Chromosome contacts in activated T cells identify autoimmune disease-candidate genes

Oliver S Burren[1,2]*, Arcadio Rubio García[2,3]*, Biola-Maria Javierre[4]*, Daniel B Rainbow[2,3]*, Jonathan Cairns[4], Nicholas J Cooper[2], John J Lambourne[5], Ellen Schofield[2], Xaquin Castro Dopico[2], Ricardo C Ferreira[2,3], Richard Coulson[2], Frances Burden[5,6], Sophia P Rowlston[5,6], Kate Downes[5,6], Steven W Wingett[4], Mattia Frontini[5,6,7], Willem H Ouwehand[5,6,7,8], Peter Fraser[4], Mikhail Spivakov[4], John A Todd[2,3#], Linda S Wicker[2,3#], Antony J Cutler[2,3#], Chris Wallace[1,2,9#]

[1]Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge, CB2 0SP, UK

[2]JDRF/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, NIHR Cambridge Biomedical Research Centre, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, CB2 0XY, UK

[3]Current address: JDRF/Wellcome Trust Diabetes and Inflammation Laboratory, Nuffield Department of Medicine, University of Oxford, The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK

[4]Nuclear Dynamics Programme, The Babraham Institute, Babraham Research Campus, Cambridge, CB22 3AT, UK

[5]Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Long Road, Cambridge, CB2 0PT, UK

[6]National Health Service Blood and Transplant, Cambridge Biomedical Campus, Long Road, Cambridge, CB2 0PT, UK

[7]British Heart Foundation Centre of Excellence, Division of Cardiovascular Medicine, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 0QQ, UK

[8]Department of Human Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1HH, UK

[9]MRC Biostatistics Unit, University of Cambridge, Cambridge Institute of Public Health, Cambridge Biomedical Campus, Cambridge, CB2 0SR, UK

*These authors contributed equally to this work.

[#]These authors contributed equally to this work.

Autoimmune disease-associated variants are preferentially found in regulatory regions in immune cells, particularly $CD4^+$ T cells. Linking such regulatory regions to gene promoters in disease-relevant cell contexts facilitates identification of candidate disease genes. Here we show that, within four hours, activation of $CD4^+$ T cells invokes changes in histone modifications and enhancer RNA transcription that correspond to altered expression of the interacting genes identified by promoter capture Hi-C (PCHi-C). By integrating PCHi-C data with genetic associations for five autoimmune diseases we prioritised 252 candidate genes, of which 116 were related to activation-sensitive interactions. This included *IL2RA*, where allele-specific expression analyses were consistent with its interaction-mediated regulation, illustrating the utility of the approach.

Genome-wide association studies (GWAS) in the last decade have associated 324 distinct genomic regions to at least one and often several autoimmune diseases (http://www.immunobase.org). The majority of associated variants lie outside genes[1] and presumably tag regulatory variants acting on nearby or more distant genes[2,3]. Progress from GWAS discovery to biological interpretation has been hampered by lack of systematic methods to define the gene(s) regulated by a given variant. The use of Hi-C[4] and capture Hi-C to link GWAS identified variants to their target genes for breast cancer[5] and autoimmune diseases[6] using cell lines, has highlighted the potential for mapping long range interactions in advancing our understanding of disease association. The observed cell specificity of these interactions indicates a need to study primary disease-relevant human cells, and investigate the extent to which cell state may affect inference.

Integration of GWAS signals with cell-specific chromatin marks has highlighted the role of regulatory variation in immune cells[7], and in particular $CD4^+$ T cells, in autoimmune diseases[8]. $CD4^+$ T cells are at the centre of the adaptive immune system and exquisite control of activation is required to guide $CD4^+$ T cell fate through selection, expansion and differentiation into one of a number of specialised subsets. Additionally, the prominence of variants in physical proximity to genes associated with T cell regulation in autoimmune disease GWAS and the association of human leukocyte antigen haplotypes have suggested that control of T cell activation is a key etiological pathway in development of many autoimmune diseases[9].

Here, we explored the effect of activation on $CD4^+$ T cell gene expression, chromatin states and chromosome conformation. PCHi-C was used to map promoter interacting regions (PIRs), and to relate activation-induced changes in gene expression to changes in chromosome conformation and transcription of PCHi-C linked enhancer RNAs (eRNAs). We also fine mapped the most probable causal variants for five autoimmune diseases, autoimmune thyroid disease (ATD), coeliac disease (CEL), rheumatoid

arthritis (RA), systemic lupus erythematosus (SLE) and type 1 diabetes (T1D). We integrated these sources of information to derive a systematic prioritisation of candidate autoimmune disease genes.

Results

A time-course expression profile of early CD4$^+$ T cell activation

We profiled gene expression in CD4$^+$ T cells from 20 individuals across a 21 hour activation time-course, and identified eight distinct gene modules by clustering these profiles (**Fig. 1, Supplementary Table 1**). This time-course focused on much earlier events than previous large time-course studies (eg 6 hours - 8 days[10]) and highlights the earliest changes that are either not seen or are returning towards baseline by 6 hours (**Supplementary Fig. 1**). Gene set enrichment analysis using MSigDB Hallmark gene sets[11] demonstrated that these modules captured temporally distinct aspects of CD4$^+$ T cell activation. For example, negative regulators of TGF-beta signalling were rapidly upregulated, and returned to baseline by 4 hours. Interferon responses, inflammatory responses and IL-2 and STAT5 signalling pathways showed a more sustained upregulation out beyond 6 hours, while fatty acid metabolism was initially downregulated in favour of oxidative phosphorylation.

PCHi-C captures dynamic enhancer-promoter interactions

We examined activated and non-activated CD4$^+$ T cells in more detail at the four-hour time point, at which the average fold change of genes related to cytokine signalling and inflammatory response was maximal, using total RNA sequencing, histone mark chromatin immunoprecipitation sequencing (ChIP-seq) and PCHi-C. Of 8,856 genes identified as expressed (see Methods) in either condition (non-activated or activated), 25% were up- or down-regulated (1,235 and 952 genes respectively, FDR<0.01, **Supplementary Table 2**). We used PCHi-C to characterise promoter interactions in activated and non-activated CD4$^+$ T cells. Our capture design baited 20,676 *Hind*III fragments (median length 4 kb) which contained the promoters of 29,131 annotated genes, 18,202 of which are protein coding (**Supplementary Table 3**). We detected 283,657 unique PCHi-C interactions with the CHiCAGO pipeline[12], with 55% found in both activation states, and 22% and 23% found only in non-activated and

only in activated $CD4^+$ T cells, respectively (**Supplementary Table 4**). 11,817 baited promoter fragments were involved in at least one interaction, with a median distance between interacting fragments of 324 kb. Each interacting promoter fragment connected to a median of eight promoter interacting regions (PIRs), while each interacting PIR was connected to a median of one promoter fragment (**Supplementary Fig. 2**).

We compared our interaction calls to an earlier ChIA-PET dataset from non-activated $CD4^+$ T cells[13] and found we replicated over 50% of the longer range interactions (100 kb or greater), with replication rates over ten-fold greater for interactions found in non-activated $CD4^+$ T cells compared to interactions found only in erythroblasts or megakaryocytes (**Supplementary Fig. 3**). We also compared histone modification profiles in interacting fragments in $CD4^+$ T cells to interacting fragments found in erythroblasts or megakaryocytes. Both promoter fragments and, to a lesser extent, PIRs were enriched for histone modifications associated with transcriptionally active genes and regulatory elements (H3K27ac, H3K4me1, H3K4me3; **Supplementary Fig. 4**), and changes in H3K27ac modifications at both promoter fragments and PIRs correlated with changes in gene expression upon activation. PIRs, but not promoter fragments, showed significant overlap with regions previously annotated as enhancers[14].

We found that absolute levels of gene expression correlated with the number of PIRs (**Supplementary Fig. 5a**, rho=0.81), consistent with recent observations[13]. We defined a subset of PCHi-C interactions that were specifically gained or lost upon activation (2,334 and 1,866 respectively, FDR<0.01) and found that the direction of change (gain or loss) at these differential interactions agreed with the direction of differential expression (up- or down-regulated) at the module level (**Fig. 2**). We further found that dynamic changes in gene expression upon activation correlated with changing numbers of PIRs. Notably, the effect was asymmetrical, with a gained interaction having approximately twice the effect of a lost

interaction (**Fig. 3a**). Given the >6 hour median half life of mRNAs expressed in T cells[15] (**Supplementary Fig. 5b)**, it is possible that the relatively weaker effects of lost interactions are due to the persistence of downregulated transcripts at the early stages of T cell activation.

As we sequenced total RNA without a poly(A) selection step, we were able to detect regulatory region RNAs (regRNAs), which are generally non-polyadenylated and serve as a mark for promoter and enhancer activity[16]. We defined 6,147 "expressed" regRNAs (see Methods) that mapped within regulatory regions defined by a 15 state ChromHMM[17] model (**Supplementary Fig. 6**) and validated them by comparison to an existing cap analysis of gene expression (CAGE) dataset[18] which has been successfully used to catalog active enhancers.[19] We found 2,888/3,897 (74%) regRNAs expressed in non-activated cells overlap CAGE defined enhancers. This suggests that the combination of chromatin state annotation and total RNA-seq data presents an alternative approach to capture active enhancers.

Almost half (48%) of expressed regRNAs showed differential expression after activation (2,254/681 up/down regulated; FDR<0.01). To determine whether activity at these regRNAs could be related to that at PCHi-C linked genes, we focused attention on a subset of 640 intergenic regRNAs, which correspond to a definition of eRNAs[20]. Of these, 404 (63%) overlapped PIRs detected in $CD4^+$ T cells and we found significant agreement in the direction of fold changes at eRNAs and their PCHi-C linked protein coding genes in activated $CD4^+$ T cells ($p<0.0001$, **Fig. 3b**). We also observed a synergy between eRNA expression and the effect of a PIR on expression with a gain or loss of a PIR overlapping a differentially regulated eRNA having the strongest effect on gene expression (**Fig. 3c**), supporting a sequential model of gene activation[21]. While eRNA function is still unknown[20], our results demonstrate the detection, by PCHi-C, of condition-specific connectivity between promoters and enhancers involved coordinating gene regulation.

PCHi-C-facilitated mapping of candidate disease causal genes

We defined an experimental framework to integrate PCHi-C interactions with GWAS data to map candidate disease causal genes for autoimmune diseases (**Fig. 4**). First, to confirm that PCHi-C interactions inform interpretation of autoimmune disease GWAS, we tested whether PIRs were enriched for autoimmune disease GWAS signals in $CD4^+$ T cells, compared to non-lymphoid PIRs. We used *blockshifter*, which accounts for correlation between (1) neighbouring variants in the GWAS data and (2) neighbouring *Hind*III fragments in the interacting data by rotating one dataset with respect to the other. This method appropriately controls type 1 error rates, in contrast to methods based on counting associated SNP/PIRs which ignore correlation, such as a Fisher's exact test (**Supplementary Fig. 7**). We found autoimmune GWAS signals were enriched in $CD4^+$ T cell PIRs compared to non-autoimmune GWAS signals (Wilcoxon $p = 2.5x10^{-7}$) and preferentially so in activated *versus* non-activated cells (Wilcoxon $p = 4.8x10^{-5}$; **Fig. 4**).

Next, we fine-mapped causal variants for five autoimmune diseases using genetic data from a dense targeted genotype array, the ImmunoChip (ATD, CEL, RA, T1D), and summary data from GWAS data imputed to 1000 Genomes Project (RA, SLE; **Supplementary Table 5**). For the ImmunoChip datasets, with full genotype data, we used a Bayesian fine mapping approach[22] which avoids the need for stepwise regression or assumptions of single causal variants and which provides a measure of posterior belief that any given variant is causal by aggregating posterior support over models containing that variant. Variant-level results are given in **Supplementary Table 6**, and show that of 73 regions with genetic association signals to at least one disease (106 disease associations), ten regions have strong evidence that they contain more than one causal variant (posterior probability > 0.5), among them the well studied region on chromosome 10 containing the candidate gene *IL2RA*[22]. For the GWAS summary data, we

make the simplifying assumption that there exists a single causal variant in any LD-defined genetic region and again generate posterior probabilities that each variant is causal[23]. To integrate these variant level data with PCHi-C interactions and prioritize protein coding genes as candidate causal genes for each autoimmune disease, we calculated gene-level posterior support by summing posterior probabilities over all models containing variants in PIRs for a given gene promoter, within the promoter fragment or within its immediate neighbour fragments. Neighbouring fragments are included because of limitations in the ability of PCHi-C to detect very proximal interactions (within a region consisting of the promoter baited fragment and one *Hind*III fragment either side). The majority of gene scores were close to 0 (99% of genes have a score <0.05) and we chose to use a threshold of 0.5 to call genes prioritised for further investigation. Having both ImmunoChip and summary GWAS data for RA allowed us to compare the two methods. Overlap was encouraging: of 24 genes prioritised for RA from ImmunoChip, 19 had a gene score > 0.5 in the GWAS prioritisation, a further four had GWAS scores > 0.3. The remaining gene, *MDN1,* corresponded to a region which has a stronger association signal in the RA-ImmunoChip than RA-GWAS dataset, which may reflect the greater power of direct genotyping versus imputation, given that the RA-ImmunoChip signal is mirrored in ATD and T1D (**Supplementary Fig. 8**). We prioritised a total of 252 unique protein coding genes, 116 of which related to activation sensitive interactions (**Supplementary Table 7**, **Fig. 4)**. Of 135 prioritised genes which could be related through interactions to a known susceptibility region, 64 (48%) lay outside that disease susceptibility region. The median distance from peak signal to prioritised gene was 152 kb. Note that prioritisation can be one (variant)-to-many (genes) because a single PIR can interact with more than one promoter, and promoter fragments can contain more than one gene promoter. Note also that the score reflects both PCHi-C interactions and the strength and shape of association signals (**Supplementary Fig. 9**), therefore a subset of prioritised genes relate to an aggregation over sub-genomewide significant GWAS signals. This is therefore a "long" list of prioritised genes which requires further filtering (**Table 1**). One hundred and

eighty six (of 252) prioritised genes were expressed in at least one activation state; we highlight specifically the subset of 120 expressed genes which can be related to a genome-wide significant GWAS signal through proximity of a genome-wide significant SNP ($p<5x10^{-8}$) to a PIR. Of these, 83 were differentially expressed, 49 related to activation-sensitive interactions and 29 showed overlap of GWAS fine-mapped variants with an expressed eRNA (**Supplementary Table 7**).

Taken together, our results reflect the complexity underlying gene regulation, and the context-driven effects that common autoimmune disease-associated variants may have on candidate genes. Our findings are consistent with, and extend, previous observations[7,8] and we highlight six examples which exemplify both activation-specific and activation-invariant interactions.

PCHi-C may prioritise additional genes lying some distance from peak association signals. For example, CEL has been associated with a region on chromosome 1q31.2, for which *RGS1* has been named as a causal candidate due to proximity of associated variants to its promoter[24]. Sub-genome-wide significant signals for T1D (min. $p=1.5x10^{-6}$) across the same SNPs which are associated with CEL have been interpreted as a colocalising T1D signal in the region[25]. *RGS1* has recently been shown to have a role in the function of T follicular helper cells in mice[26], the frequencies of which and their associated IL-21 production have been shown to be increased in T1D patients[27]. However, our analysis also prioritises, in activated T cells, the strong functional candidate genes *TROVE2* and *UCHL5*, over half a megabase distant and with three intervening genes not prioritised, for CEL and T1D (**Fig. 5**). *UCHL5* encodes ubiquitin carboxyl-terminal hydrolase-L5 a deubiquitinating enzyme that stabilizes several Smad proteins and TGFBR1, key components of the TGF-beta1 signalling pathway[28,29]. *TROVE2* is significantly upregulated upon activation (FDR=0.005) and encodes Ro60, an RNA binding protein that indirectly regulates type-I IFN-responses by controlling endogenous Alu RNA levels[30]. A global anti-inflammatory effect for *TROVE2* expression would fit with its effects on gut (CEL) and pancreatic islets (T1D).

A similar situation is seen in the chromosome 1q32.1 region associated with T1D in which *IL10* has been named as a causal candidate gene[31]. Together with *IL10*, prioritised through proximity of credible SNPs to the *IL10* promoter, we prioritised other *IL10* gene family members *IL19*, *IL20* and *IL24* as well as two members of a conserved three-gene immunoglobulin-receptor cluster (*FCMR* and *PIGR,* **Supplementary Fig. 10**). While *IL19, IL20* and *PIGR* were not expressed in $CD4^+$ T cells, *FCMR* was down- and *IL24* and *IL10* were up-regulated following $CD4^+$ T cell activation. IL-10 is recognised as an important anti-inflammatory cytokine in health and disease[32] and candidate genes *FCMR* and *IL24* are components of a recently identified proinflammatory module in Th17 cells[33]. At this, and other regions, we found candidate causal variants interacting with multiple genes. Parallel results have demonstrated co-regulation of multiple PCHi-C interacting genes by a single variant[34], suggesting that disease related variants may act on multiple genes simultaneously, consistent with the finding that regulatory elements can interact with multiple promoters[35–37]. This region also shows that clusters of multiple adjacent PIRs can be detected for the same promoter. It remains to be further validated whether all PIRs detected within such clusters correspond to 'causal' interactions or whether some such PIRs are 'bystanders' of strong interaction signals occurring in their vicinity. The use of PCHi-C nonetheless adds considerable resolution compared to simply considering topologically associating domains (TADs), which have a median length of 415 kb in naive $CD4^+$ T cells[34] compared to a median of 37.5 kb total PIR length per gene in non-activated $CD4^+$ T cells (**Supplementary Fig. 11**).

Three neighbouring genes on chromosome 16q24.1, *EMC8*, *COX4I1* and *IRF8*, were prioritised, the last only in activated T cells, for two diseases: RA and SLE (**Supplementary Fig. 12**). Candidate causal variants for SLE and RA fine-mapped to distinct PIRs, yet all these PIRs interact with the same gene promoters, suggesting that interactions, possibly specific to different $CD4^+$ T cell subsets, may allow us to

unite discordant GWAS signals for related diseases[6,38,39]. *EMC8* and *COX4I1* RNA expression was relatively unchanged by activation, whereas *IRF8* expression was upregulated 97-fold, coincident with the induction of 16 intergenic *IRF8* PIRs, four of which overlap autoimmune disease fine-mapped variants. Although the dominant effect of *IRF8* is to control the maturation and function of the mononuclear phagocytic system[40], a T cell-intrinsic function regulating $CD4^+$ Th17 differentiation has been proposed[41]. Our data further link the control of Th17 responses to susceptibility to autoimmune disease including RA and SLE[42].

