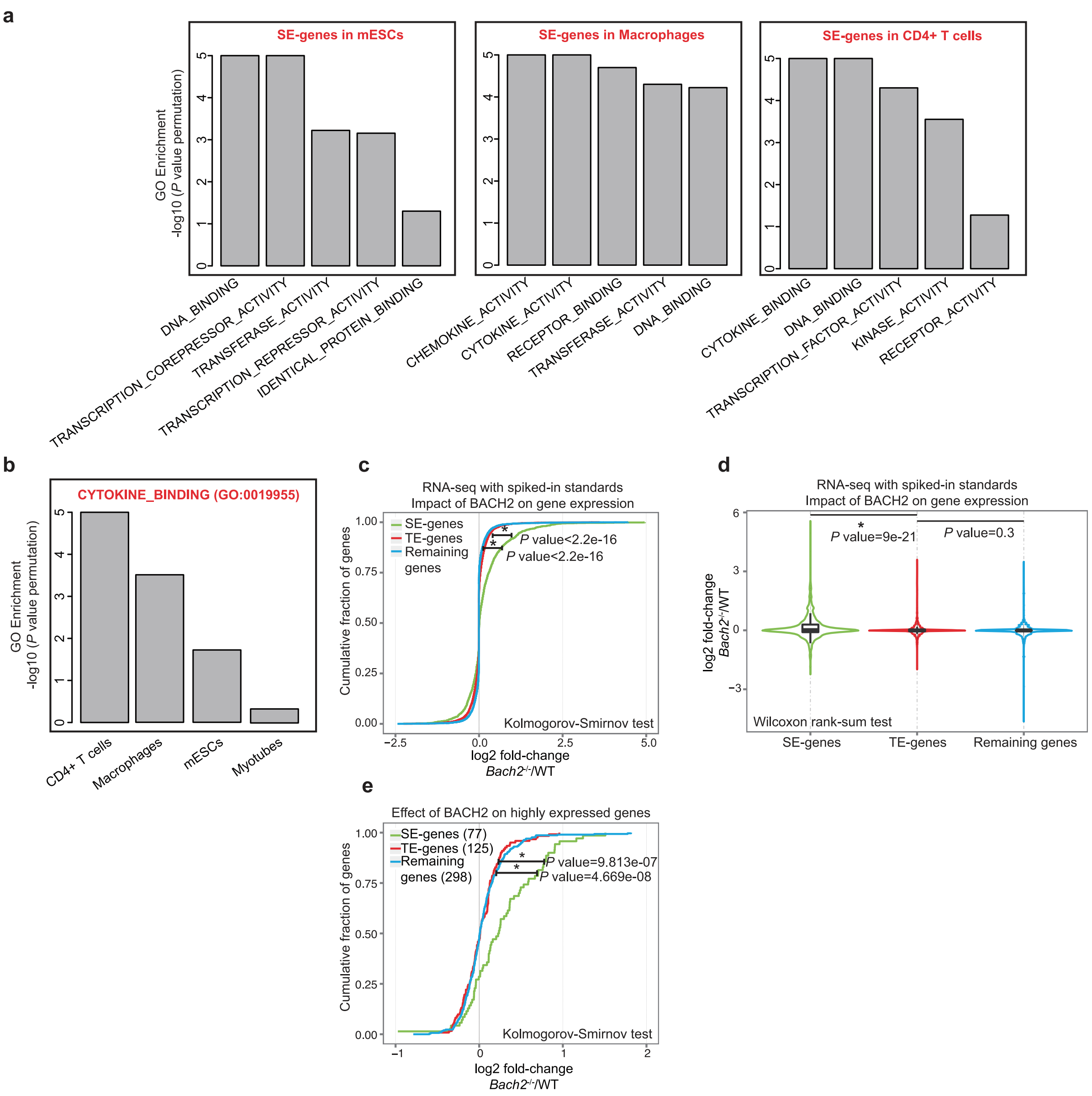
25. Ghisletti, S. *et al.* Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity* **32,** 317–328 (2010).
26. Nakayamada, S. *et al.* Early Th1 cell differentiation is marked by a Tfh cell-like transition. *Immunity* **35,** 919–931 (2011).
