Simultaneous analysis of open chromatin, promoter interactions and gene expression in stimulated T cells implicates causal genes for rheumatoid arthritis

Jing Yang[1§], Amanda McGovern[2§], Paul Martin[2,3], Kate Duffus[2], Peyman Zarrineh[1], Andrew P Morris[2], Antony Adamson[4], Peter Fraser[5*], Magnus Rattray[1*] & Stephen Eyre[2,6*]

[§]Equal contribution, Joint first authors

*Equal contribution, Joint senior authors

1. Division of Informatics, Imaging & Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, M13 9PT, UK.
2. Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, Manchester Academic Health Science Centre, University of Manchester, Manchester, M13 9PT, UK.
3. Lydia Becker Institute of Immunology and Inflammation, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, M13 9PT, UK.
4. The Genome Editing Unit, Faculty of Biology, Medicine and Health, University of Manchester, Manchester M13 9PT, UK.
5. Department of Biological Science, Florida State University, 319 Stadium Drive, Tallahassee, FL 32306-4295, USA.
6. NIHR Manchester Biomedical Research Centre, Manchester University NHS Foundation Trust, Manchester, UK

*email: Magnus Rattray (magnus.rattray@manchester.ac.uk), Stephen Eyre (steve.eyre@manchester.ac.uk)

Abstract

Genome-wide association studies have identified genetic variation contributing to complex disease risk but assigning causal genes and mechanisms has been more challenging, as disease-associated variants are often found in distal regulatory regions with cell-type specific behaviours. Here, the simultaneous correlation of ATAC-seq, Hi-C, Capture Hi-C and nuclear RNA-seq data, in the same stimulated T-cells over 24 hours, allowed the assignment of functional enhancers to genes. We show how small magnitude changes in DNA interaction and activity dynamics are correlated with much larger changes to dynamics in gene expression and that the strongest correlations are observed within 200 kb of promoters. Using rheumatoid arthritis as an exemplar T-cell mediated disease, we demonstrate interactions of expression quantitative trait locus SNPs with target genes and confirm assigned genes or show complex interactions for 20% of disease associated loci. Finally, we confirm one of the putative causal genes using CRISPR/Cas9.

Introduction

It is now well established that the vast majority of SNPs implicated in common complex diseases from genome-wide association studies (GWAS) are found outside protein coding exons and are enriched in both cell type and stimulatory dependent regulatory regions[1,2]. The task of assigning these regulatory enhancers to their target genes is non-trivial. First, since they can act over long distances, often 'skipping' genes[3]. Second, they can behave differently dependent on cellular context[4,5], including chronicity of stimulation[6]. To translate GWAS findings in complex disease genetics, one of the pivotal tasks is therefore to link the genetic changes that are associated with disease risk to genes, cell types and direction of effect.

Popular methods to link these 'disease enhancers' to genes is to determine physical interactions, with methods such as Hi-C[7], use quantitative trait analysis[4] or exam correlated states[8], with techniques such as ChIP-seq and ATAC-seq, linked to gene expression. The vast majority of these studies, to date, have investigated these epigenomic profiles at either discrete time points[9,10] (e.g. baseline and/or after stimulation), and/or by combining data from different experiments (e.g. ATAC-seq and Hi-C)[9].

Over 100 genetic loci have been associated with rheumatoid arthritis (RA), a T-cell mediated autoimmune disease. Of these, 14 loci have associated variants that are protein-coding and 13 have robust evidence through eQTL studies to implicate the target gene. The remainder are thought to map to regulatory regions, with so far unconfirmed gene targets, although we, and others, have previously shown interactions with disease implicated enhancers and putative causal genes[3,11,12].

Here we have combined simultaneously measured ATAC-seq, Hi-C, Capture Hi-C (CHi-C) and nuclear RNA-seq data in stimulated primary T cells (Fig. 1), to define the complex relationship between DNA activity, interactions and gene expression. We then go on to incorporate fine-mapped associated variants from a T-cell driven complex genetic disease, and validate long range interactions with CRISPR/Cas9, to assign SNPs, genes and direction of effect to rheumatoid arthritis implicated loci.