Other notable examples include *CCR7* and *RARA,* prioritised for T1D through a GWAS signal which maps to chromosome 17q21.2 (**Supplementary Fig. 13a**) and *AHR,* which was prioritised in rheumatoid arthritis (RA), specifically in activated T cells rather than non-activated T cells (**Supplementary Fig. 13b**). Both *CCR7* and *RARA* are strong functional candidates with key roles in trafficking of $CD4^+$ T cells and immune homeostasis[43] and modulating T cell differentiation[44], respectively. *AHR is* a high affinity receptor for toxins in cigarette smoke that has been linked to RA previously through differential expression in synovial fluid of patients, though not through GWAS[45]. Our analysis prioritises *AHR* for RA due to a sub-genome-wide significant signal (rs71540792, $p=2.9x10^{-7}$) and invites attempts to validate the genetic association in additional RA patients.

Interaction-mediated regulation of *IL2RA* expression

We focused on the gene *IL2RA* and attempted to confirm predicted functional effects of fine-mapped variants on *IL2RA* expression. *IL2RA* encodes CD25, a component of the key trimeric cytokine receptor that is essential for high-affinity binding of IL-2, regulatory T cell survival and T effector cell differentiation and function[46]. Multiple variants in and near *IL2RA* have been associated with a number of autoimmune diseases[31,47–49]. We have previously fine-mapped genetic causal variants for T1D and

multiple sclerosis (MS) in the *IL2RA* region[22], identifying five groups of SNPs in intron 1 and upstream of *IL2RA,* each of which is likely to contain a single disease causal variant. Out of the group of eight SNPs previously denoted "A"[22], three (rs12722508, rs7909519 and rs61839660) are located in an area of active chromatin in intron 1, within a PIR that interacts with the *IL2RA* promoter in both activated and non-activated $CD4^+$ T cells (**Fig. 6a**). These three SNPs are also in LD with rs12722495 (r^2>0.86) that has previously been associated with differential surface expression of CD25 on memory T cells[39] and differential responses to IL-2 in activated Tregs and memory T cells[50]. We measured the relative expression of *IL2RA* mRNA in five individuals heterozygous across all group "A" SNPs and homozygous across most other associated SNP groups (**Supplementary Table 9**), in a four-hour activation time-course of $CD4^+$ T cells. Allelic imbalance was observed consistently for two reporter SNPs in intron 1 and in the 3' UTR in non-activated $CD4^+$ T cells in each individual (**Fig. 6b; Supplementary Fig. 14a**), validating a functional effect of the PCHi-C-derived interaction between this PIR and the *IL2RA* promoter in non-activated $CD4^+$ T cells . While the allelic imbalance was maintained in non-activated cells cultured for 2-4 hours, the imbalance was lost in cells activated under our *in vitro* conditions. Since increased CD25 expression with rare alleles at group "A" SNPs has previously been observed on memory $CD4^+$ T cells but not the naive or Treg subsets that are also present in the total $CD4^+$ T cell population[39], we purified memory cells from 8 group "A" heterozygous individuals and confirmed activation-induced loss of allelic imbalance of IL2RA mRNA expression in this more homogeneous population (**Fig. 6c**, **Supplementary Fig. 14b**; Wilcoxon p=0.007). *IL2RA* is one of the most strongly upregulated genes upon $CD4^+$ T cell activation, showing a 65-fold change in expression in our RNA-seq data. Concordant with the genome-wide pattern (**Fig. 3**), the *IL2RA* promoter fragment gains PIRs that accumulate H3K27ac modifications upon activation and these, as well as potentially other changes marked by an increase in H3K27ac modification at rs61839660 and across the group A SNPs in intron 1, could account for the loss of allelic imbalance. These results emphasise the importance of

steady-state CD25 levels on $CD4^+$ T cells for the disease association mediated by the group A SNPs, levels which will make the different subsets of $CD4^+$ T cells more or less sensitive to the differentiation and activation events caused by IL-2 exposure *in vivo*[51].

Discussion

Our results illustrate the dramatic global changes in chromosome conformation in a single cell type in response to a single activation condition, in addition to providing support for the candidacy of certain genes and sequences in GWAS regions as causal for disease. Recent attempts to link GWAS signals to variation in gene expression in primary human cells have sometimes found only limited overlap[52–54]. One explanation may be that these experiments miss effects in specific cell subsets or states, especially given the transcriptional diversity between the many subsets of memory $CD4^+$ T cells.[55] We highlight the complex nature of disease association at the *IL2RA* region where additional PIRs for *IL2RA* gained upon activation overlap other fine-mapped disease-causal variants (**Fig. 6a**), suggesting that other allelically-imbalanced states may exist in activated cells, which may also correspond to altered disease risk. For example, the PIR containing rs61839660, a group A SNP, also contains an activation eQTL for *IL2RA* expression in $CD4^+$ T effectors[56] marked by rs12251836, which is unlinked to the group A variants and was not associated with T1D[56]. Furthermore, rs61839660 itself has recently been reported as a QTL for methylation of the *IL2RA* promoter as well as an eQTL for *IL2RA* expression in whole blood[57]. The differences between CD25 expression in different T cell subsets[58,59], and the rapid activation-induced changes in gene and regulatory expression, chromatin marks and chromosome interactions we observe, imply that a large diversity of cell types and states will need to be assayed to fully understand the identity and effects of autoimmune disease causal variants.

It will be challenging to assay this diversity of cell types and states in large numbers of individuals for traditional eQTL studies, particularly for cell-type or condition-specific eQTLs that have been shown to generally have weaker effects[60,61]. Allele-specific expression (ASE) is a more powerful design to quantify the effects of genetic variation on gene expression with modest sample sizes[62] and the targeted ASE we adopt enables testing individual variants or haplotypes at which donors are selected to be heterozygous,

while controlling for other potentially related variants at which donors are selected to be homozygous. By using statistical fine mapping of GWAS data, integrated with PCHi-C, to highlight both likely disease causal variants and their potential target genes, we enable the design of such targeted ASE analyses. This systematic experimental framework offers an alternative approach to candidate causal gene identification for variants with cell state-specific functional effects, with achievable sample sizes.

Online Methods

$CD4^+$ T cell purification and activation, preparation for genomics assays

$CD4^+$ T cells were isolated from whole blood using RosetteSep (STEMCELL technologies, Canada) according to the manufacturer's instructions. Purified $CD4^+$ T cells (average =96.5% pure, range 92.9 - 98.7%) were washed in X-VIVO 15 supplemented with 1% AB serum (Lonza, Switzerland) and penicillin/streptomycin (Invitrogen, UK) and plated in 96-well CELLSTAR U-bottomed plates (Greiner Bio-One, Austria) at a concentration of 2.5 x 10^5 cells / well. Cells were left untreated or stimulated with Dynabeads human T activator CD3/CD28 beads (Invitrogen, UK) at a ratio of 1 bead : 3 cells for 2-21 hours at 37°C and 5% CO_2. Cells were harvested, centrifuged, supernatant removed and either, (i) resuspended in RLT buffer (RNeasy micro kit, Qiagen, Germany) for RNA-seq (0.75-1 x 10^6 $CD4^+$ T cells / pool and activation state) or microarray (1 x 10^6 $CD4^+$ T cells / donor / timepoint and activation state) (ii) fixed in formaldehyde for capture Hi-C (44-101 x 10^6 $CD4^+$ T cells / pool and activation state) or ChIP-seq (16-26 x 10^6 $CD4^+$ T cells / pool and activation state) as detailed in[34].

ChIP-seq was carried out according to BLUEPRINT protocols[63]. Formaldehyde fixed cells were lysed, sheared and DNA sonicated using a Bioruptor Pico (Diagenode). Sonicated DNA was pre-cleared (Dynabeads Protein A, Thermo Fisher) and ChIP performed using BLUEPRINT validated antibodies and the IP-Star automated platform (Diagenode). Libraries were prepared and indexed using the iDeal library preparation kit (Diagenode) and sequenced (Illumina HiSeq, paired-end).

For PCHi-C[34], DNA was digested overnight with *Hind*III, end labeled with biotin-14-dATP and ligated in preserved nuclei. De-crosslinked DNA was sheared to an average size of 400 bp, end-repaired and adenine-tailed. Following size selection (250-550 bp fragments), biotinylated ligation fragments were

immobilized, ligated to paired-end adaptors and libraries amplified (7-8 PCR amplification rounds). Biotinylated 120-mer RNA baits targeting both ends of *Hind*III restriction fragments that overlap Ensembl-annotated promoters of protein-coding, noncoding, antisense, snRNA, miRNA and snoRNA transcripts were used to capture targets. After enrichment, the library was further amplified (4 PCR cycles) and sequenced on the Illumina HiSeq 2500 platform.

PCHi-C interaction calls

Raw sequencing reads were processed using the HiCUP pipeline[64] and interaction confidence scores were computed using the CHiCAGO pipeline[12] as previously described[34]. We considered the set of interactions with high confidence scores (> 5) in this paper.

Raw PCHi-C read counts from 3 replicates and 2 conditions were transformed into a matrix, and a trimmed mean of M-values normalization was applied to account for library size differences. Subsequently, a voom normalization was applied to log-transformed counts in order to estimate precision weights per contact, and differential interaction estimates were obtained after fitting a linear model on a paired design, using the limma Bioconductor R package[65].

Microarray measurement of gene expression

We recruited 20 healthy volunteers from the Cambridge BioResource. Total $CD4^+$ T cells were isolated from whole blood within 2 hours of venepuncture by RosetteSep (StemCell technologies). To assess the transcriptional variation in response to TCR stimulation, 10^6 $CD4^+$ T cells were cultured in U-bottom 96-well plates in the presence or absence of human T activator CD3/CD28 beads at a ratio of 1 bead : 3 cells. Cells were harvested at 2 , 4, 6 or 21 hours post-stimulation, or after 0, 6 or 21 hours in the absence of stimulation. Three samples from the 6 hour unstimulated time point were omitted from the study due

to insufficient cell numbers, and a further four samples were dropped after quality control, resulting in a total of 133 samples that were included in the final analysis. RNA was isolated using the RNAeasy kit (Qiagen) according to the manufacturer's instructions.

cDNA libraries were synthesized from 200 ng total RNA using a whole-transcript expression kit (Ambion) according to the manufacturer's instructions and hybridized to Human Gene 1.1 ST arrays (Affymetrix). Microarray data were normalized using a variance stabilizing transformation[66] and differential expression was analysed in a paired design using limma[65]. Genes were clustered into modules using WGCNA[67]. Clustering code is available at https://github.com/chr1swallace/cd4-pchic/blob/master/make_modules.R.

ChIP sequencing and regulatory annotation

ChIP sequencing reads for all histone modification assays and control experiments were mapped to the reference genome using BWA-MEM[68], a Burrows-Wheeler genome aligner. Samtools[69] was employed to filter secondary and low-quality alignments (we retained all read pair alignments with PHRED score > 40 that matched all bits in SAM octal flag 3, and did not match any bits in SAM octal flag 3840). The remaining alignments were sorted, indexed and a whole-genome pileup was produced for each histone modification, sample and condition triple.

We used ChromHMM[17], a multivariate hidden Markov model, to perform a whole-genome segmentation of chromatin states for each activation condition (**Supplementary Table 8**). First, we binarized read pileups for each chromatin mark pileup using the corresponding control experiment as a background model. Second, we estimated the parameters of a 15-state hidden Markov model (a larger state model resulted in redundant states) using chromosome 1 data from both conditions. Parameter learning was

re-run five times using different random seeds to assess convergence. Third, a whole-genome segmentation was produced for each condition by running the obtained model on the remaining chromosomes. Each state from the obtained model was manually annotated, and states indicating the presence of promoter or enhancer chromatin tags were selected (E4-E11, **Supplementary Fig. 6**). Overlapping promoter or enhancer regions in non-activated and activated genome segmentations were merged to create a $CD4^+$ T cell regulatory annotation. Thus, we defined 53,534 regulatory regions (**Supplementary Table 8**).

RNA sequencing

Total RNA was isolated using the RNeasy kit (Qiagen) and the concentrations and integrity were quantified using Bioanalyzer (Agilent); all samples reached RINs > 9.8. Two pools of RNA were generated from three and four donors and for each experimental condition. cDNA libraries were prepared from 1ug total RNA using the stranded NEBNext Ultra Directional RNA kit (New England Biolabs), and sequenced on HiSeq (Illumina) at an average coverage of 38 million paired-end reads/sample. RNA sequencing reads were trimmed to remove traces of library adapters by matching each read with a library of contaminants using Cutadapt[70], a semi-global alignment algorithm. Owing to our interest in detecting functional enhancers, which constitute transcription units on their own right, we mapped reads to the human genome using STAR[71], a splicing-aware aligner. This frees us from relying on a transcriptome annotation which would require exact boundaries and strand information for all features of interest, something not available in case of promoters and enhancers.

After alignment, we employed Samtools[69] to discard reads with an unmapped pair, secondary alignments and low-quality alignments. The resulting read dataset, with an average of 33 million paired-end reads/sample, was sorted and indexed. We used FastQC (v0.11.3,

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to ensure all samples had regular GC content (sum of deviations from normal includes less than 15% of reads), base content per position (difference A vs T and G vs C less than 10% at all positions) and kmer counts (no imbalance of kmers with $p < 0.01$) as defined by the tool. We augmented Ensembl 75 gene annotations with regulatory region definitions obtained from our ChIP-seq analysis described above, and defined them as present in both genome strands due to their bidirectional transcription potential. For each RNA-seq sample, we quantified expression of genomic and regulatory features in a two-step strand-aware process using HTSeq[72]. For each gene we counted the number of reads that fell exactly within its exonic regions, and did not map to other genomic elements. For each regulatory feature we counted the number of reads that fell exactly within its defined boundaries, and did not map to other genomic or regulatory elements.

By construction, this quantification scheme counts each read at most once towards at most one feature. Furthermore, strand information is essential to be able to assign reads to features in regions with overlapping annotations. For example, distinguishing intronic eRNAs from pre-mRNA requires reads originating from regulatory activity in the opposite strand from the gene.

Feature counts were transformed into a matrix, and a trimmed mean of M-values normalization was applied to account for library size differences, plus a filter to discard features below an expression threshold of < 0.4 counts per million mapped reads in at least two samples, a rather low cutoff, to allow for regulatory RNAs to enter differential expression calculations. This threshold equates to approximately 15 reads, given our mapped library sizes of ~35 million paired-end reads. A voom normalization was applied to log-transformed counts in order to estimate precision weights per gene, and differential expression estimates were obtained after fitting a linear model on a paired design, using the limma Bioconductor R package[65]. There was a strong correlation (rho=0.81) between microarray and

RNA-seq fold change estimates at 4 hours.

Comparison of regRNAs to FANTOM CAGE data

We compared expressed regRNA regions detected in our non-activated $CD4^+$ T cell samples versus those found using CAGE-seq by the FANTOM5 Consortium. RNA-seq, using a regulatory reference obtained from chromatin states, yields 17,175 features expressed with at least 0.4 counts per million in both non-activated $CD4^+$ T cell samples. Among those, 3,897 correspond to regulatory regions. Unstimulated $CD4^+$ samples from FANTOM5 (http://fantom.gsc.riken.jp/5/datafiles/latest/basic/human.primary_cell.hCAGE/, samples 10853, 11955 and 11998) contain 266,710 loci expressed (with at least one read) in all 3 samples.

We found 13,178 of our 17,175 expressed $CD4^+$ T cell features overlap expressed loci in CAGE data (77%). Conversely, 243,596/266,710 CAGE loci overlap $CD4^+$ T cell features (91%). Similarly 2,888/3,897 expressed regRNAs overlap expressed loci in CAGE data (74%).

Comparison of PCHi-C and ChIA-PET interactions

We downloaded supplementary table 1 from http://www.nature.com/cr/journal/v22/n3/extref/cr201215x1.xlsx[13] and counted the overlaps of PCHi-C interactions from $CD4^+$ T cells and comparitor cells (megakaryoctyes and erythroblasts) in distance bins. R code to replicate the analysis is at https://github.com/chr1swallace/cd4-pchic/blob/master/chepelev.R. Calling interactions requires correction for the expected higher density of random collisions at shorter distances[73] which are explicitly modelled by CHICAGO[12] used in this study but not in the ChIA-PET study[13]. As a result, we expected a higher false positive rate from the ChIA-PET data at shorter distances.

Regression of gene expression against PIR count and eRNA expression

We related measures of gene expression (absolute log2 counts or log2 fold change) to numbers of PIRs or numbers of PIRs overlapping specific features using linear regression. We used logistic regression to relate agreement between fold change direction at PCHi-C linked protein coding genes and eRNAs. We used robust clustered variance estimates to account for the shared baits for some interactions across genes with the same prey. Enrichment of chromatin marks in interacting baits and prey were assessed by logistic regression modelling of a binary outcome variable (fragment overlapped specific chromatin peak) against a fragment width and a categorical explanatory variable (whether the *Hind*III fragment was a bait or prey and the cell state the interaction was identified in), using block bootstrapping of baited fragments (https://github.com/chr1swallace/genomic.autocorr) to account for spatial correlation between neighbouring fragments.

GWAS summary statistics

We used a compendium of 31 GWAS datasets [34] (**Supplementary Table 5**). Briefly we downloaded publicly available GWAS statistics for 31 traits. Where necessary we used the *liftOver* utility to convert these to GRCh37 coordinates. To remove spurious association signals, we removed variants with $P< 5 \times 10^{-8}$ for which there were no variants in LD ($r^2>0.6$ using 1000 genomes EUR cohort as a reference genotype panel) or within 50 kb with $P<10^{-5}$. We masked the MHC region (GRCh37:chr6:25-35Mb) from all downstream analysis due to its extended LD and known strong and complex associations with autoimmune diseases.