27. Wei, G. *et al.* Genome-wide analyses of transcription factor GATA3-mediated gene regulation in distinct T cell types. *Immunity* **35,** 299–311 (2011).
28. Hawkins, R. D. *et al.* Global chromatin state analysis reveals lineage-specific enhancers during the initiation of human T helper 1 and T helper 2 cell polarization. *Immunity* **38,** 1271–1284 (2013).
29. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
30. Zang, C. *et al.* A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25,** 1952–1958 (2009).
31. Shen, L., Shao, N., Liu, X. & Nestler, E. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics* **15,** 284 (2014).
32. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genet.* **44,** 981–990 (2012).



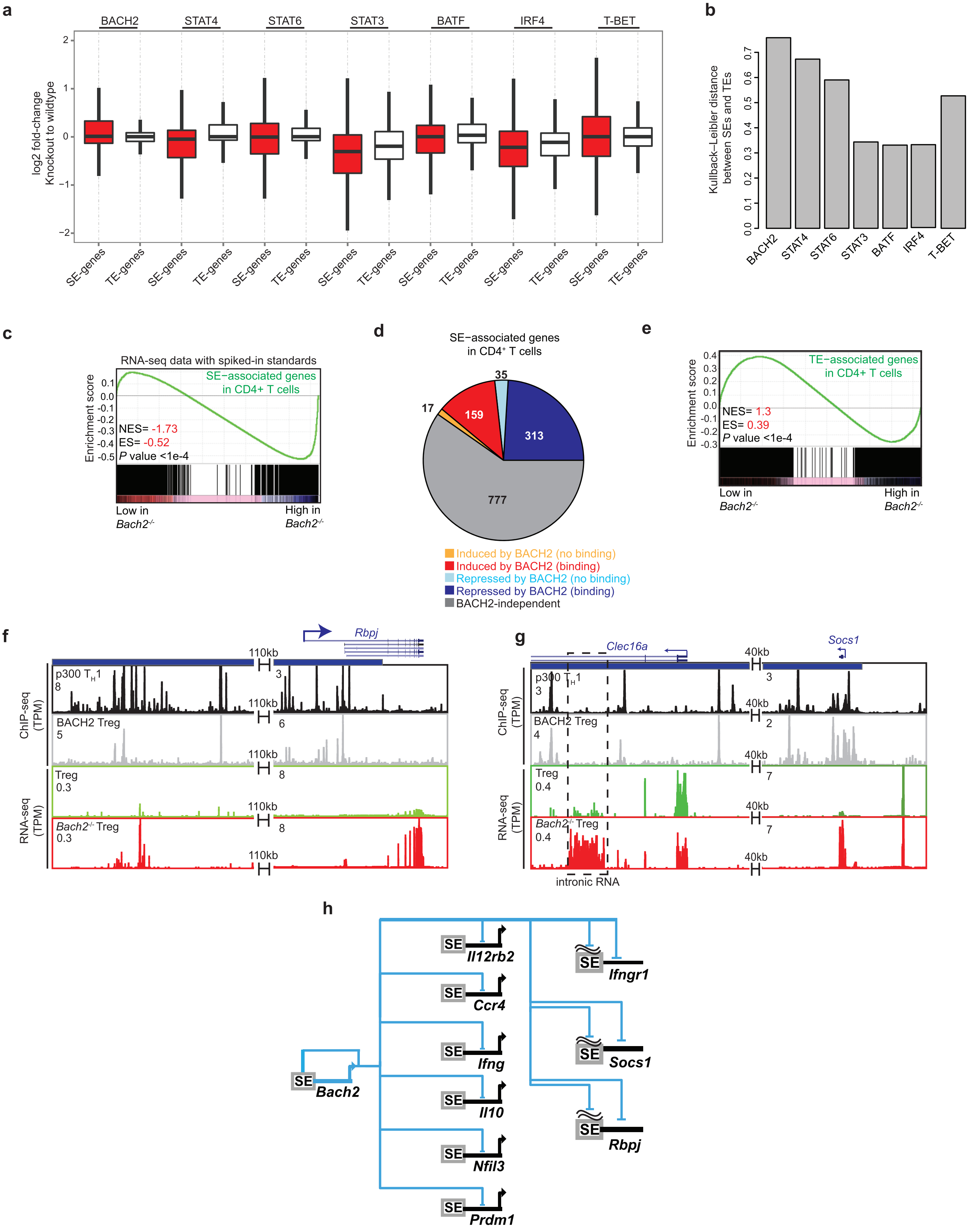
Extended Data Figure 1 | SE structures are lineage-specific. a, Histone acetyltransferase p300 is distributed asymmetrically across the genome in CD4$^+$ T cells with a subset of enhancers (SEs) containing exceptionally high amounts of p300 binding. Graph demonstrates the ranked distribution of p300 binding measured by ChIP-seq in T_H2 and T_H17 cells. **b**, Closely related CD4$^+$ T-cell populations have distinct SE landscapes. Common and cell-type-specific SE domains in T-cell subsets are illustrated for various fractions of overlapping genomic regions ($f = 0.1, 0.3, 0.5$ and 0.7). The overlapping pattern of SEs across CD4$^+$ T cells was statistically significant when these annotations were shuffled across the genome (P value $< 10^{-5}$). **c**, Lineage-specific presence of SEs for master transcription factor genes in T cells. Genomic loci of genes encoding T-BET, GATA3 and ROR-γt exhibit SE structures in T_H1, T_H2 and T_H17 cells, respectively. Black bar represents the genomic location of SEs. **d**, The genomic locus of the gene encoding gp130, *Il6st*, accommodates an SE with a high level of transcription. Black bar represents the genomic location of SEs.



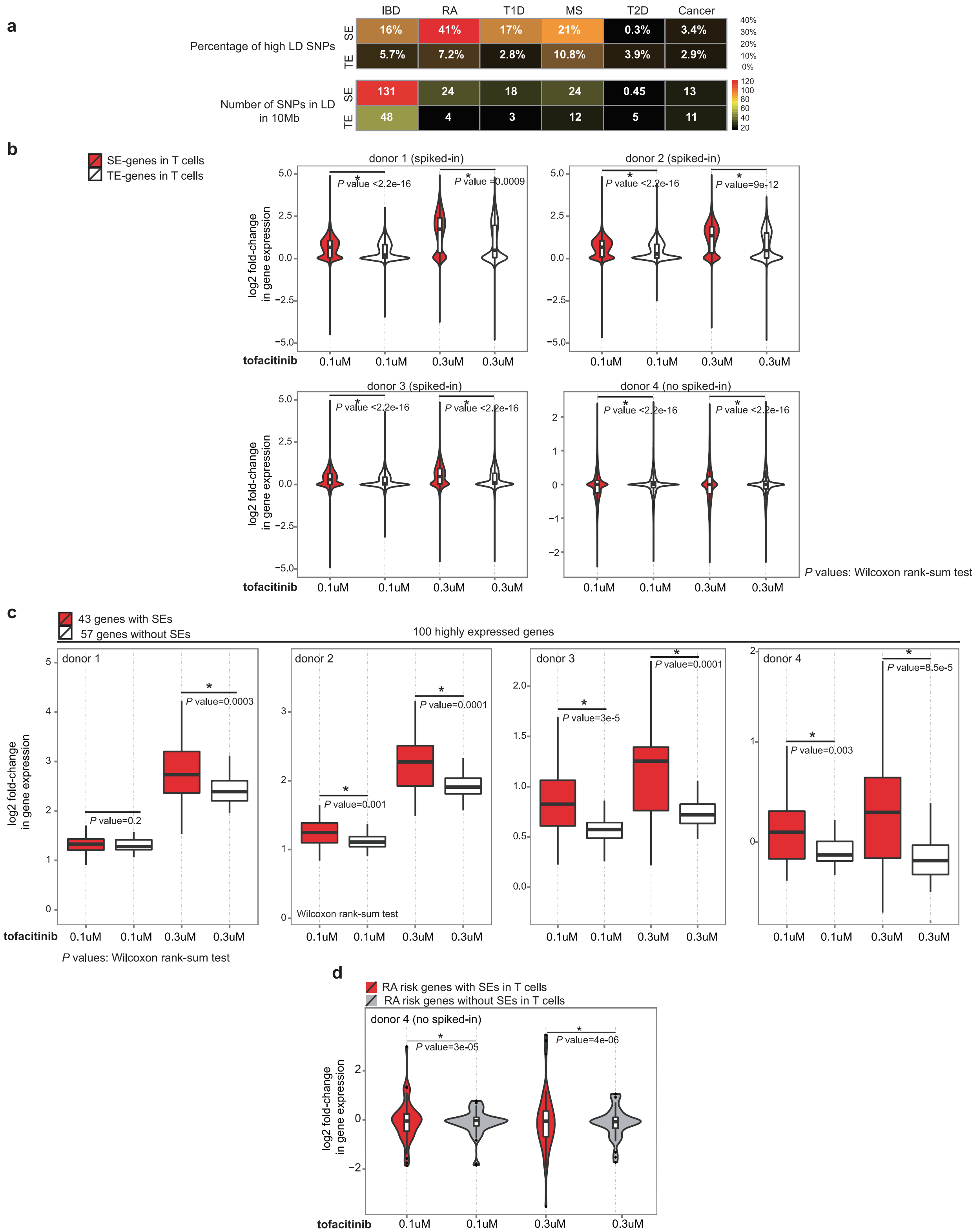