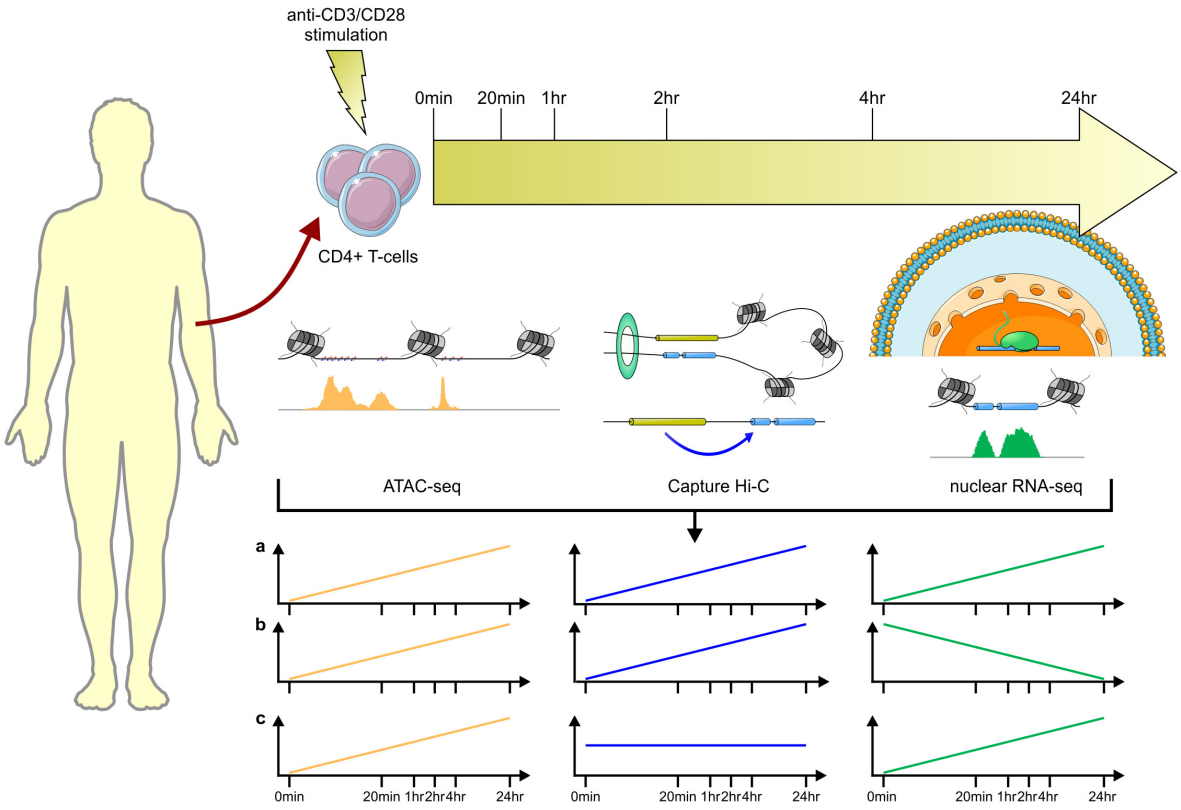


Fig. 1 **Schematic of the study design**. ATAC-seq, CHi-C and nuclear RNA-seq experiments were carried out for unstimulated and stimulated CD4+ T-cell samples at time 0 mins, 20 mins, 1 hr, 2 hrs, 4 hrs and 24 hrs. Time course profiles were created by aligning features (ATAC-seq peaks and CHi-C interactions) across time and counting reads supporting each feature at each time point.

Results

High-quality data from sequenced libraries

A total of 116.7 ± 28.5 million reads per sample were mapped for RNA-seq by STAR[13] with alignment rates over 98% across six time points: 0 mins, 20 mins, 1 hr, 2 hrs, 4 hrs and 24 hrs after stimulation with CD3/CD28. A total of 76,359 ATAC-seq peaks were obtained for both stimulated and unstimulated conditions across the same six time points. Among these peaks, 24,203 peaks are shared across all stimulated and unstimulated conditions, 6,287 peaks are unique to unstimulated state (time 0 mins) and 45,869 peaks only appear after stimulation. 74,583 peaks were retained with peak sizes within a 5-95 percentile range of 123 bp to 1414 bp after merging peaks across the six time points, with each peak appearing in at least one time point. A total of 7,081 baits were designed to capture 5,124 genes for CHi-C, among which 6,888 baits were successfully recovered with 97% on target. On average, 90.9±20.2 million unique di-tags were generated from two individuals for five experimental time points: 0 mins (unstimulated), and then 20 mins, 1 hr, 4 hrs and 24 hrs. A total of 271,398 CHi-C interactions were generated from the time course data and interactions were retained as features when at least

one time point showed a significant interaction. Of these interactions, 94% occurred within the same chromosome and 57% were within 5 Mb distances of promoters.