Comparison of GWAS data and PIRs requires dense genotyping coverage. For GWAS which did not include summary statistics imputed for non-genotyped SNPs, we used a poor man's imputation (PMI) method [34] to impute. We imputed p values at ungenotyped variants from 1000 Genomes EUR phase 3 by

replacing missing values with those of their nearest proxy variant with r^2>0.6, if one existed. Variants that were included in the study but did not map to the reference genotype set were also discarded.

To calculate posterior probabilities that each SNP is causal under a single causal variant assumption, we divided the genome into linkage disequilibrium blocks of 1cM based on data from the HapMap project (http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/2011-01_phaseII_B37/). For each region excluding the MHC we used code modified from *Giambartolomei et al.*[74] to compute approximate Bayes factors for each variant using the Wakefield approximation[75], and thus posterior probabilities that each variant was causal as previously proposed[76].

Testing of the enrichment of GWAS summary statistics in PIRs using *blockshifter*

We used the *blockshifter* method [34] (https://github.com/ollyburren/CHIGP) to test for a difference between variant posterior probability distributions in *Hind*III fragments with interactions identified in test and control cell types using the mean posterior probability as a measure of central location. *Blockshifter* controls for correlation within the GWAS data due to LD and interaction restriction fragment block structure by employing a rotating label technique similar to that described in GoShifter[77] to generate an empirical distribution of the difference in means under the null hypothesis of equal means in the test and control set. Runs of one or more PIRs (separated by at most one *Hind*III fragment) are combined into 'blocks', that are labeled unmixed (either test or control PIRs) or mixed (block contains both test and control PIRs). Unmixed blocks are permuted in a standard fashion by reassigning either test or control labels randomly, taking into account the number of blocks in the observed sets. Mixed blocks are permuted by conceptually circularising each block and rotating the labels. A key parameter is the gap size - the number of non-interacting *Hind*III fragments allowed within a single block, with larger gaps allowing for more extended correlation.

We used simulation to characterise the type 1 error and power of *blockshifter* under different conditions and to select an optimal gap size. Firstly, from the Javierre *et al.* dataset[34] we selected a test (Activated or Non Activated $CD4^+$ T Cells) and control (Megakaryocyte or Erythroblast) set of PIRs with a CHiCAGO score > 5, as a reference set for *blockshifter* input.

Using the European 1000 genomes reference panel, we simulated GWAS summary statistics , under different scenarios of GWAS/PIR enrichment. We split chromosome 1 into 1cM LD blocks and used reference genotypes to compute a covariance matrix for variants with minor allele frequency above 1%, Σ. GWAS Z scores can be simulated as multivariate normal with mean μ and variance Σ[78]. Each block may contain no causal variants ($GWAS_{null}$, $\mu = 0$) or one ($GWAS_{alt}$). For $GWAS_{alt}$ blocks, we pick a single causal variant, *i*, and calculate the expected non-centrality parameter (NCP) for a 1 degree of freedom chi-square test of association at this variant and its neighbours. This framework is natural because the NCP at any variant *j* can be expressed as the NCP at the causal variant multiplied by the r^2 between variants *i* and *j*[79]. In each case we set the NCP at the causal variant to 80 to ensure that each causal variant was genome-wide significant ($P < 5 \times 10^{-8}$). μ is defined as the square root of this constructed NCP vector.

For all scenarios we randomly chose 50 $GWAS_{alt}$ blocks leaving the remaining 219 $GWAS_{null}$. Enrichment is determined by the preferential location of simulated causal variants within test PIRs. In all scenarios, each causal variant has a 50% chance of lying within a PIR, to mirror a real GWAS in which we expect only a proportion of causal variants to be regulatory in any given cell type. Under the enrichment-null scenario, used to confirm control of type 1 error rate, the remaining variants were assigned to PIRs without regard for whether they were identified in test or control tissues. To examine power, we

considered two different scenarios with PIR-localised causal variants chosen to be located specifically in test PIRs with either 50% probability, scenario power (1), or 100%, scenario power (2). Note that a PIR from the test set may also be in the control set, thus, as with a real GWAS, not all causal variants will be informative for this test of enrichment.

For each scenario we further considered variable levels of genotyping density, corresponding to full genotyping (everything in 1000 Genomes), HapMap imputation (the subset of SNPs also in Stahl et al. REF dataset) or genotyping array (the subset of SNPs also on the Illumina 550k array). Where genotyping density is less than full, we used our proposed poor man's imputation (PMI) strategy to fill in Z scores for missing SNPs.

We ran *blockshifter*, with 1000 null permutations, for each scenario and PMI condition for 4000 simulated GWAS, with a *blockshifter* superblock gap size parameter (the number of contiguous non-PIR *Hind*III fragments allowed within one superblock) of between 1 and 20 and supplying numbers of cases and controls from the RA dataset[48].

For comparison we also investigated the behaviour of a naive test for enrichment for the null scenario. We computed a 2x2 table variants according to test and control PIR overlap, and whether a variant's posterior probability of causality exceeded an arbitrary threshold of 0.01, and Fisher's exact test to test for enrichment.

Enrichment of GWAS summary statistics in CD4$^+$ and activated CD4$^+$ PIRs

We compared the following sets using all GWAS summary statistics, with a superblock gap size of 5 (obtained from simulations above) and 10,000 permutations under the null:-

- Total $CD4^+$ Activated + Total $CD4^+$ NonActivated (test) versus Endothelial precursors + Megakaryocytes (control)
- Total $CD4^+$ Activated (test) versus Total $CD4^-$ NonActivated (control).

Variant posterior probabilities of inclusion, full genotype data (ImmunoChip)

We carried out formal imputation to 1000 Genomes Project EUR data using IMPUTE2 [80] and fine-mapped causal variants in each of the 179 regions where a minimum $p < 0.0001$ was observed using a stochastic search method which allows for multiple causal variants in a region, (https://github.com/chr1swallace/GUESSFM)[22]. The posterior probabilities for models that contained variants which overlapped PIRs for each gene were aggregated to compute PIR-level marginal posterior probabilities of inclusion.

Variant posterior probabilities of inclusion, summary statistics

Where we have only summary statistics of GWAS data already imputed to 1000 Genomes, we divided the genome into linkage disequilibrium blocks of 0.1cM based on data from the HapMap project (http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/2011-01_phaseII_B37/). For each region excluding the MHC we use code modified from *Giambartolomei et al.*[74] to compute approximate Bayes factors for each variant using the Wakefield approximation[75], and thus posterior probabilities that each variant was causal assuming at most one causal variant per region as previously proposed[76].

Computation of gene prioritisation scores

We used the COGS method[34] (https://github.com/ollyburren/CHIGP) to prioritise genes for further analysis. We assign variants to the first of the following three categories it overlaps for each annotated gene, if any

1. coding variant: the variant overlaps the location of a coding variant for the target gene.

2. promoter variant: the variant lies in a region baited for the target gene or adjacent restriction fragment.

3. PIR variant: the variant lies in a region overlapping any PIR interacting with the target gene.

We produced combined gene/category scores by aggregating, within LD blocks, over models with a variant in a given set of PIRs (interacting regions), or over *Hind*III fragments baited for the gene promoter and immediate neighbours (promoter regions), or over coding variants to generate marginal probabilities of inclusion (MPPI) for each hypothesised group. We combine these probabilities across LD blocks, *i*, using standard rules of probability to approximate the posterior probability that at least one LD block contains a causal variant:

$$gene\ score = 1 - \prod_{i \in LD\ blocks} (1 - [score\ for\ i])$$

Thus the score takes a value between 0 and 1, with 1 indicating the strongest support. We report all results with score > 0.01 in **Supplementary Table 7**, but focus in this manuscript on the subset with scores > 0.5.

Because COGS aggregates over multiple signals, a gene may be prioritised because of many weak signals or few strong signals in interacting regions. To predict the expected information for future users of this method, we considered the subset of 76 input regions with genome-wide significant signals ($p<5x10^{-8}$) in ImmunoChip datasets. We prioritised at least one gene with a COGS score > 0.5 in 35 regions, with a median of three genes/region (interquartile range, IQR = 1.5-4). Equivalent analysis of the genome-wide significant GWAS signals prioritised a median of two genes/region (interquartile range = 1-3). This suggests that this algorithm might be expected to prioritise at least one gene in about half the genomewide significant regions input when run on a relevant cell type.

Whilst components 1 and 2 are fixed for a given gene and trait the contribution of variants overlapping PIRs varies depending on the tissue context being examined. We developed a hierarchical heuristic method to ascertain for each target gene which was the mostly likely component and cell state. Firstly for each gene we compute the gene score due to genic effects (components 1 + 2) and interactions (component 3) using all available tissue interactions for that gene. We use the ratio of gene effects score to interactions score in a similar manner to a Bayes factor to decide whether one is more likely. If gene effect is more likely (gene.score ratio >3) we iterate and compare if the gene score due to coding variants (component 1) is more likely than for promoter variants (component 2). Similarly if an interaction is more likely we compare interaction gene scores for activated vs non-activated cells. If at any stage no branch is substantially preferred over its competitor (ratio of gene scores < 3) we return the previous set as most likely, otherwise we continue until a single cell state/set is chosen. In this way we can prioritize genes based on the overall score and label as to a likely mechanism for candidate causal variants.

Allele-specific expression assays

Total $CD4^+$ T cells were isolated from five donors and activated as described above and were harvested after 0, 2 and 4 hours in RLT Plus buffer. Selected donors were heterozygous at all eight group A SNP and, homozygous for group C and F SNPs. Two and three of the donors were homozygous for the group D and E SNP groups, respectively (**Supplementary Table 9**). Memory $CD4^+$ T cells were sorted from cryopreserved PBMC as viable, αβ TCR^+, $CD4^+$, $CD45RA^-$,$CD127^+$, $CD27^+$ cells using a FACSAria III cell sorter (BD Biosciences). Sorted cells were either activated for 4 hours in culture as described above or resuspended directly in RLT plus buffer post-sort. Total RNA was extracted using Qiagen RNeasy Micro plus kit and cDNA was synthesised using Superscript III reverse transcriptase (Thermo Fisher) according to manufacturer's instructions. To perform allele-expression experiments we used a modified version of a previously described method for quantifying methylation in bisulfite sequence data[81]. A

two-stage PCR was used, the first round primers were designed to flank the variant of interest using Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/) and adaptor sequences were added to the primers (Sigma), shown as lowercase letters (rs61839660_ASE_F tgtaaaacgacggccagtGCACACACCTATCCTAGCCT, rs61839660_ASE_R caggaaacagctatgaccCCCACAGAATCACCCACTCT, product size 114bp; rs12244380_ASE_F tgtaaaacgacggccagtTTCGTGGGAGTTGAGAGTGG, rs12244380_ASE_R caggaaacagctatgaccTTAAAAGAGTTCGCTGGGCC, product size 180bp; rs12722495_ASE_F tgtaaaacgacggccagtGTGAGTTTCAATCCTAAGTGCGA, rs12722495_ASE_R caggaaacagctatgaccATTAAGCGGACTCTCTGGGG, product size 97bp). The first round PCR contains 10 µl of Qiagen multiplex PCR mastermix, 0.5 µl of 10 nmol forward primer, 0.5 µl of 10 nmol reverse primer, 4 µl of cDNA and made up to 20 µl with ultra-pure water. The PCR cycling conditions were 95ºC for 15 minutes hot start, followed by 30 cycles of the following steps: 95ºC for 30 seconds, 60ºC for 90 seconds and 72ºC for 60 seconds, finishing with a 72ºC for 10 minutes cycle. The first round PCR product was cleaned using AmpureXP beads (Beckman Coulter) according to manufacturer's instructions. To add Illumina sequence compatible ends to the individual first round PCR amplicons, additional primers were designed to incorporate P1 and A sequences plus sample-specific index sequences in the A primer, through hybridisation to adapter sequence present on the first round gene-specific primers. Index sequences are as published[81]. The second-round PCR contained 8 µl of Qiagen multiplex PCR mastermix, 2.0 µl of ultra-pure water, 0.35 µl of each forward and reverse index primer, 5.3 µl of Ampure XP-cleaned first-round PCR product. The PCR cycling conditions were 95ºC for 15 minutes hotstart, followed by 7 cycles of the following steps: 95ºC for 30 seconds, 56ºC for 90 seconds, 72ºC for 60 seconds, finishing with 72ºC for 10 minutes cycle. All PCR products were pooled at equimolar concentrations based on quantification on the Shimadzu Multina. AmpureXP beads were used to remove unincorporated primers from the product pool. We used the Kapa Bioscience library quantification kit to

accurately quantify the library according to manufacturer's instructions before sequencing on an Illumina MiSeq v3 reagents (2 x 300 bp reads).

Statistical analysis of allele-specific expression data

Sequence data was processed using the Methpup package (https://github.com/ollyburren/Methpup) to extract counts of each allele at rs12722495, and rs12244380 (**Supplementary Table 10**). Individuals were part of a larger cohort genotyped on the ImmunoChip and were phased using snphap (https://github.com/chr1swallace/snphap) to confirm which allele at each SNP was carried on the same chromosome as A2=rs12722495:C or A1=rs12722495:T. Allelic imbalance was quantified as the ratio A2/A1 and was averaged across replicates within individuals using a geometric mean. Allelic ratios in cDNA and gDNA were compared using Wilcoxon rank sum tests. P values are shown in **Fig. 6b** and **Supplementary Fig. 13**. Full details are in https://github.com/chr1swallace/cd4-pchic/blob/master/IL2RA-ASE.R.

References

1. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–1195 (2012).
2. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507,** 371–375 (2014).
3. McGovern, A. *et al.* Capture Hi-C identifies a novel causal gene, IL20RA, in the pan-autoimmune genetic susceptibility region 6q23. *Genome Biol.* **17,** 212 (2016).
4. Xu, Z. *et al.* HiView: an integrative genome browser to leverage Hi-C results for the interpretation of GWAS variants. *BMC Res. Notes* **9,** 159 (2016).
5. Dryden, N. H. *et al.* Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res.* **24,** 1854–1868 (2014).
6. Martin, P. *et al.* Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat. Commun.* **6,** 10069 (2015).
7. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).
8. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2015).
9. Benacerraf, B. & McDevitt, H. O. Histocompatibility-linked immune response genes. *Science* **175,** 273–279 (1972).
10. Gustafsson, M. *et al.* A validated gene regulatory network and GWAS identifies early regulators of T cell-associated diseases. *Sci. Transl. Med.* **7,** 313ra178 (2015).
11. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (2005).
12. Cairns, J. *et al.* CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data.

Genome Biol. **17,** 127 (2016).

13. Chepelev, I., Wei, G., Wangsa, D., Tang, Q. & Zhao, K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res.* **22,** 490–503 (2012).
14. Vahedi, G. *et al.* Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature* **520,** 558–562 (2015).
15. Raghavan, A. *et al.* Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Res.* **30,** 5529–5538 (2002).
16. Lam, M. T. Y., Li, W., Rosenfeld, M. G. & Glass, C. K. Enhancer RNAs and regulated transcriptional programs. *Trends Biochem. Sci.* **39,** 170–182 (2014).
17. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9,** 215–216 (2012).
18. Schmidl, C. *et al.* The enhancer and promoter landscape of human regulatory and conventional T-cell subpopulations. *Blood* **123,** e68–e78 (2014).
19. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).
20. Li, W., Notani, D. & Rosenfeld, M. G. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat. Rev. Genet.* **17,** 207–223 (2016).
21. Levine, M., Cattoglio, C. & Tjian, R. Looping back to leap forward: transcription enters a new era. *Cell* **157,** 13–25 (2014).
22. Wallace, C. *et al.* Dissection of a Complex Disease Susceptibility Region Using a Bayesian Stochastic Search Approach to Fine Mapping. *PLoS Genet.* **11,** e1005272 (2015).
23. Bowes, J. *et al.* Dense genotyping of immune-related susceptibility loci reveals new insights into the genetics of psoriatic arthritis. *Nat. Commun.* **6,** 6046 (2015).

24. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* **43,** 1193–1201 (2011).

25. Smyth, D. J. *et al.* Shared and distinct genetic variants in type 1 diabetes and celiac disease. *N. Engl. J. Med.* **359,** 2767–2777 (2008).

26. Caballero-Franco, C. & Kissler, S. The autoimmunity-associated gene RGS1 affects the frequency of T follicular helper cells. *Genes Immun.* **17,** 228–238 (2016).

27. Ferreira, R. C. *et al.* IL-21 production by CD4+ effector T cells and frequency of circulating follicular helper T cells are increased in type 1 diabetes patients. *Diabetologia* **58,** 781–790 (2015).

28. Nan, L. *et al.* Ubiquitin carboxyl-terminal hydrolase-L5 promotes TGFβ-1 signaling by de-ubiquitinating and stabilizing Smad2/Smad3 in pulmonary fibrosis. *Sci. Rep.* **6,** 33116 (2016).

29. Wicks, S. J. *et al.* The deubiquitinating enzyme UCH37 interacts with Smads and regulates TGF-beta signalling. *Oncogene* **24,** 8080–8084 (2005).

30. Hung, T. *et al.* The Ro60 autoantigen binds endogenous retroelements and regulates inflammatory gene expression. *Science* **350,** 455–459 (2015).

31. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47,** 381–386 (2015).

32. Moore, K. W., de Waal Malefyt, R., Coffman, R. L. & O'Garra, A. Interleukin-10 and the interleukin-10 receptor. *Annu. Rev. Immunol.* **19,** 683–765 (2001).

33. Gaublomme, J. T. *et al.* Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity. *Cell* **163,** 1400–1412 (2015).

34. Javierre, B. M. *et al.* Lineage-specific genome architecture links disease variants to target genes. *submitted* (2016).

35. Hughes, J. R. *et al.* Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* **46,** 205–212 (2014).

36. Martin, P. *et al.* Identifying Causal Genes at the Multiple Sclerosis Associated Region 6q23 Using Capture Hi-C. *PLoS One* **11,** e0166923 (2016).

37. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep.* **17,** 2042–2059 (2016).

38. Meddens, C. A. *et al.* Systematic analysis of chromatin interactions at disease associated loci links novel candidate genes to inflammatory bowel disease. *Genome Biol.* **17,** 247 (2016).

39. Dendrou, C. A. *et al.* Cell-specific protein phenotypes for the autoimmune locus *IL2RA* using a genotype-selectable human bioresource. *Nat. Genet.* **41,** 1011–1015 (2009).

40. Hambleton, S. *et al.* IRF8 mutations and human dendritic-cell immunodeficiency. *N. Engl. J. Med.* **365,** 127–138 (2011).