Extended Data Figure 2 | Transcription factor enrichment at SEs. **a**, Lineage-specific transcription factors are enriched at cell-type-specific SEs. Binding patterns of STAT4, STAT6 and STAT3 revealed preferential binding at T_H1-, T_H2- and T_H17-specific SE regions, respectively. Furthermore, transcription factors T-BET, GATA3, and HIF-1α and ROR-γt were enriched at lineage-specific SEs. Strong binding of BATF, BACH2 and IRF4 was present in SEs of all three cell types. Maps of cell-type-specific SEs were constructed as described in Fig. 1b. Normalization of *y*-axis takes into account the variable sizes of genomic regions and also the corresponding library size (that is, the total read count) (Methods). **b**, CTCF binding demarcates the boundaries of SEs. Normalized binding profile of CTCF protein revealed the enrichment of CTCF at boundaries of SE regions. **c**, Comparing the enrichment of transcription factors at constituent enhancers of SEs and TEs reveals the preferential binding of STAT3 at SEs while other transcription factors demonstrated comparable binding at SEs and TEs.



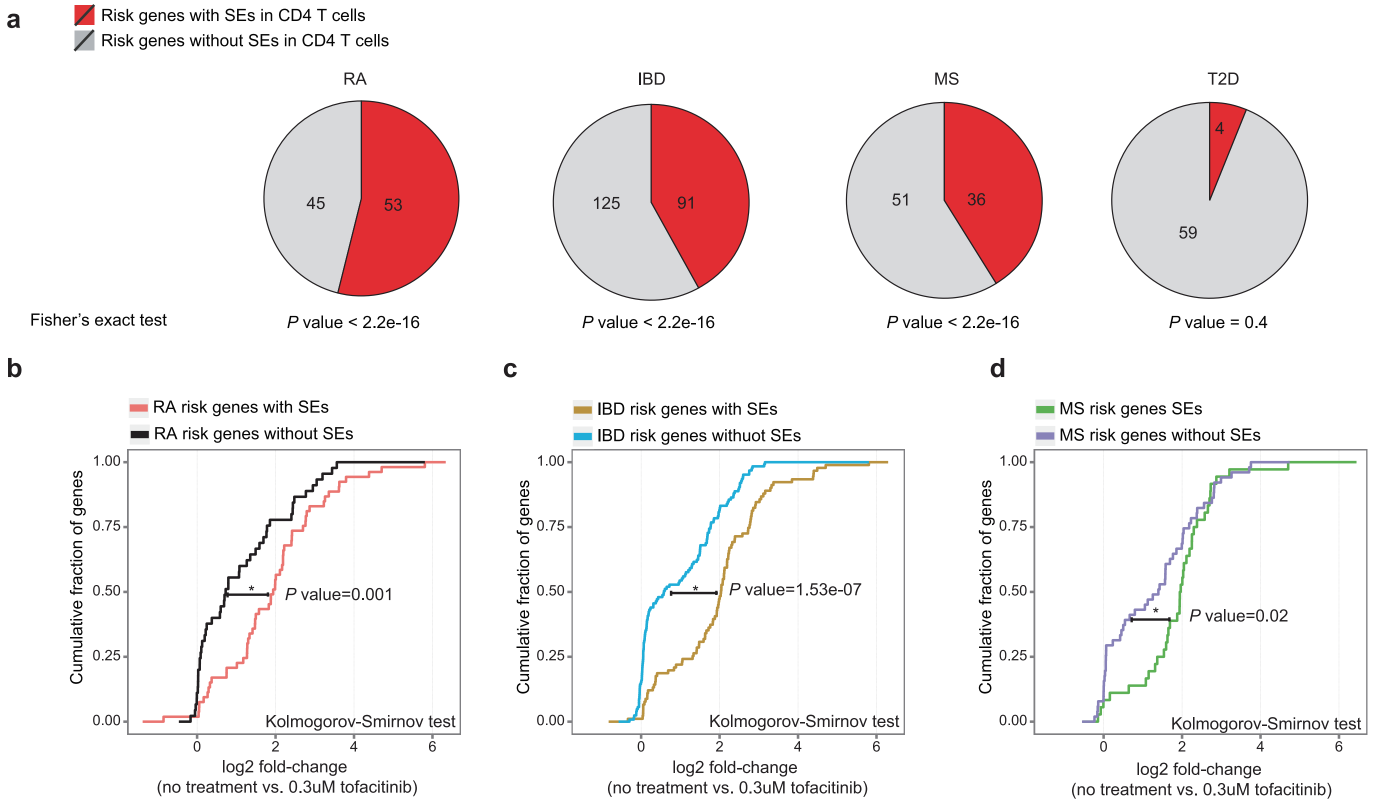
Extended Data Figure 3 | Identity of SE-associated genes. **a**, SEs delineate genes that have a central role in the biology of specific cell lineages. Gene ontology (GO) functional categories relevant to cytokine binding are enriched at SE-associated genes in T cells. In ES cells, SE structures primarily encompass DNA-binding proteins and transcriptional repressor functions. In macrophages, chemokine and cytokine activity were the most prominent categories. Using a complementary approach to that described in Fig. 2a, we characterized genes in proximity to SEs. The top GO molecular functions using GSEA were chosen. To calculate the statistical significance of these gene categories, we shuffled the SE regions around the genome 10^5 times, delineating the gene sets in proximity to the random genomic domains. We then assessed the relative proportion of a gene set captured in the actual data versus the shifted SEs. $-\mathrm{Log}_{10}$ *P* values for this permutation test are reported in the bar graph. **b**, GO functional category relevant to cytokines binding is enriched at SE-associated genes in