Data consistency with previous studies

We compared our data with published datasets from the same cell-type and stimulation. Our CHi-C data, both unstimulated or stimulated for 4 hrs, demonstrated good consistency with published data[10] (Supplementary Fig. 1a). When restricting all the interactions from both unstimulated and stimulated to those that share the same baits, we found 57% of interactions (27,794/48,581) to overlap (by at least 1 bp) between our study and previously identified interactions[10] and this increases to 73% for interactions within 5 Mb of promoters (26,836/36,706) and 87% within 200 kb of promoters (8,367/9,631). This strongly suggests that the interactions between promoters and active enhancers within 200kb are consistent, robust and reproducible between studies. We found 18,162 genes with evidence of CD4+ T cell expression in at least one time point in our RNA-seq dataset and these include 96% (4,903/5,124) of the genes included on the CHi-C design. We considered genes classified as "Persistent repressed", "Early induced", "Intermediate induced I", "Intermediate induced II" and "Late induced" in a previous study[4], and found that these genes exhibited similar patterns of expression in our RNA-seq data (Supplementary Figure 2e-i). Comparison of our ATAC-seq data to CD4+ baseline (unstimulated) and 48 hrs after stimulation peaks from a similar dataset[9] revealed strong concordance (Supplementary Table 2): 71% of peaks (21,549/30,403) from our unstimulated data overlapped with their baseline peaks[9], while 75% of our peaks (22,911/30,593) at 24 hrs after stimulation overlapped with their peaks at 48 hrs after stimulation[9]. We also observed a similar magnitude increase in the number of ATAC-seq peaks for merged unstimulated and stimulated data.

Chromatin conformation dynamics

It is well established that gene expression changes with time after stimulation in CD4+ cells[4] and we find similar changes to previous studies, with a range of dynamic expression profiles corresponding to genes activated early, intermediate or late, or repressed (Supplementary Fig. 2e-i). It is less well established how chromatin structure changes post stimulation, in the form of A/B compartments, topologically associating domains (TADs) and individual interactions, or how enhancer activity and open chromatin changes over time.

Based on Hi-C matrices with resolutions of 40 kb, 1,230 TADs were recovered from our study with an average size of 983.2 kb. On average, 84% of TADs intersect between replicates and 80% of TADs intersect across different time points with reciprocal 90% region overlap, showing that most of the differences observed between the called TADs are due to experimental variation rather than conformational changes. Interestingly, we do see that the percentage falls to 72% when comparing the TAD overlap before 24 hrs to 24 hrs, illustrating more substantial dynamic changes in TADs over longer times (Supplementary Table 3). Fig. 2b shows the Stratum adjusted Correlation Coefficient (SCC) between Hi-C datasets[14] and shows a slight but significant reduction in correlation as the time separation of experiments increases, consistent with our observations regarding TADs.