41. Ouyang, X. *et al.* Transcription factor IRF8 directs a silencing programme for TH17 cell differentiation. *Nat. Commun.* **2,** 314 (2011).

42. Patel, D. D. & Kuchroo, V. K. Th17 Cell Pathway in Human Immunity: Lessons from Genetics and Therapeutic Interventions. *Immunity* **43,** 1040–1051 (2015).

43. Förster, R. *et al.* CCR7 coordinates the primary immune response by establishing functional microenvironments in secondary lymphoid organs. *Cell* **99,** 23–33 (1999).

44. Hall, J. A., Grainger, J. R., Spencer, S. P. & Belkaid, Y. The role of retinoic acid in tolerance and immunity. *Immunity* **35,** 13–22 (2011).

45. Nguyen, N. T. *et al.* Aryl hydrocarbon receptor antagonism and its role in rheumatoid arthritis. *J. Exp. Pharmacol.* **7,** 29–35 (2015).

46. Liao, W., Lin, J.-X. & Leonard, W. J. Interleukin-2 at the crossroads of effector responses, tolerance, and immunotherapy. *Immunity* **38,** 13–25 (2013).

47. International Multiple Sclerosis Genetics Consortium *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476,** 214–219 (2011).

48. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506,** 376–381 (2014).

49. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–124 (2012).

50. Garg, G. *et al.* Type 1 diabetes-associated IL2RA variation lowers IL-2 signaling and contributes to diminished CD4+CD25+ regulatory T cell function. *J. Immunol.* **188,** 4644–4653 (2012).

51. Ballesteros-Tato, A. Beyond regulatory T cells: the potential role for IL-2 to deplete T-follicular helper cells and treat autoimmune diseases. *Immunotherapy* **6,** 1207–1220 (2014).

52. Guo, H. *et al.* Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Hum. Mol. Genet.* **24,** 3305–3313 (2015).

53. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167,** 1398–1414.e24 (2016).

54. Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* (2016). doi:10.1038/ng.3737

55. Duhen, T., Duhen, R., Lanzavecchia, A., Sallusto, F. & Campbell, D. J. Functionally distinct subsets of human FOXP3+ Treg cells that phenotypically mirror effector Th cells. *Blood* **119,** 4430–4440 (2012).

56. Ye, C. J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* **345,** 1254665 (2014).

57. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49,** 131–138 (2017).

58. Hua, J., Davis, S. P., Hill, J. A. & Yamagata, T. Diverse Gene Expression in Human Regulatory T Cell Subsets Uncovers Connection between Regulatory T Cell Genes and Suppressive Function. *J.*

Immunol. **195,** 3642–3653 (2015).

59. Pekalski, M. L. *et al.* Postthymic expansion in human CD4 naive T cells defined by expression of functional high-affinity IL-2 receptors. *J. Immunol.* **190,** 2554–2566 (2013).
60. Wright, F. A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46,** 430–437 (2014).
61. Brown, C. D., Mangravite, L. M. & Engelhardt, B. E. Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs. *PLoS Genet.* **9,** e1003649 (2013).
62. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* **48,** 206–213 (2016).
63. Kostadima. Cell type specific chromatin architecture defines erythropoiesis and megakaryopoiesis. *submitted* (2016).
64. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res.* **4,** 1310 (2015).
65. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).
66. Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1,** S96–104 (2002).
67. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).
68. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
69. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

70. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).

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

72. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).

73. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489,** 109–113 (2012).

74. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10,** e1004383 (2014).

75. Wakefield, J. Bayes factors for genome-wide association studies: comparison with P-values. *Genet. Epidemiol.* **33,** 79–86 (2009).

76. The Wellcome Trust Case Control Consortium *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44,** 1294–1301 (2012).

77. Trynka, G. *et al.* Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *Am. J. Hum. Genet.* **97,** 139–152 (2015).

78. Liu, J. Z. *et al.* A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **87,** 139–145 (2010).

79. Chapman, J. M., Cooper, J. D., Todd, J. A. & Clayton, D. G. Detecting Disease Associations due to Linkage Disequilibrium Using Haplotype Tags: A Class of Tests and the Determinants of Statistical Power. *Hum. Hered.* **56,** 18ㄱㄆㄷ31 (2003).

80. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39,** 906–913 (2007).

81. Rainbow, D. B. *et al.* Epigenetic analysis of regulatory T cells using multiplex bisulfite sequencing. *Eur. J. Immunol.* **45,** 3200–3203 (2015).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements

This work was funded by the JDRF (9-2011-253), the Wellcome Trust (089989, 091157, 107881), the UK Medical Research Council (MR/L007150/1, MC_UP_1302/5), the UK Biotechnology and Biological Sciences Research Council (BB/J004480/1) and the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre. The research leading to these results has received funding from the European Union's 7th Framework Programme (FP7/2007-2013) under grant agreement no.241447 (NAIMIT). The Cambridge Institute for Medical Research (CIMR) is in receipt of a Wellcome Trust Strategic Award (100140).

We thank all study participants and family members.

We thank the Wellcome Trust for funding the AITD UK national collection; all doctors and nurses in Birmingham, Bournemouth, Cambridge, Cardiff, Exeter, Leeds, Newcastle and Sheffield for recruitment of patients and J. Franklyn, S. Pearce (Newcastle) and P. Newby (Birmingham) for preparing and providing DNA samples on Graves' disease patients.

This research utilizes resources provided by the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Human Genome Research Institute (NHGRI), National Institute of Child Health and Human Development (NICHD), and JDRF and supported by U01 DK062418.

We gratefully acknowledge the participation of all Cambridge NIHR BioResource volunteers, and thank the Cambridge BioResource staff for their help with volunteer recruitment. We thank the National Institute for Health Research Cambridge Biomedical Research Centre and NHS Blood and Transplant for funding. Further information can be found at www.cambridgebioresource.org.uk

We thank the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (funded by Wellcome Trust grant reference 090532/Z/09/Z) for the generation of the sequencing data.

We thank Stephen Eyre for helpful comments on the manuscript, and N. Soranzo and the HaemGen consortium for sharing blood trait GWAS summary statistics.

The authors acknowledge the assistance and support of the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre. Helen Stevens, Meeta Maisuria-Armer, Pamela Clarke, Gillian Coleman, Sarah Dawson, Simon Duley, Jennifer Denesha and Trupti Mistry for sample processing. Judy Brown, Lynne Adshead, Amie Ashley, Anna Simpson and Niall Taylor for laboratory administration and procurement support. Vin Everett and Sundeep Nanuwa for logistical and web development.
We thank investigators of published ImmunoChip studies for making available their raw genotyping data (David van Heel, celiac disease; Stephen Eyre, rheumatoid arthritis; Matthew Simmonds, Stephen Gough, Jayne Franklyn, and Oliver Brand, autoimmune thyroid disease).

Author Contributions

Study conceived by: CW, JAT, LSW, PF, and led by CW. Interpreted the data: CW, OSB, AJC, ARG, DR, LSW, JAT. Sequence data analysis: ARG. HiCUP analysis: SW. ASE experiments and analysis:

DR. Microarray experiments and analysis: XCD,RCF, RC, CW. Statistical analysis: OSB, JC, NJC, CW, ARG. Laboratory experiments: AJC, BJ, DR, JJL, FB, SPR, KD. Wrote the paper: CW, OSB, AJC, ARG and contributed to writing: JAT, LSW, MS. Revised the paper: all authors. Genetic association data processing: CW, OSB, ES. Supervised capture Hi-C experiments: MS and PF. Supervised cell experiments: AJC, MF, WO, PF, JAT and LW.

Data availability

The following datasets were generated:

- PCHi-C data are available as raw sequencing reads (submission to EGA in progress) or CHICAGO-called interactions (**Supplementary Table 4**) and are available for interactive exploration via http://www.chicp.org
- RNA-seq and ChIPseq data are available as raw sequencing reads (submission to EGA in progress).
- Microarray data are available at ArrayExpress, https://www.ebi.ac.uk/arrayexpress, accession number E-MTAB-4832
- Processed datasets are available as Supplementary Tables
- Code used to analyse the data are available from https://github.com/chr1swallace/cd4-pchic except where other URLs are indicated in Methods

Reprints and permissions information is available at www.nature.com/reprints.

Competing financial interests.

The authors declare no competing financial interests.

Correspondence and requests for materials should be addressed to cew54@cam.ac.uk.

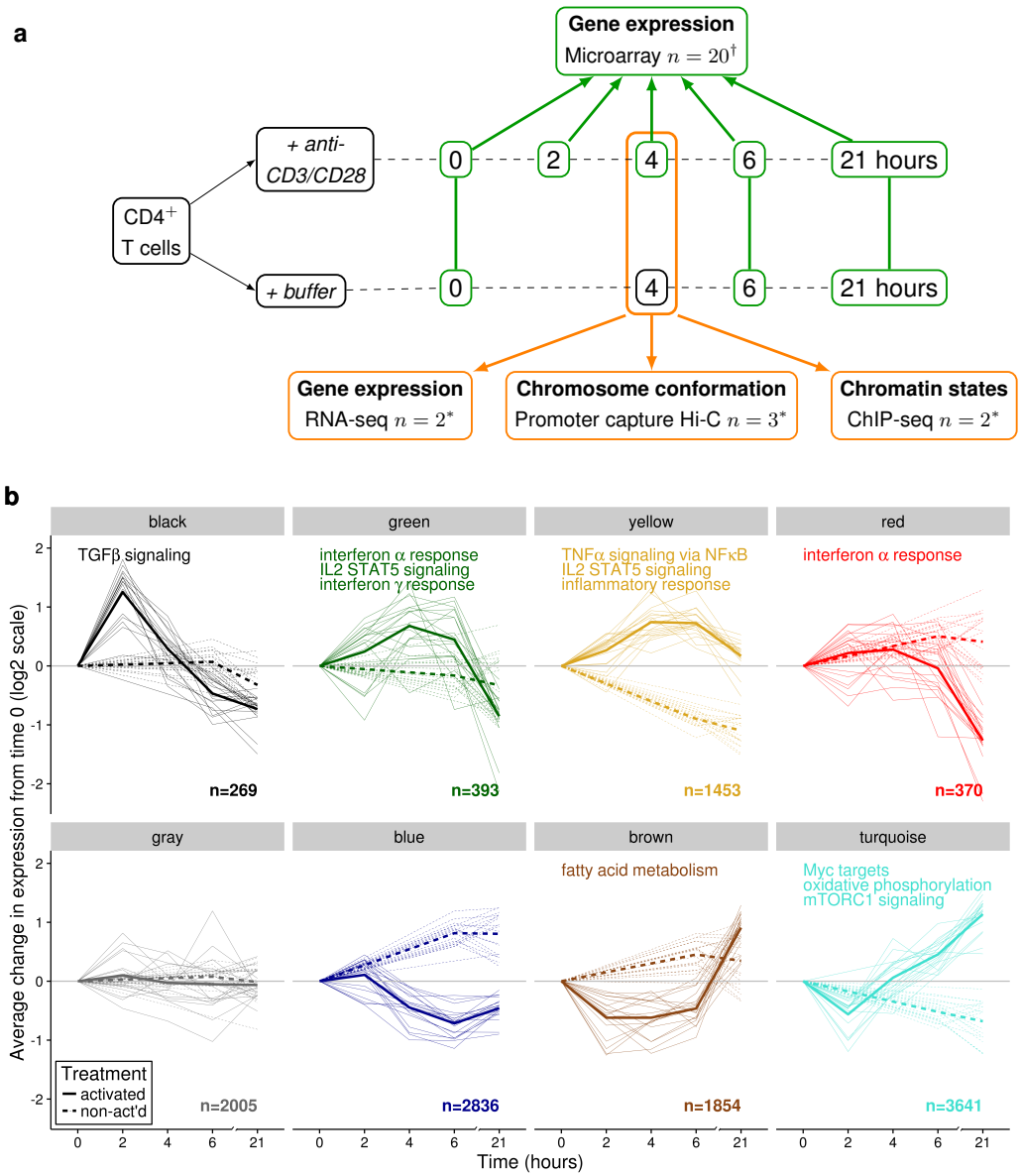


Fig. 1 a: Summary of genomic profiling of CD4^{+} T cells during activation with anti-CD3/CD28 beads. We examined gene expression using microarray in activated and non-activated CD4^{+} T cells across 21 hours, and assayed cells in more detail at the four hour time point using ChIP-seq, RNA-seq and PCHi-C. n gives the number of individuals or pools* assayed. **b**: Eight modules of co-regulated genes were identified, and eigengenes are plotted for each individual (solid lines=activated, dashed lines=non-activated), with heavy lines showing the average eigengene across individuals. We characterized these modules by gene set enrichment analysis within the MSigDB HALLMARK gene sets, and where significant gene sets were found, up to three are shown per module. n is the number of genes in each module.

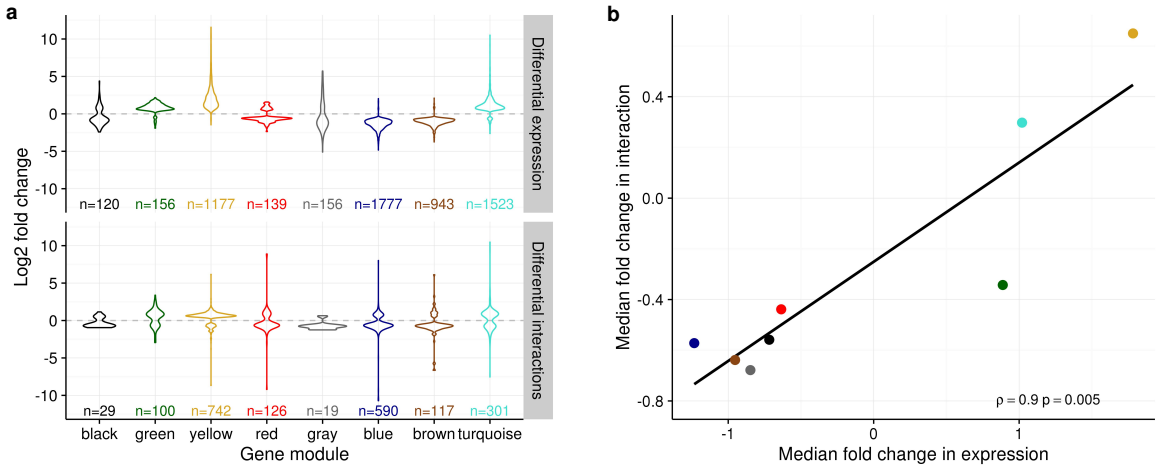


Fig. 2 Change in PCHi-C interactions correlate with change in gene expression a: Distribution of significant (FDR<0.01) fold changes induced by activation of $CD4^+$ T cells in (top) gene expression, and (bottom) differential PCHi-C interactions for differentially expressed genes in by module. **b**: Median significant expression and interaction fold changes by module are correlated (Spearman rank correlation).

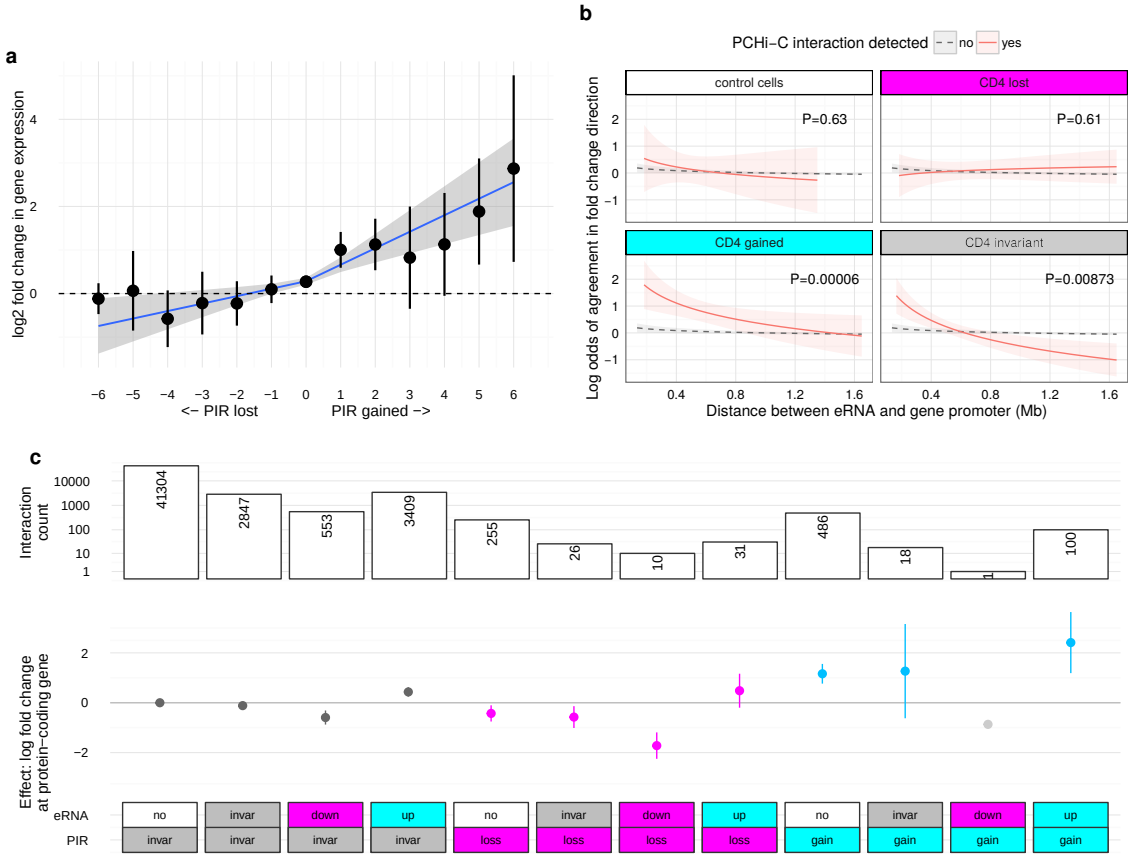


Fig. 3 PCHi-C interactions and enhancer activity predict change in gene expression. **a**: Change in gene expression at protein coding genes ($\log_2$ fold change, y axis) correlates with the number of PIRs gained or lost upon activation (x axis). **b**: Fold change at transcribed sequence within intergenic regulatory regions (eRNAs) was more likely to agree with the direction of protein coding gene fold change when the two are linked by PCHi-C (red) in activated $CD4^+$ T cells compared to pairs of eRNA and protein coding genes formed without regard to PCHi-C derived interactions (background, grey, $p < 10^{-4}$). Interactions were categorised as control (present only in megakaryocytes and erythroblasts, our control cells), invariant (invar; present in non-activated and activated $CD4^+$ T cells), loss (present in non-activated but not activated $CD4^+$ T cells, and significantly down-regulated at FDR<0.01) or gain (present in activated but not non-activated $CD4^+$ T cells, and significantly up-regulated at FDR<0.01). **c**: gain or loss of PIRs upon activation predicts change in gene expression, with the effect more pronounced if accompanied by up- or down-regulation at an eRNA. Points show estimated effect on gene expression of each interaction and lines the 95% confidence interval. PIRs categorised as in **b**. eRNAs categorised as no (undetected), invariant (invar, detected in non-activated and activated $CD4^+$ T cells, differential expression FDR≥0.01), up (up-regulated; FDR<0.01) or down (down-regulated; FDR<0.01). Bar plot (top) shows the number of interactions underlying each estimate. Note that eRNA=down, PIR=gain (light gray) has only one observation so no confidence interval can be formed and is shown for completeness only.