T cells and, to a lesser extent, in macrophages but not in mouse ES cells (mESCs) or myotubes. To explore whether “cytokine binding” is specific to the SE structure in $\mathrm{CD4^+}$ T cells, we explored its association within the SE structures of other cell types. The GO molecular function associated with cytokine binding (GO:0019955) was chosen. To calculate the statistical significance of this gene category, we shuffled the SE regions of mouse ES cells, macrophage, myotubes and $\mathrm{CD4^+}$ T cells around the genome 10^5 times, delineating the gene sets in proximity to the random genomic domains associated with each cell type. We then assessed the relative proportion of the gene set captured in the actual data versus the shifted SEs. *P* values for this permutation test are reported in the bar graph. **c**, **d**, BACH2 preferentially represses SE genes. Wild-type and *Bach2*-deficient $\mathrm{CD4^+}$ T cells were polarized to induced regulatory T cells ($\mathrm{iT_{reg}}$ cells) and were subjected to total RNA extraction. RNA standards ‘spiked-in’ were added in proportion to the number of cells present in the sample. The resulting transcriptome data measured by RNA-seq were processed by using standard normalization methods and then renormalized based on spiked-in reads (RPKM) (see Methods). Transcript abundance measured by RNA-seq was evaluated in wild-type and *Bach2*-deficient cells at SE- and TE-associated genes compared to remaining genes (RPKM). Cumulative distribution (**c**) and violin plots (**d**) show the ($\log_2$) fold change in gene expression for wild-type versus *Bach2*-deficient cells for these three groups of genes. SE genes are preferentially affected by loss of BACH2 compared with TE genes (*P* value $< 2.2 \times 10^{-16}$, Kolmogorov–Smirnov test) or remaining genes (*P* value $< 2.2 \times 10^{-16}$, Kolmogorov–Smirnov test). *P* values for the violin plots (**d**) were calculated using the Wilcoxon rank-sum test. **e**, BACH2 selectively affects SE genes and such selectivity remains statistically significant when controlling for the higher levels of gene