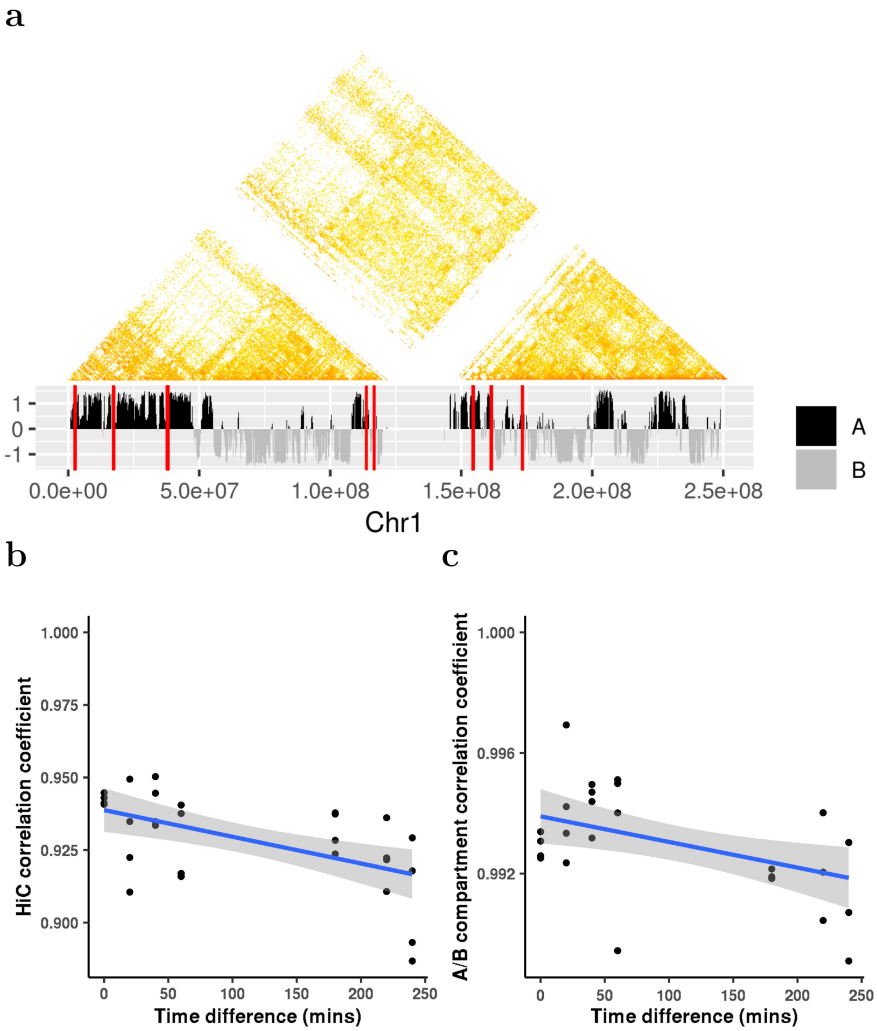


Fig. 2 **Illustration of Hi-C dynamics**. **a**, Hi-C interaction matrix (100 kb resolution) of replicate 1 for chr1 at time 0 mins (upper) with corresponding A/B compartments (lower), where red lines represent positons of SNPs. **b**, Correlation changes between Hi-C data with respect to differences between times, where the blue line is the fitted linear line for the correlation coefficients with grey area representing 95% confidence region of the linear fitting. **c**, Correlation changes of A/B compartments with respect to the differences between times, where the blue line and the shaded area share the same information as conveyed in plot **b**.

We recovered 1,136 A compartments and 1,266 B compartments merged across the time course data, with the maximum compartment sizes being 39.5 Mb and 34.5 Mb, respectively. Fig. 2c shows the correlation between A/B compartment allocations, demonstrating a slight but significant reduction between experiments as time separation increases.

These results are broadly consistent with other studies, demonstrating how the higher chromatin conformation states, in the form of A/B compartments and TADS, is largely invariant between cell types[11]. Here, we demonstrate similar levels of consistency in a single cell-type post stimulation. There was also, as expected[7], a

high degree of correlation between A/B compartments and marks of histone activity, such as H3K27ac, with A compartments overlapping histone marks of strong DNA activity (Supplementary Fig. 4a-c).