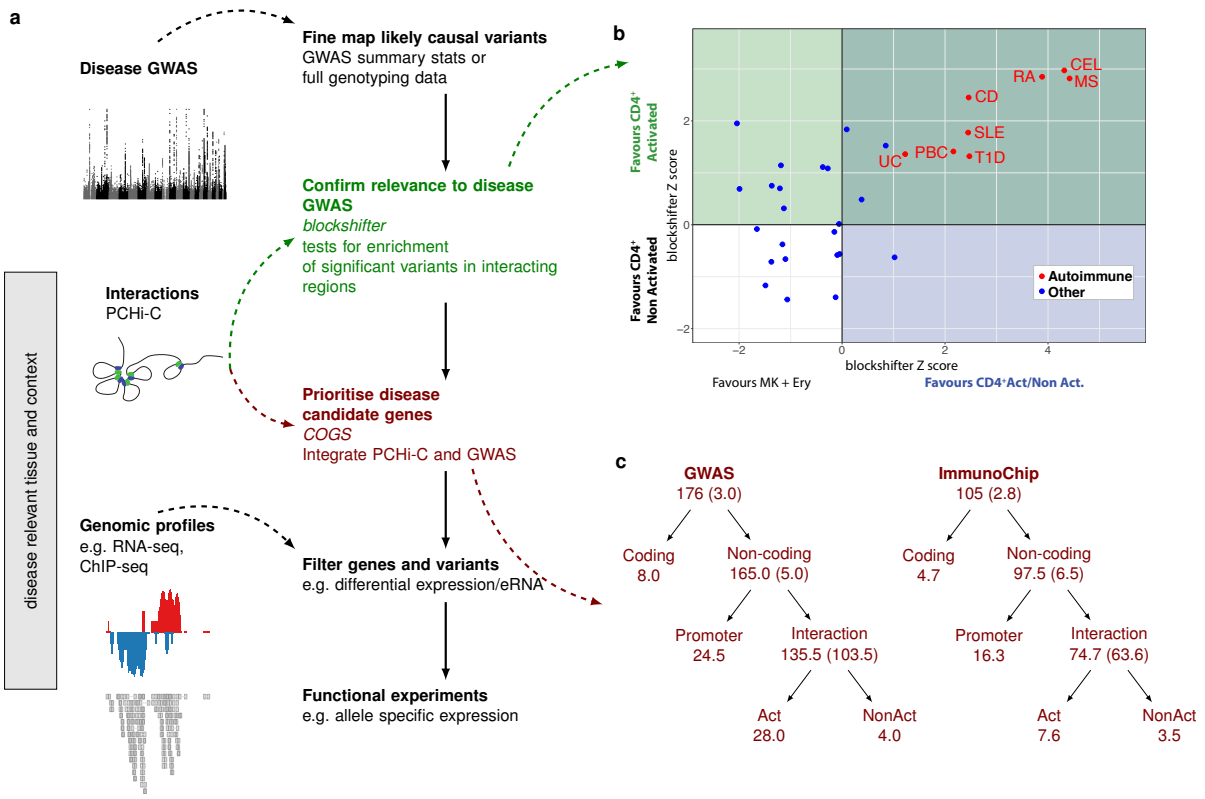


Fig. 4. a: An experimental framework for identifying disease causal genes. Before prioritising genes, enrichment of GWAS signals in PCHi-C interacting regions should be tested to confirm the tissue and context are relevant to disease. Then, probabilistic fine mapping of causal variants from the GWAS data can be integrated with the interaction data to prioritise candidate disease causal genes, a list which can be iteratively filtered using genomic datasets to focus on (differentially) expressed genes and variants which overlap regions of open or active chromatin. **b**: Autoimmune disease GWAS signals are enriched in PIRs in CD4$^+$ T cells generally compared to control cells (blockshifter Z score, x axis) and in PIRs in activated compared to non-activated CD4$^+$ T cells (blockshifter Z score, y axis). Text labels correspond to datasets described in Supplementary Data 5. **c**: Genes were prioritised with a COGS score>0.5 across five autoimmune diseases using genome-wide (GWAS) or targeted genotyping array (ImmunoChip) data. The numbers at each node give the number of genes prioritised at that level. Where there is evidence to split into one of two non-overlapping hypotheses ($\log_{10}$ ratio of gene scores>3), the genes cascade down the tree. Act and NonAct correspond to gene scores derived using PCHi-C data only in activated or non-activated cells, respectively. Where the evidence does not confidently predict which of the two possibilities is more likely, genes are stuck at the parent node (number given in brackets). When the same gene was prioritised for multiple diseases, we assigned fractional counts to each node, defined as the proportion of the n diseases for which the gene was prioritised at that node. Because of duplicate results between GWAS and ImmunoChip datasets, the total number of prioritised genes is 252 (see **Table 1**).

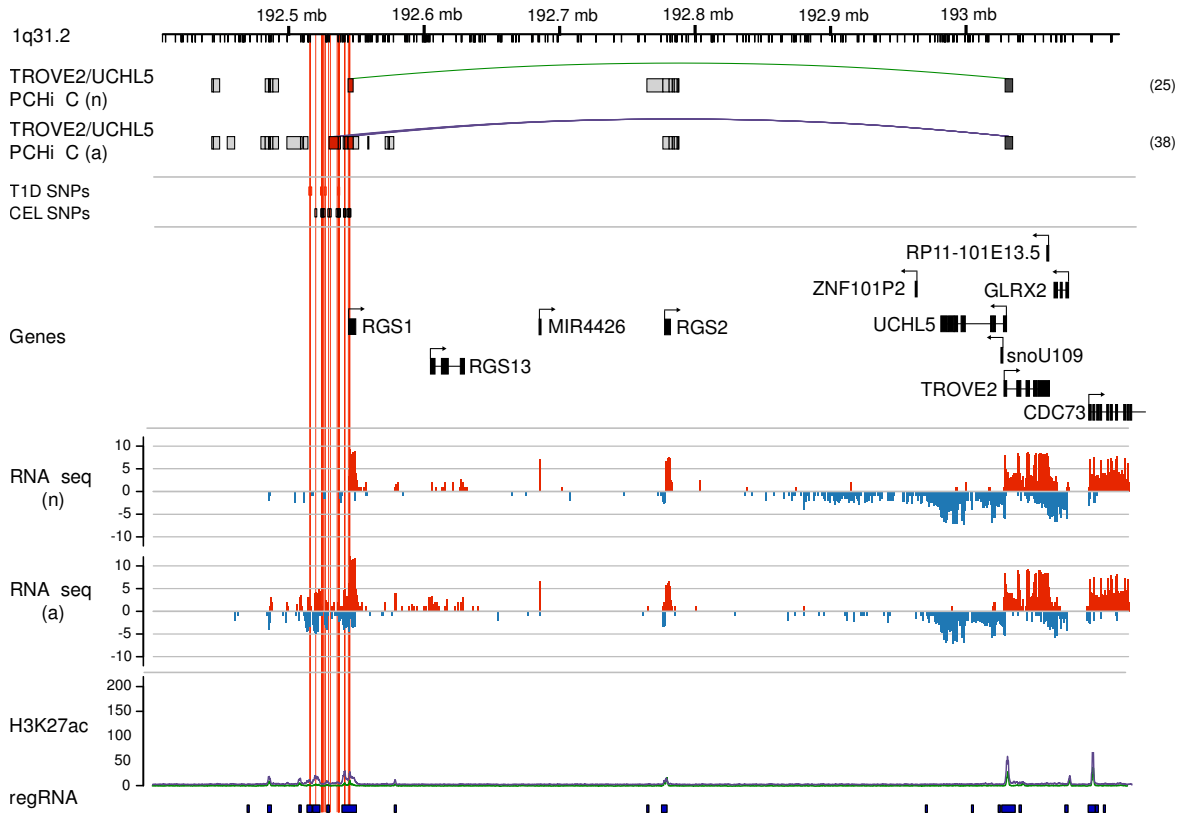


Fig. 5. *TROVE2* and *UCLH5* on chromosome 1 are prioritised for T1D and CEL. The ruler shows chromosome location, with *Hind*III sites marked by ticks. The top tracks show PIRs for prioritised genes in non-activated (n) and activated (a) CD4$^+$ T cells. Green and purple lines are used to highlight those PIRs containing credible SNPs from our fine mapping. The total number of interacting fragments per PCHi-C bait is indicated in parentheses for each gene in each activation state. Dark grey boxes indicate promoter fragments; light grey boxes, PIRs containing no disease associated variants; and red boxes, PIRs overlapping fine mapped disease associated variants. The position of fine mapped variants area indicated by red boxes and vertical red lines. Gene positions and orientation (ensembl v75) are shown above log2 read counts for RNA-seq forward (red) and reverse (blue) strand. H3K27ac background-adjusted read count is shown in non-activated (green line) and activated (purple line) and boxes on the regRNA track show regions considered through ChromHMM to have regulatory marks.

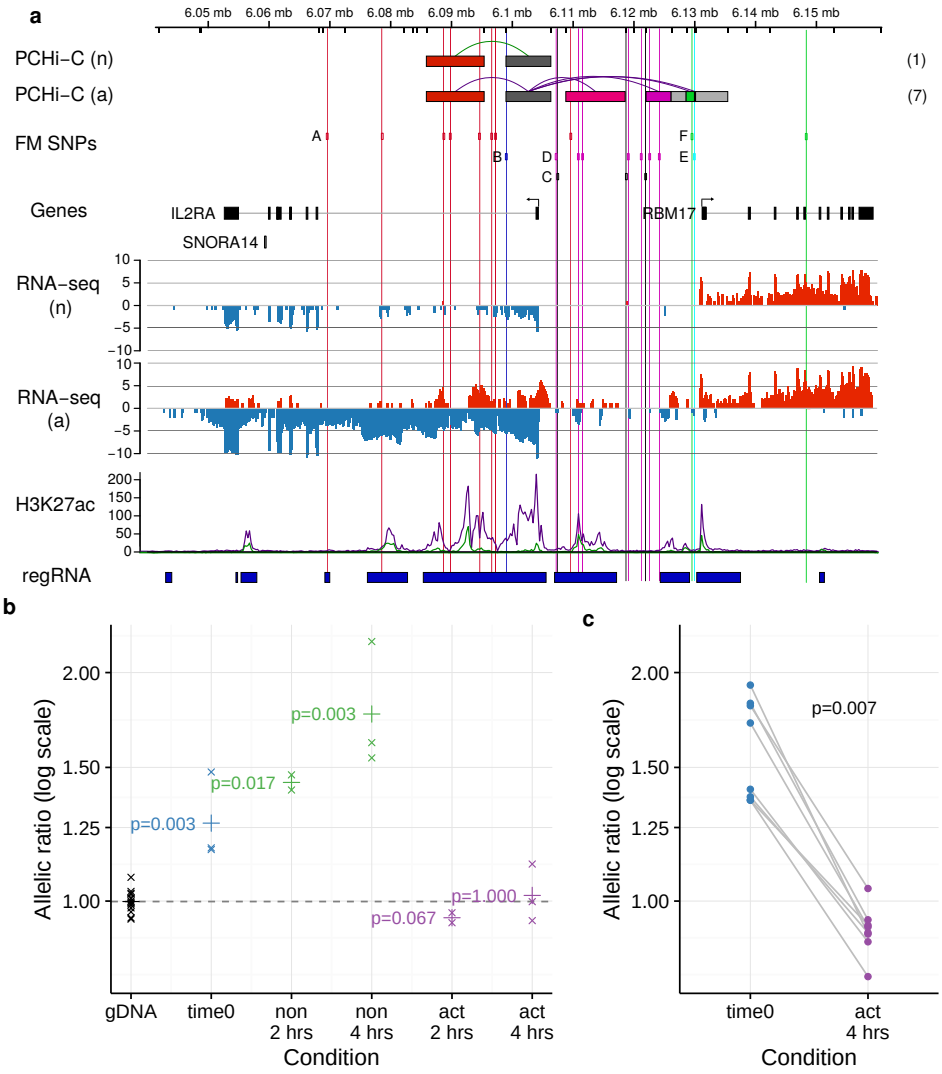
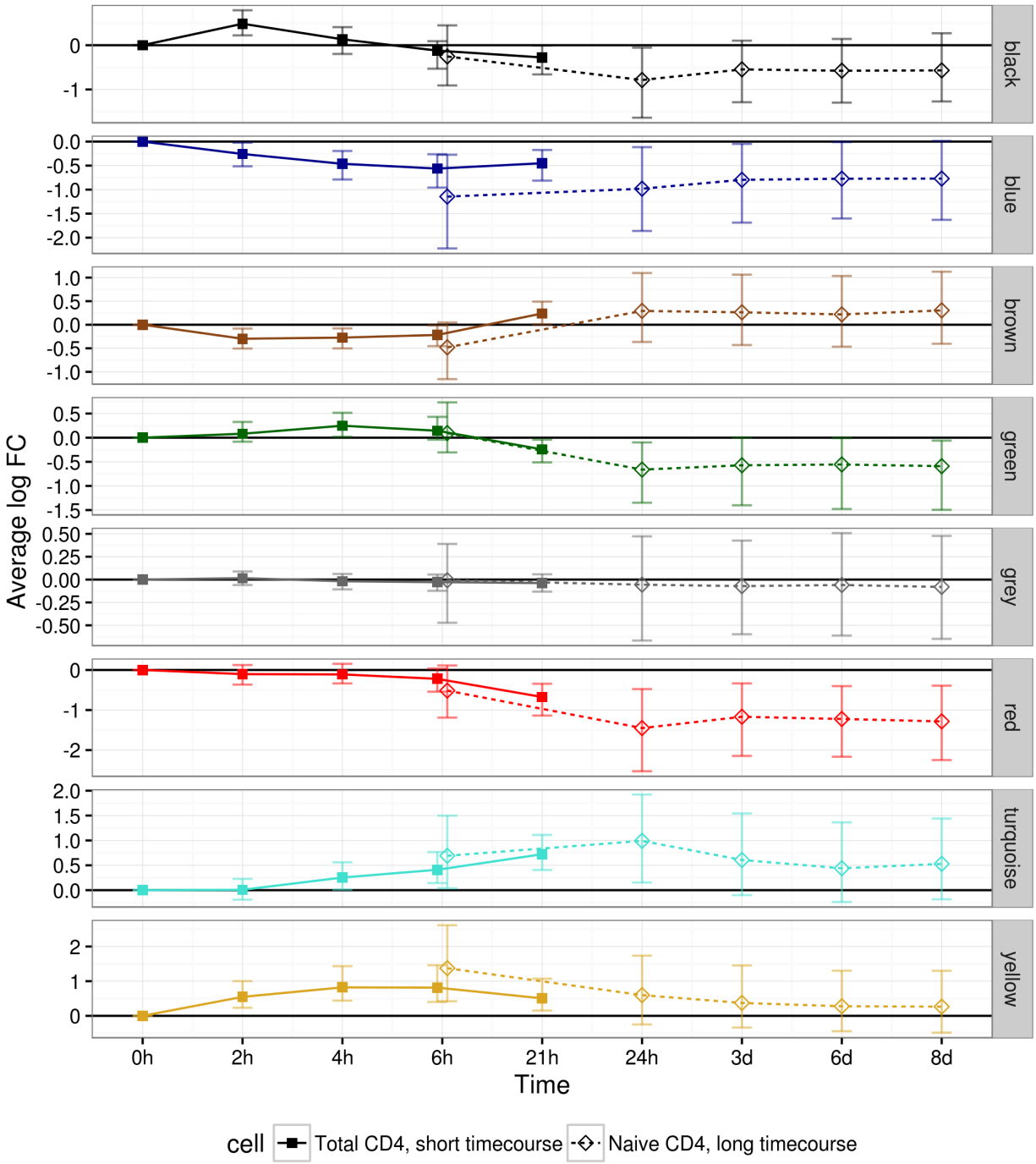

a
6.05 mb
6.06 mb
6.07 mb
6.08 mb
6.09 mb
6.1 mb
6.11 mb
6.12 mb
6.13 mb
6.14 mb
6.15 mb
PCHi-C (n)
(1)
PCHi-C (a)
(7)
FM SNPs
A
B
C
D
E
F
Genes
IL2RA
SNORA14
RBM17
RNA-seq (n)
RNA-seq (a)
H3K27ac
regRNA
b
Allelic ratio (log scale)
p=0.003
p=0.017
p=0.003
p=1.000
p=0.067
gDNA
time0
non 2 hrs
non 4 hrs
act 2 hrs
act 4 hrs
Condition
c
p=0.007
time0
act 4 hrs
Condition