expression for the SE genes. Genes were ranked based on their transcriptional activity in $\mathrm{T_{reg}}$ cells. We focused on the top 500 highly expressed genes and explored the effect of BACH2 on three categories among them: genes with SEs (77), with TEs (125), and without either SEs or TEs (298). Expression levels among these three categories of genes were comparable (Wilcoxon rank-sum test, *P* value = 0.644). However, BACH2 selectively affected highly expressed SE genes in contrast to those with TEs or no enhancers (Kolmogorov–Smirnov test, *P* value $= 9.813 \times 10^{-7}$ and 4.669×10^{-8}).


a
BACH2
STAT4
STAT6
STAT3
BATF
IRF4
T-BET
log2 fold-change
Knockout to wildtype
2
1
0
−1
−2
SE-genes
TE-genes
b
Kullback–Leibler distance
between SEs and TEs
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
BACH2
STAT4
STAT6
STAT3
BATF
IRF4
T-BET
c
RNA-seq data with spiked-in standards
SE−associated genes
in CD4+ T cells
Enrichment score
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
NES= -1.73
ES= -0.52
P value <1e-4
Low in
Bach2-/-
High in
Bach2-/-
d
SE−associated genes
in CD4+ T cells
35
17
159
313
777
Induced by BACH2 (no binding)
Induced by BACH2 (binding)
Repressed by BACH2 (no binding)
Repressed by BACH2 (binding)
BACH2-independent
e
TE−associated genes
in CD4+ T cells
Enrichment score
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
NES= 1.3
ES= 0.39
P value <1e-4
Low in
Bach2-/-
High in
Bach2-/-
f
Rbpj
110kb
ChIP-seq
(TPM)
p300 TH1
8
3
BACH2 Treg
5
6
RNA-seq
(TPM)
Treg
0.3
8
Bach2-/- Treg
0.3
8
g
Clec16a
Socs1
40kb
ChIP-seq
(TPM)
p300 TH1
3
3
BACH2 Treg
4
2
RNA-seq
(TPM)
Treg
0.4
7
Bach2-/- Treg
0.4
7
intronic RNA
h
SE
Il12rb2
Ccr4
Ifng
Il10
Nfil3
Prdm1
Bach2
Ifngr1
Socs1
Rbpj

Extended Data Figure 4 | BACH2 acts as a guardian transcription factor. **a**, **b**, Loss of BACH2, STAT4 and STAT6 has the most selective impact on the expression of SE genes. **a**, The fold change in expression (in RPKM) between wild-type and knockout samples was calculated for SE genes and an equal number of randomly selected TE genes. **b**, For each transcription factor, the difference between SEs and TEs was quantitated using Kullback–Leibler distance. The larger difference between SEs and TEs for BACH2, STAT4 and STAT6 suggests the more selective impact of these transcription factors on SEs. STAT4 and T-BET transcriptome data were under T_H1, STAT6 under T_H2, STAT3, BATF and IRF4 under T_H17 and BACH2 under iT_{reg} conditions. **c**, SE-associated genes in $CD4^+$ T cells are repressed by BACH2. To ensure accurate inference of the effect of BACH2 on the transcriptome, spiked-in RNA standards were added. The gene set enrichment analysis (GSEA) of SE-associated genes revealed that SE genes were enriched in genes repressed by BACH2 when transcript levels were renormalized using spiked-in RNA standards. **d**, BACH2 acts as a repressor of SE-associated genes. Comparison of the transcriptome data measured by RNA-seq in wild-type and *Bach2*-deficient cells (DE-seq analysis for three wild-type and knockout samples, FDR < 0.05 and fold change > 1.5) revealed that 348 SE genes were repressed while 176 were induced by this protein. Integration of BACH2 binding data measured by ChIP-seq characterized the direct targets of BACH2. **e**, The GSEA of TE-associated genes revealed that TE genes are not enriched in genes repressed by BACH2. **f**, **g**, BACH2-associated transcriptional repression at some SE domains correlates with the downregulation of nearby genes such as *Rbpj* (**f**) and *Socs1* (**g**). **h**, Genes and noncoding transcripts endowed with SE architecture in $CD4^+$ T cells are tightly and negatively controlled by the 'guardian' transcription factor BACH2, which itself has a rich cassette of regulatory elements. Examples were selected based on direct binding of BACH2 at the gene body or SE regions measured by ChIP-seq.



Extended Data Figure 5 | RA risk genes with SE structure are selectively targeted by a JAK inhibitor, tofacitinib. a, Genetic variants in high linkage disequilibrium (LD) with SNPs associated with autoimmune disorders such as RA, IBD, MS and T1D exhibit preferential enrichment in SEs versus TEs in human CD4 T cells. Variants in LD with SNPs in each disease