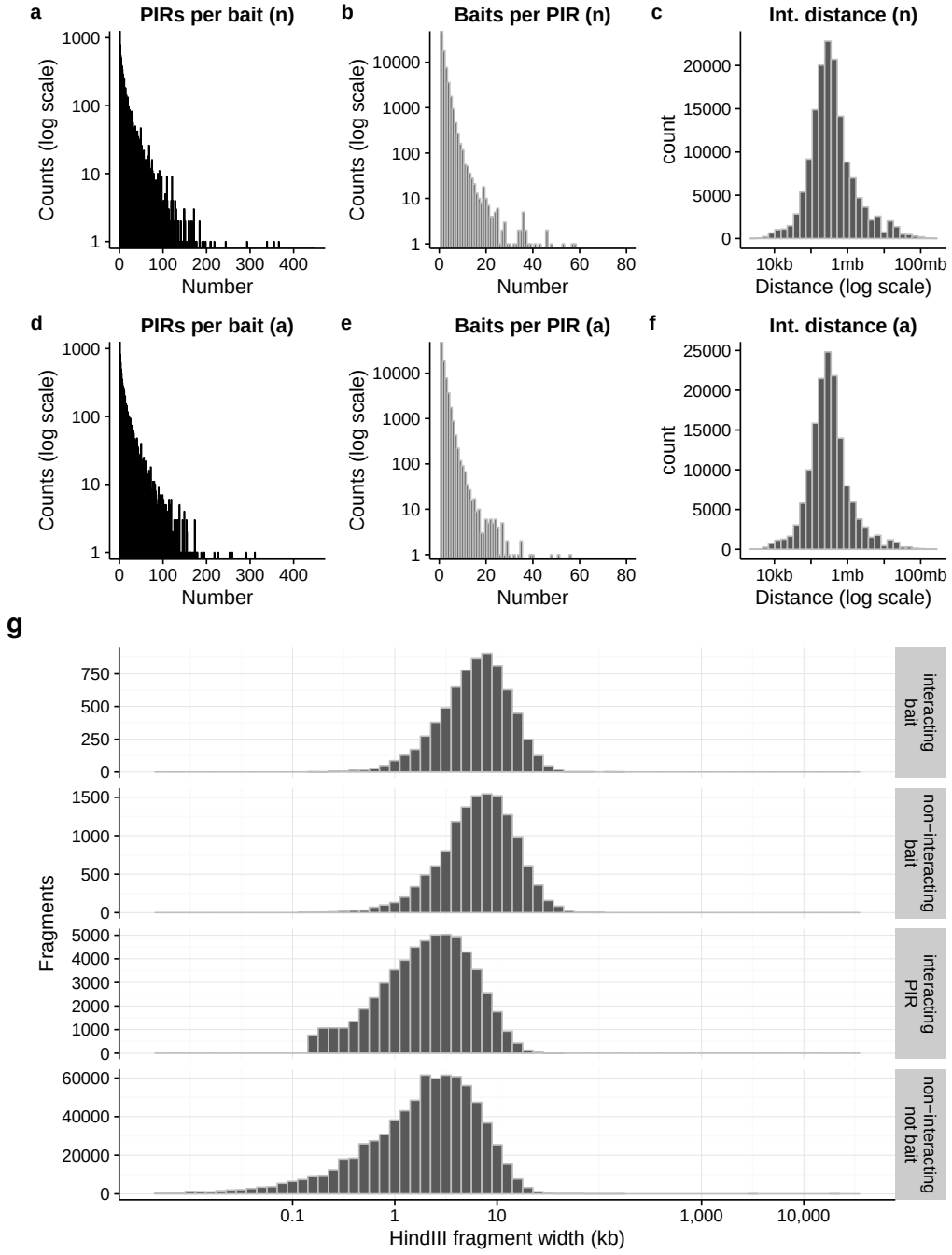
Fig. 6. PCHi-C interactions link the *IL2RA* promoter to autoimmune disease associated genetic variation, which leads to expression differences in *IL2RA* mRNA. a: The ruler shows chromosome location, with *Hind*III sites marked by ticks. The top tracks show PIRs for prioritised genes in non-activated (n) and activated (a) $CD4^+$ T cells. Green and purple lines are used to highlight those PIRs containing credible SNPs for the autoimmune diseases T1D and MS fine mapped on chromosome 10p15[22]. Six groups of SNPs (A-F) highlighted in Wallace et al.[22] are shown, although note that group B was found unlikely to be causal. The total number of interacting fragments per PCHi-C bait is indicated in parentheses for each gene in each activation state. Dark grey boxes indicate promoter fragments; light grey boxes, PIRs containing no disease associated variants; and coloured boxes, PIRs overlapping fine mapped disease associated variants. PCHi-C interactions link a region overlapping group A in non-activated and activated $CD4^+$ T cells to the *IL2RA* promoter (dark grey box) and regions overlapping groups D and F in activated $CD4^+$ T cells only. RNA-seq reads ($\log_2$ scale, red=forward strand, blue=reverse strand) highlight the upregulation of *IL2RA* expression upon activation and concomitant increases in H3K27ac (non-activated, n, green line; activated, a, purple line) in the regions linked to the *IL2RA* promoter. Red vertical lines mark the positions of the group A SNPs. Numbers in parentheses show the total number of *IL2RA* PIRs detected in each state. Here we show those PIRs proximal to the *IL2RA* promoter. Comprehensive interaction data can be viewed at `http://www.chicp.org`. **b**: Allelic imbalance in mRNA expression in total $CD4^+$ T cells from individuals heterozygous for group A SNPs using rs12722495 as a reporter SNP in non-activated (non) and activated (act) $CD4^+$ T cells cultured for 2 or 4 hours, compared to genomic DNA (gDNA, expected ratio=1). Allelic ratio is defined as the ratio of counts of T to C alleles. '×'=geometric mean of the allelic ratio over 2-3 replicates within each of 4-5 individuals, and p values from a Wilcoxon rank sum test comparing cDNA to gDNA are shown. '+' shows the geometric mean allelic ratio over all individuals. **c**: Allelic imbalance in mRNA expression in memory $CD4^+$ T cells differs between *ex vivo* (time 0) and four hour activated samples from eight individuals heterozygous for group A SNPs using rs12722495 as a reporter SNP. p value from a paired Wilcoxon signed rank test is shown.

Group	Description	Number of genes
1	Total	252
2	$\cdots$ Expressed	186
3	$\cdots\cdots$ Proximal GWAS significant SNP ($p < 5 \times 10^{-8}$)	120
4	$\cdots\cdots\cdots$ Prioritised gene differentially expressed upon activation	83
5	$\cdots\cdots\cdots$ Prioritisation relates to activation sensitive interactions	49
6	$\cdots\cdots\cdots$ GWAS signal overlaps expressed eRNA in at least one state	29

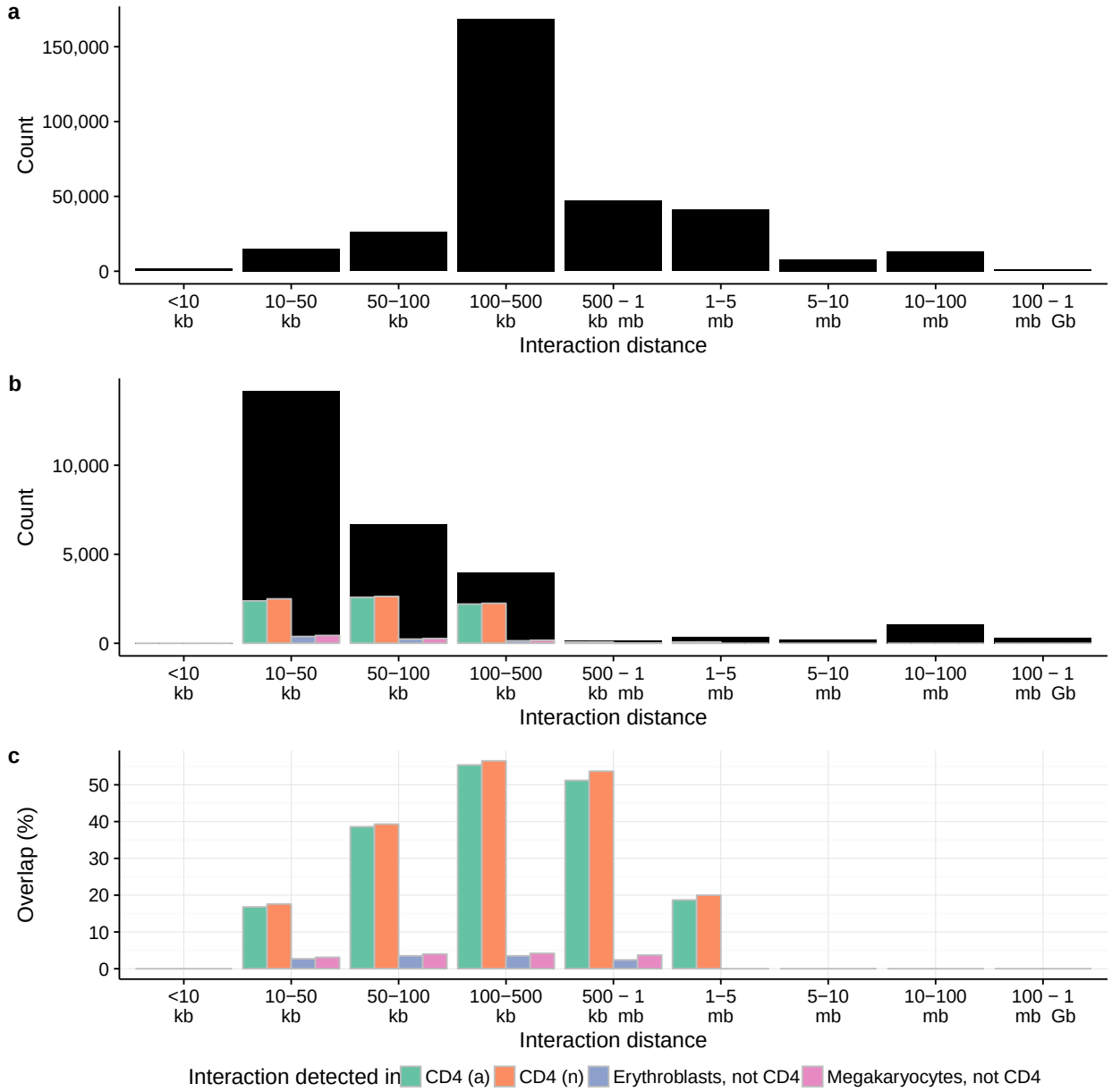
Table 1. Number of genes prioritised for autoimmune disease susceptibility under successive filters. Note that group 2 is a subset of group 1, group 3 is a subset of group 2, and groups 4, 5 and 6 are all subsets of group 3 but not necessarily of each other.



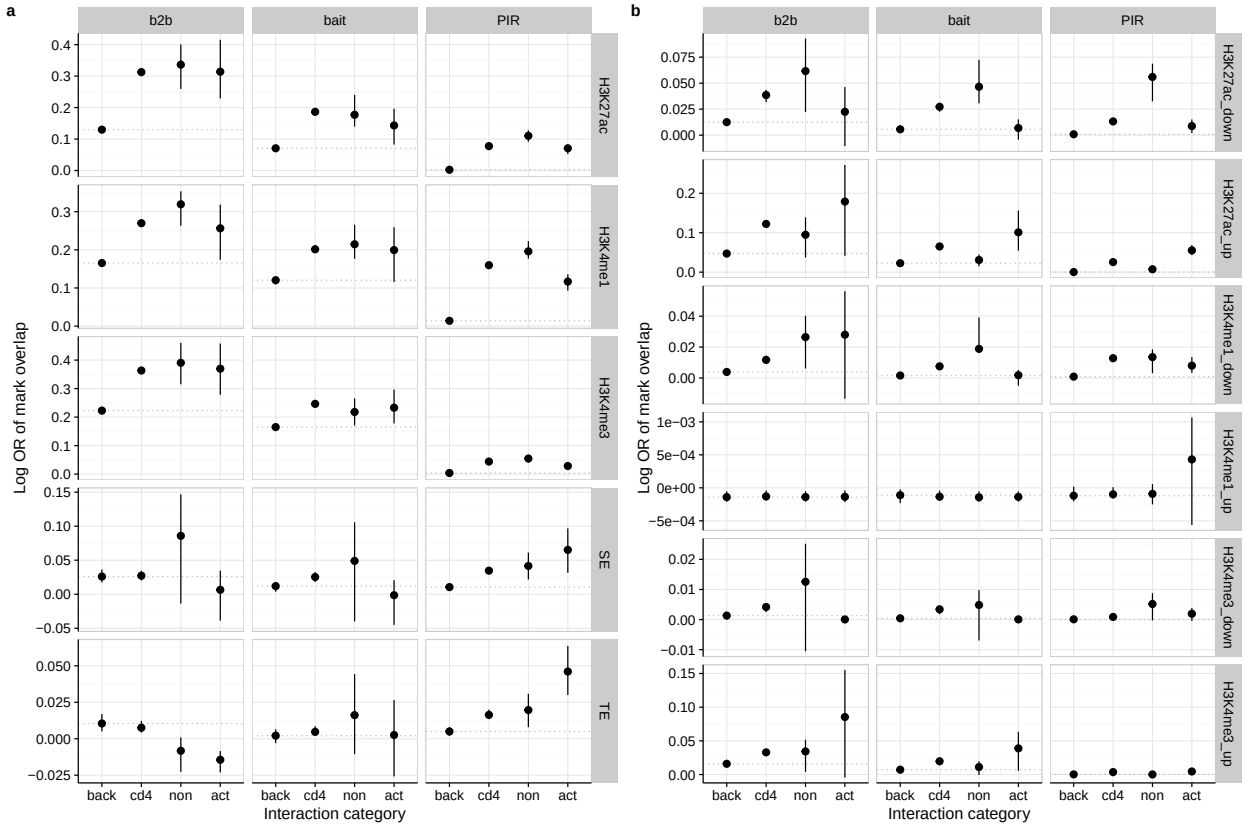
Supplementary Fig. 1: **Comparison of longer and shorter CD4$^+$ T cell activation timecourses**. Microarray timecourse summary from this experiment (solid points) overplotted with a longer timecourse from GSE60680 (open points, Gustafsson et al, 2015). Points show the median log fold change amongst genes assigned to each module at each timepoint, with the interquartile range displayed as vertical ranges around each point. The results at 6 hours are slightly horizontally offset to allow the results from the two experiments to be visually distinguished. Note the non-linear mapping of time to the x axis, which contains a mixture of hours (h) and days (d), to allow visualization of the early timepoints in particular.



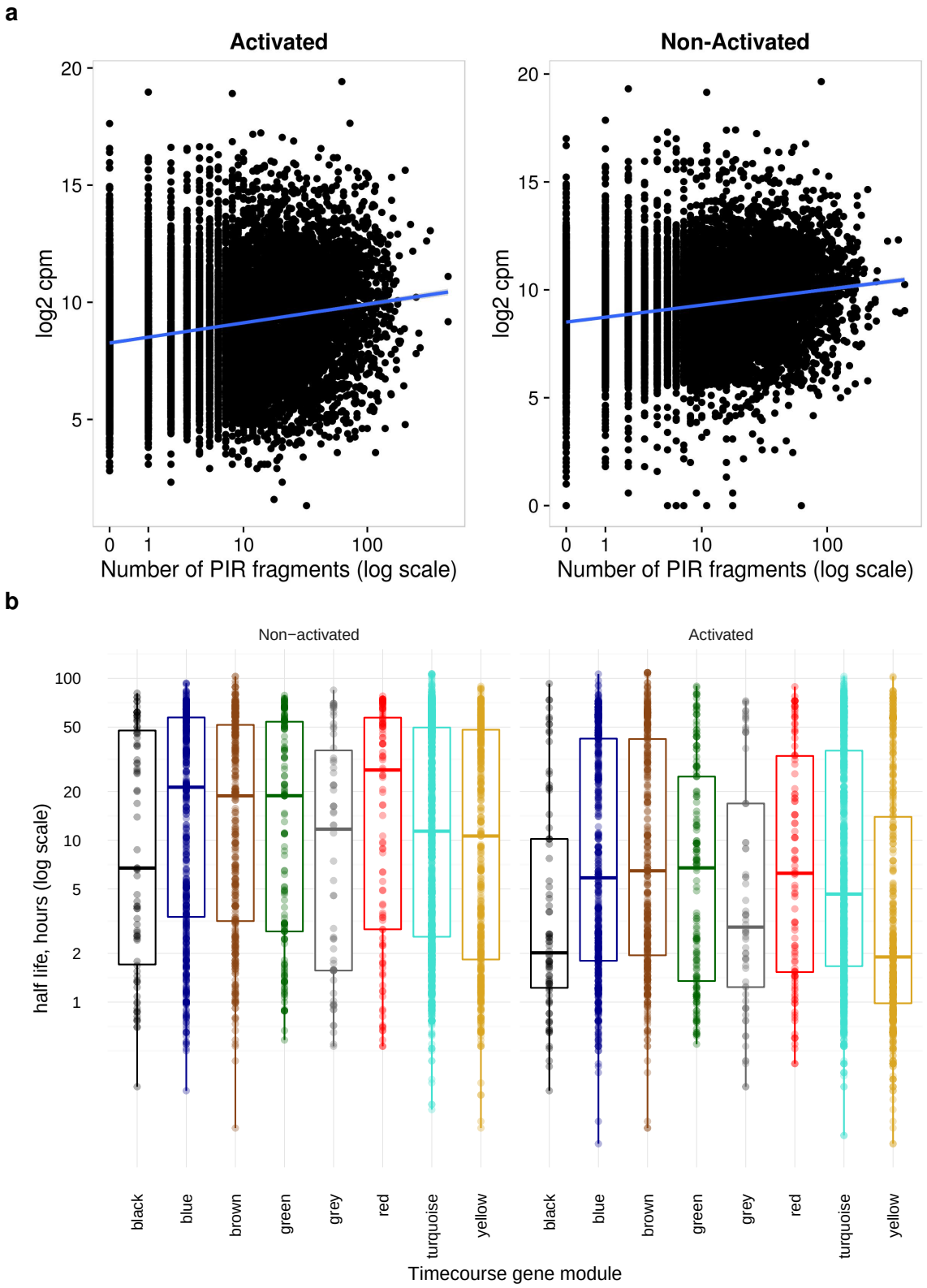
Supplementary Fig. 2: **Summary distributions of interacting fragments.** Distributions of **a**, **d** number of interacting promoter bait fragments per PIR; **b**, **e** PIRs per promoter fragment; and **c**, **f** distance between midpoints of promoter and PIR *Hind*III fragments in activated (**a-c**) and non-activated (**d-f**) $CD4^+$ T cells. **f** Width profile of *Hind*III fragments according to whether they were baited promoter fragments or not, and interacting fragments or not. **g** *Hind*III fragment length in the four categories of interacting and non-interacting baited fragments and PIRs.