were determined from the 1000 Genomes Project using $r^2 = 0.9$ and distance limit = 500 by SNAP toolbox. The heatmap depicts the percentages of SNPs and total number of SNPs per 10 Mb within SEs and TEs. **b**, Tofacitinib treatment has a selective impact on SE versus TE genes in human T cells. Violin plots depict the fold change ($\log_2$) in transcript levels due to tofacitinib treatment at SE versus TE genes in CD4$^+$ T cells. The *P* values were calculated based on the Wilcoxon signed-rank test. **c**, Highly expressed genes in T cells with SEs are selectively affected by tofacitinib. For each donor, the top 100 highly expressed genes in non-treated cells were selected and categorized as having SEs or not. The *P* values were calculated based on the Wilcoxon signed-rank test. **d**, RA risk genes with SEs are selectively targeted by a JAK inhibitor, tofacitinib. Violin plots depict the fold change ($\log_2$) after tofacitinib treatment of human CD4$^+$ T cells at RA risk genes with or without SEs (a donor with no spiked-in standard in RNA-seq). *P* values were calculated using the F-test.



Extended Data Figure 6 | Tofacitinib selectively affects autoimmune disease risk genes with SE structure in T cells. **a**, RA, IBD, and MS risk genes are linked to SEs in $CD4^+$ T cells. The candidate genes associated with RA[7], IBD[22], MS[23] and T2D[32] were chosen based on recent meta-analyses of GWAS data. More than half of RA risk genes (53/98) contained SEs in $CD4^+$ T cells. In line with the enrichment of SNPs associated with IBD and MS in T-cell SEs (Fig. 4a), around half of IBD (91/216) and MS risk genes (36/87) were associated with SEs in T cells. In contrast, T2D risk genes showed little association with SEs (4/65) (Fisher's exact test, *P* value = 0.4). **b**–**d**, RA and IBD risk genes with SEs are selectively targeted by a JAK inhibitor, tofacitinib. Cumulative plots depict the fold change in expression ($\log_2$) at RA (**b**), IBD (**c**) and MS (**d**) risk genes with or without SEs after 0.3 μM tofacitinib treatment of human $CD4^+$ T cells (*P* values, Kolmogorov–Smirnov test).