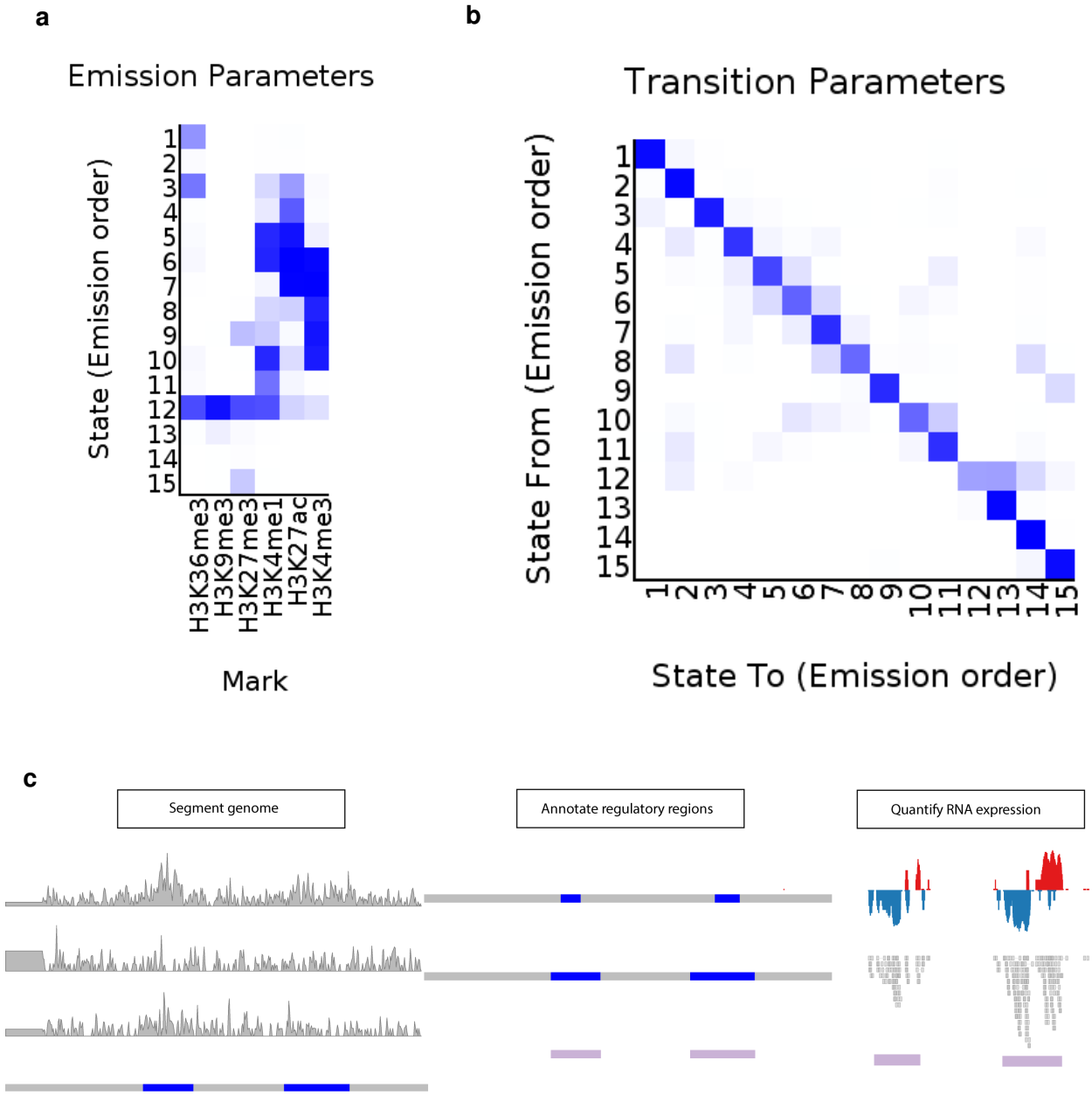
Supplementary Fig. 3: **Validation of PCHi-C by ChIA-PET**. Distance profiles of PCHi-C and ChIA-PET derived promoter-enhancer interactions in **a** PCHi-C, non-activated CD4$^+$ T cells and **b** ChIA-PET (Chepelev et al), black bars. Coloured bars show the count (**b**) or percentage (**c**) of ChIA-PET interactions recovered in the PCHi-C experiment in non-activated and activated CD4$^+$ T cells (CD4 (a) and CD4 (n), respectively) and, for comparison, two non-lymphocyte cells, ethryroblasts and megakaryocytes processed in parallel after exclusion of interactions found in either CD4$^+$ T cell. Calling interactions requires correction for the expected higher density of random collisions at shorter distances[57] which are explicitly modelled by CHICAGO[9] used in this study but not in the ChIA-PET study[12]. As a result, we expected a higher false positive rate from the ChIA-PET data at shorter distances. Indeed, while we replicated only 17% of interactions in the 10-50kb range, we replicated over 50% of the longer range interactions (>100 kb).



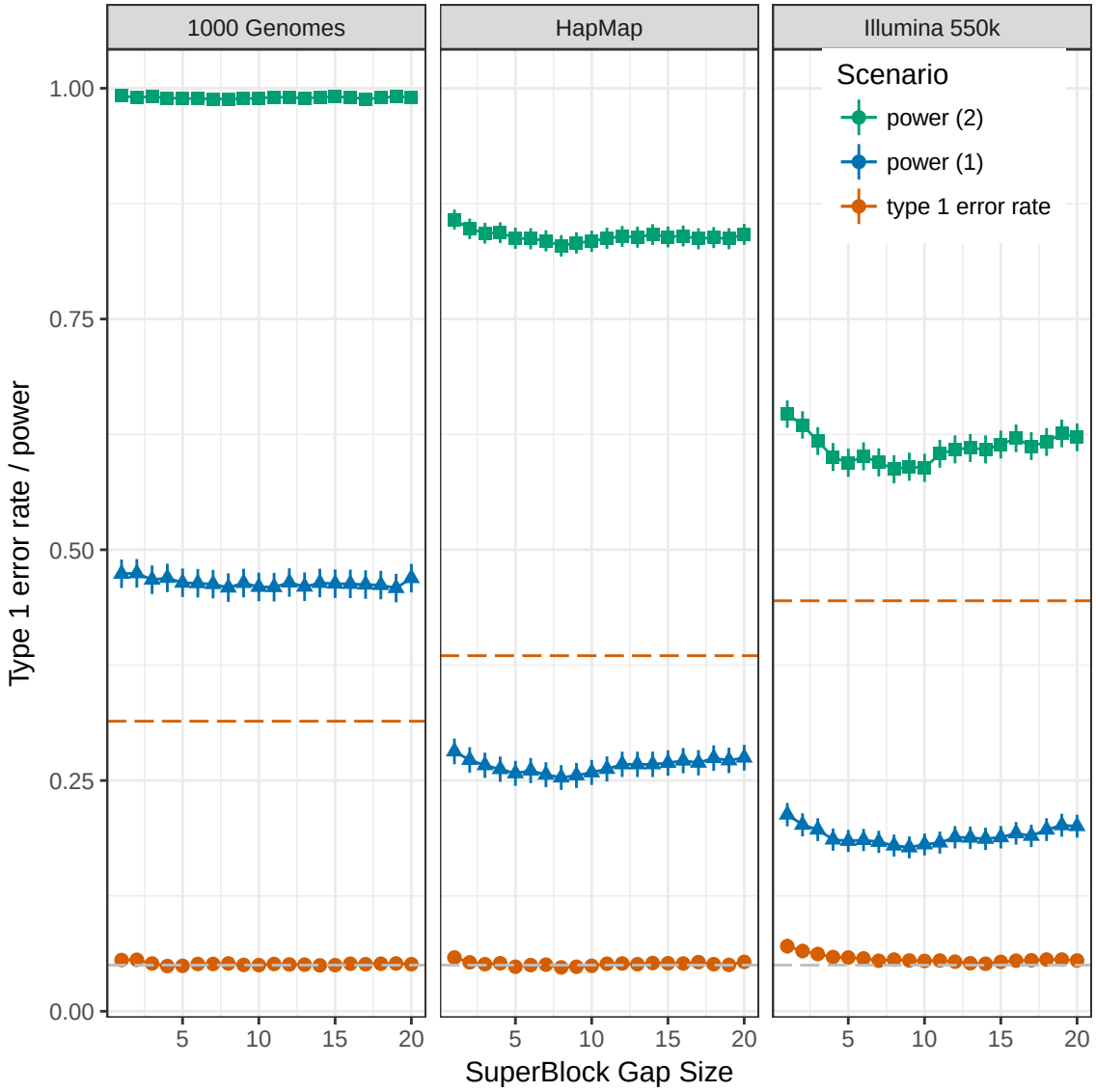
Supplementary Fig. 4: **Chromatin state profiles of interacting fragments.** Log odds ratio that bait, bait-to-bait (b2b) and PIR regions detected in background cells (back; megakayrocytes and erythroblasts), activated and non-activated $CD4^+$ T cells (cd4), specifically non-activated or activated $CD4^+$ T cells (non or act, respectively) overlap (**a**) given ChIP-seq peaks or typical (TE) or super (SE) enhancers in resting T cells as previously defined[12] and (**b**) differential (FDR$<$0.1) ChIP-seq peaks compared to non-interacting regions. Regions considered specific to activated or non-activated cells had a CHICAGO score $>$ 5 only in that cell type and were considered differential interactions in a comparative analysis of mapped sequence counts at FDR$<$0.1.



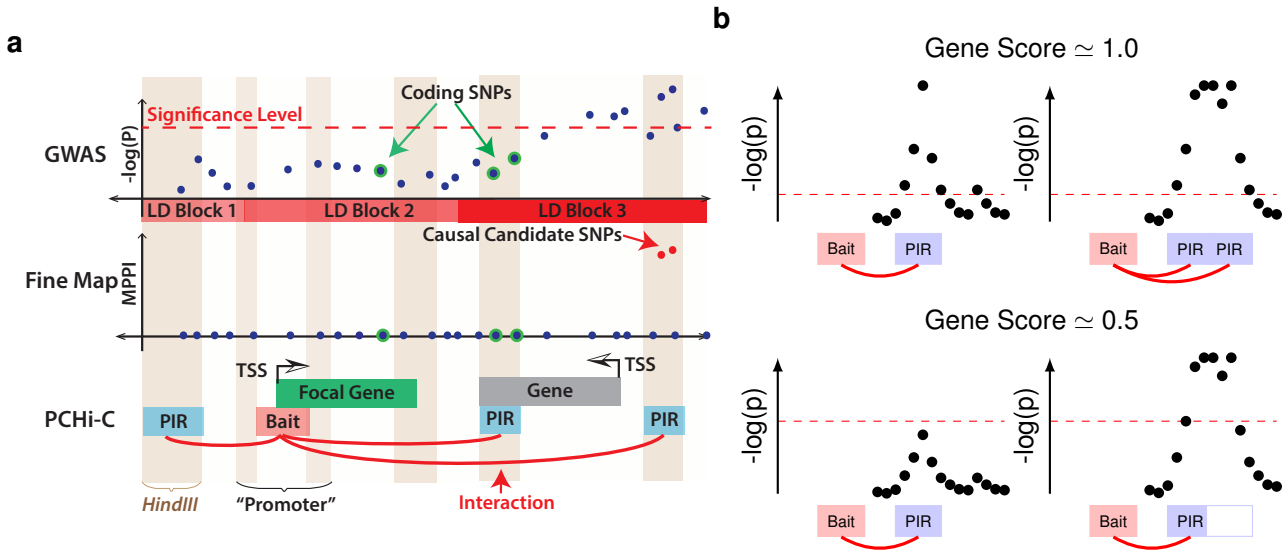
Supplementary Fig. 5: **Relationship of gene expression to PIR number and mRNA half-life. a** RNA-seq expression (counts per million mapped reads, $\log_2$ scale) shows a positive correlation with the number of PIRs indentified through PCHi-C. **b** half-life of mRNA (Raghavan et al. 2002) by gene module in non-activated and activated cells. The most dynamically regulated genes in our time-course, those in the black module, had the shortest half-life ($p = 3 \times 10^{-8}$).



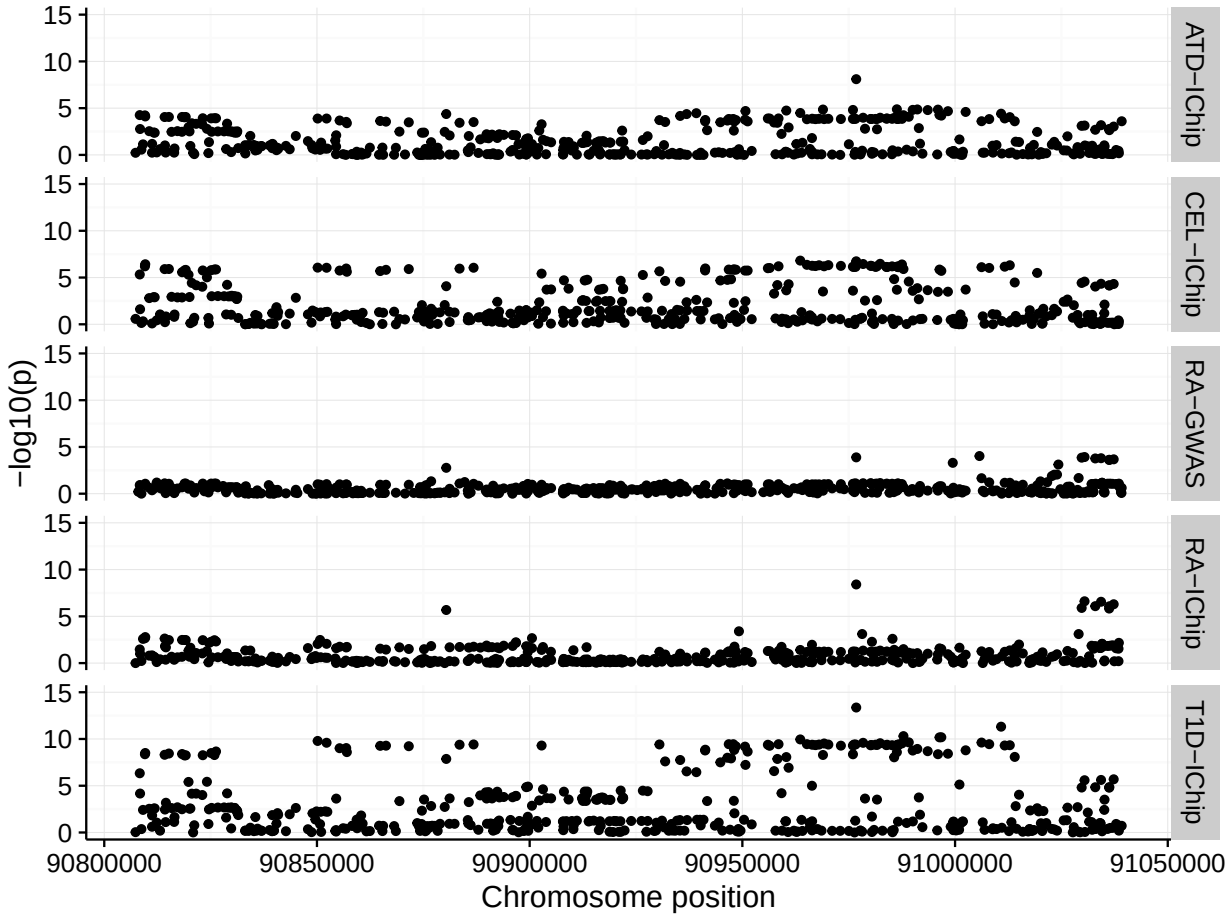
Supplementary Fig. 6: Definition and quantification of regulatory RNAs. CHROMHMM analysis of ChIP-seq marks was used to produce a whole genome segmentation into 15 states. Resulting emission (**a**) and transmission (**b**) matrices are shown. States E4-E11 were defined as regulatory. **c** Neighbouring regions containing promoter or enhancer states (E4-E11) were merged together into regulatory annotations. Expression levels of each regulatory area were quantified using RNA-seq in a strand-aware fashion, to avoid the confounding effect of overlapping genomic features.



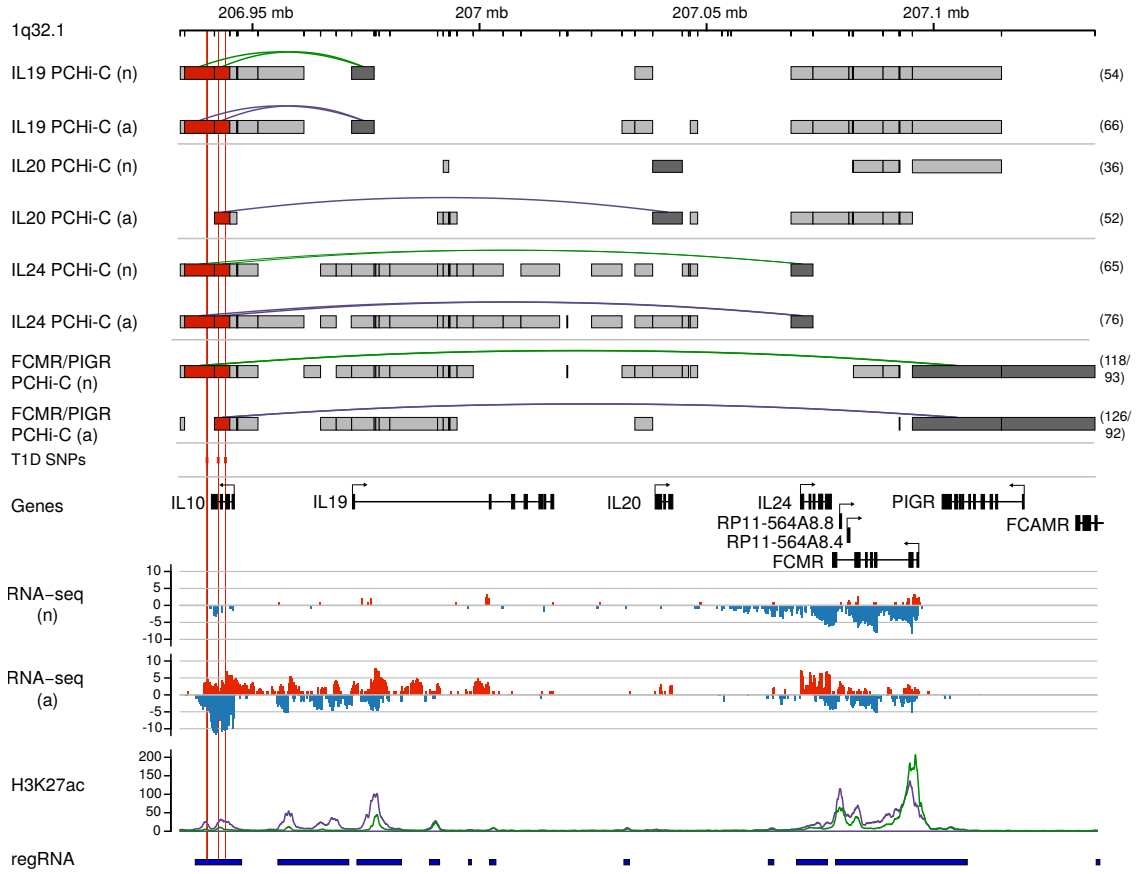
Supplementary Fig. 7: blockshifter calibration. Each panel represents a simulated genotyping density: 1000 genomes (156,082 SNPs); HapMap (44,647 SNPs input,); Ilumina 550k (10,241 SNPs input). Points represent type 1 error rates (alpha=0.05) for the null scenario (no enrichment of GWAS variants in test specific PIRs) and moderate (power 1) and strong (power 2) enrichment scenarios across 4000 simulated GWAS, with differing blockshifter 'SuperBlock' gap size parameter. Error bars represent 95% confidence intervals. Dashed red lines represent the type 1 error rate for Fisher's test of enrichment of variants in test and control PIRs. The naive application of Fisher's test leads to substantial inflation of type 1 error rate, more so in lower-density genotyping scenarios. Blockshifter maintains type 1 error rate control, although a gap size of 5 or more is required to deal with the extended correlation induced by PMI in lower density genotyping scenarios, while Blockshifter power is impacted, as expected, by genotyping density.



Supplementary Fig. 8: **Gene prioritisation using COGS.** We prioritised disease candidate causal genes by integrating GWAS data with PCHi-C interactions using the COGS algorithm. **a** The algorithm uses a Bayesian method to define the marginal posterior probability of inclusion (MPPI, middle panel) for each variant from GWAS data (top panel). We can also calculate the MPPI marginalising across PIRs (light blue, bottom panel), coding variants and promoter regions for each focal gene. *Hind*III fragments are indicated by dark/light vertical shading. **b** Note that the gene score is therefore a function of the strength of GWAS signal, how peaked/diffuse it is, and the interactions. For example, in the top row there are two strong GWAS signals, one peaked, one diffuse, but the PIRs cover all of the most strongly associated SNPs, and in each case the gene score is expected to be close to 1. In the lower left plot, the GWAS signal is less strong, not even genomewide significant, but all the most associated SNPs lie within the PIR. The score will fall, perhaps to around 0.5, reflecting the weaker evidence for disease association. In contrast, the bottom left plot shows a diffuse signal, only part of which lies within a PIR. Although we can be confident the disease is genuinely associated, only about half the fine mapped candidate causal SNPs will lie within a PIR, and the gene score will again fall, to about 0.5. The situations in the lower row are quite different, but will generate similar scores.

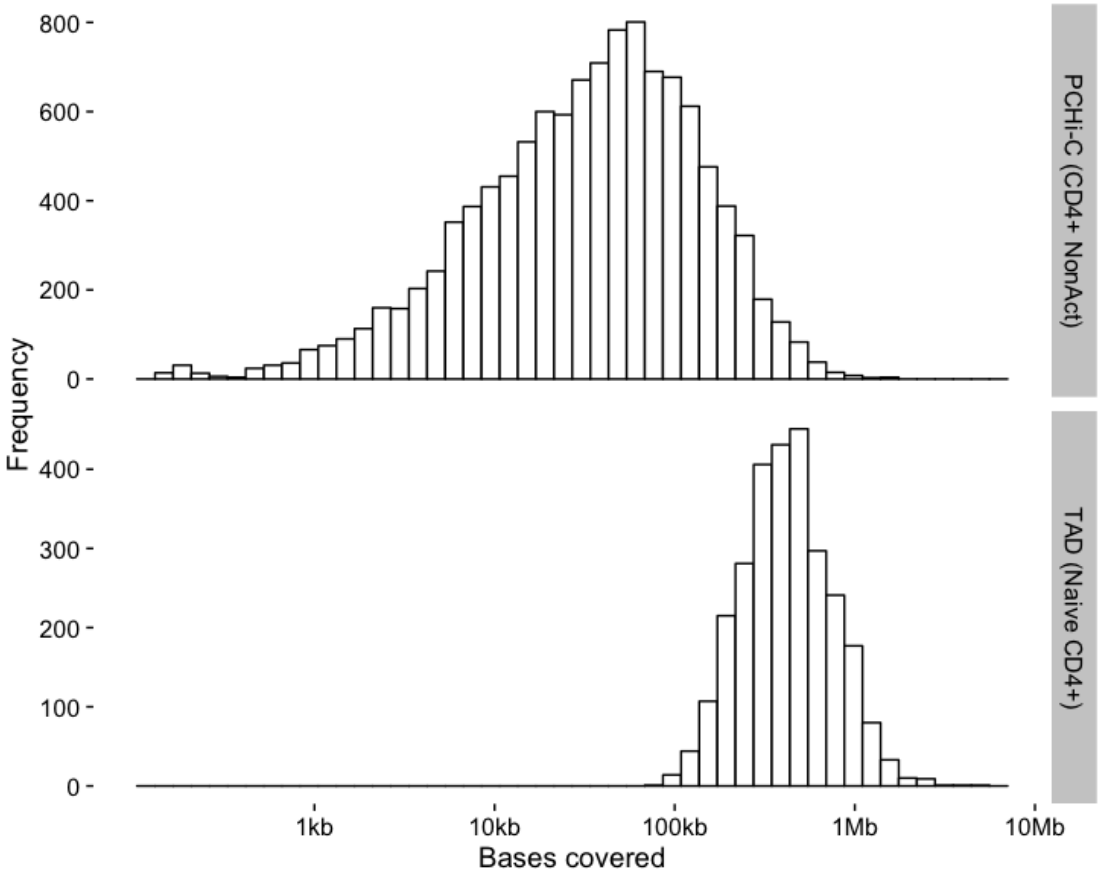


Supplementary Fig. 9: *MDN1* is prioritised for RA through ImmunoChip but not GWAS data. Similar signals are found for ATD and T1D, which also link to *MDN1*, supporting the RA-ImmunoChip result. The lack of prioritisation in the RA-GWAS dataset relates to the weaker evidence for association in this region.

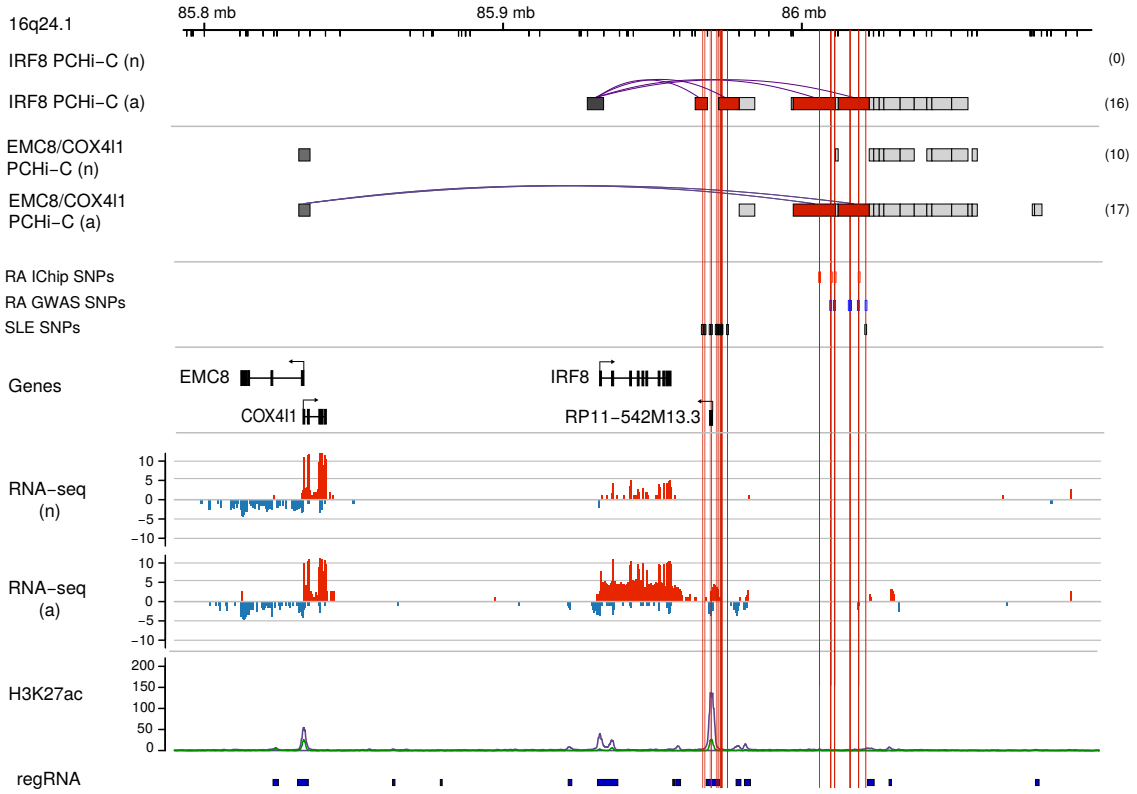


Supplementary Fig. 10: Multiple genes on chromosome 1q32.1 (*IL10*, *IL19*, *IL20*, *IL24*, *FCAMR*/*PIGR*) are prioritised for T1D, CRO and UC. For full legend see **Fig. 5**.

.

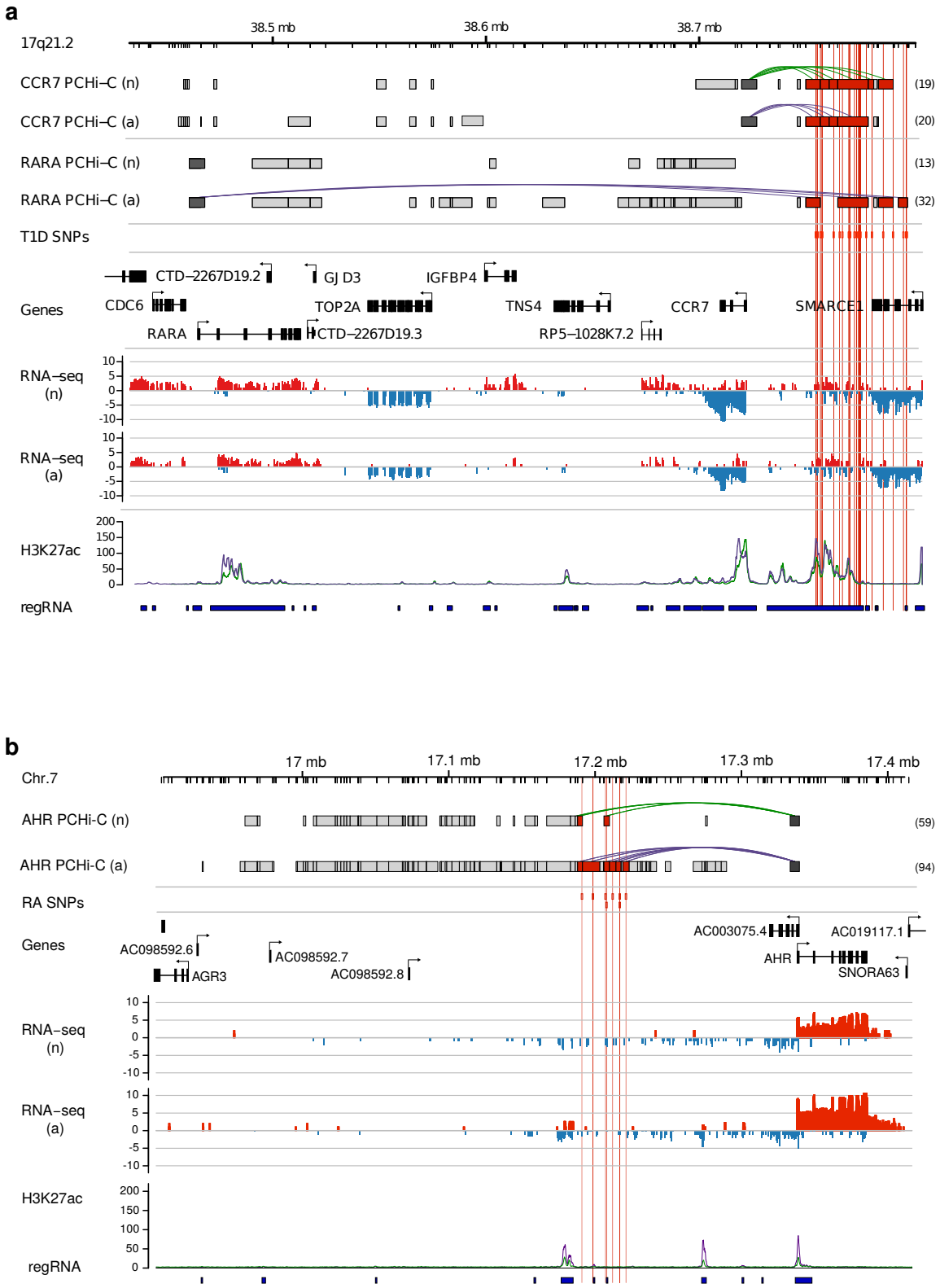


Supplementary Fig. 11: Histograms show the distribution of summed PIR length by gene in non-activated $CD4^+$ T cells (top panel) and TAD length in naive $CD4^+$ T cells. Note the x axis is drawn using a log scale and that for each gene we have included the promoter-baited fragment and its two immediate neighbours to allow that PCHi-C cannot detect very proximal interactions in this range.

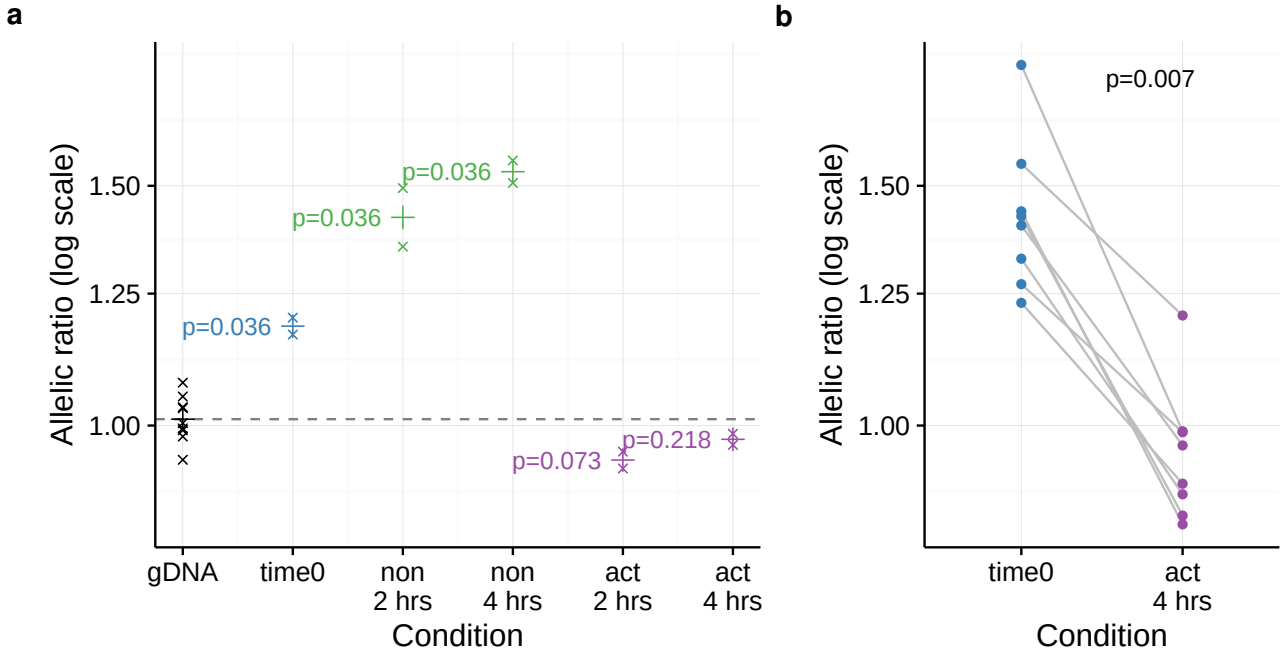


Supplementary Fig. 12: *IRF8* and *EMC8*/*COX4I1* on chromosome 16 are prioritised for RA and SLE. For full legend see **Fig. 5**.

.



Supplementary Fig. 13: **a** *CCR7* and *RARA* on chromosome 17 are prioritised for T1D. **b** *AHR* on chromosome 7 is prioritised for RA in activated CD4$^+$ T cells. For full legend see **Fig. 5**.



Supplementary Fig. 14: **Allelic imbalance in mRNA expression in individuals heterozygous for group A SNPs is confirmed with reporter SNP rs12244380 (*IL2RA* 3' UTR) a**: Allelic imbalance in mRNA expression in total CD4$^+$ T cells from individuals heterozygous for group A SNPs using rs12244380 as a reporter SNP in non-activated (non) and activated (act) CD4$^+$ T cells compared to genomic DNA (gDNA, expected ratio=1). Allelic ratio is defined as the ratio of counts of the allele carried on the chromosome carrying rs12722495:T to that carried on the chromosome carrying rs12722495:C. '×'=geometric mean of the allelic ratio over 2-3 replicates within each of 4-5 individuals, and p values from a Wilcoxon rank sum test comparing cDNA to gDNA are shown. '+' shows the geometric mean allelic ratio over all individuals. **b**: Allelic imbalance in mRNA expression in memory CD4$^+$ T cells differs between *ex vivo* (time 0) and four hour activated samples from eight individuals heterozygous for group A SNPs using rs12244380 as a reporter SNP. p value from a paired Wilcoxon signed rank test is shown.