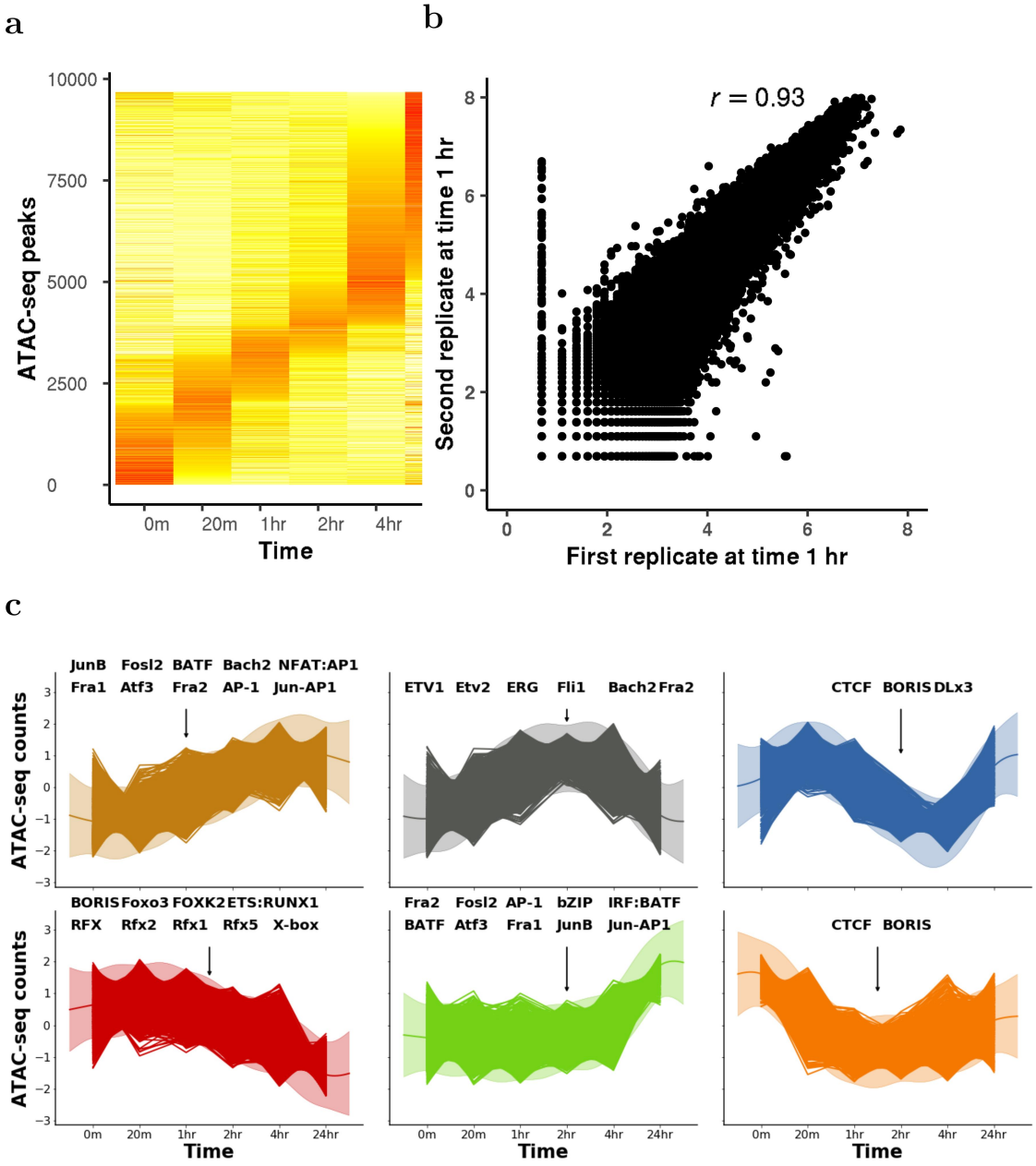


Fig. 3 **Illustration of ATAC-seq time course profile dynamics**. **a**, heatmap of ATAC-seq counts data for peaks showing evidence of temporal dynamics. **b**, ATAC-seq data correlation between first replicate and second replicate at time 1 hr. r is the Pearson Correlation coefficient. **c**, Clustering of ATAC-seq time course data using a Gaussian process mixture model. Significantly enriched DNA-binding MOTIFs in each peak (using static peaks are background) are labelled in each cluster.

In contrast to the relative invariance of TADs and A/B compartments, our CHi-C data, analysing interactions between individual restriction fragments, showed a much greater degree of dynamics. We used the Bayesian Information Criterion (BIC)[15] and a χ^2 test to compare a dynamic (Gaussian process) model to a static model[16] for CHi-C interaction counts data across time. We found 24% (63,843/271,398) of CHi-C links with evidence of change over time ($BIC_{dynamic} < BIC_{static}$) and 7.5% of interactions showed stronger evidence of change over time (20,224/271,398, χ^2 test, $p < 0.05$), among which 24% (4,837/20,224) are within 200 kb of promoters.

Open chromatin dynamics

Our ATAC-seq time course data showed good correlations across replicates (Fig. 3b). We compared a dynamic (Gaussian process) and static model for ATAC-seq time course data to identify changes in open chromatin across time and found 11% (7,852/74,583) of ATAC-seq peaks with evidence of change over 24 hrs ($BIC_{dynamic} < BIC_{static}$) with 2,780 of these peaks showing stronger evidence of change (χ^2 test, $p < 0.05$). A heatmap of ATAC-seq time course data (Fig. 3a) demonstrates six broad patterns of change (Fig.3c). Mapping Transcription Factor Binding Sites (TFBS) motifs under these broad clusters revealed a strong enrichment of transcription factors known to be important in CD4+ stimulation and differentiation. The AP-1 TFBS (e.g. BATF) motif was shown to be enriched in low to high activity, strong enrichment of ETS/RUNX1 TFBS was seen in models of high to low activity, and a strong enrichment of CTCF and BORIS motifs was observed in the models that demonstrated transient dynamics before returning to baseline after 24 hours. These findings match those reported in a previous study of ATAC-seq data in CD4+ T cells stimulated with CD3/CD28[9]. There it was demonstrated that AP-1/BATF motif was enriched in stimulated ATAC-seq peaks, ETS/RUNX in unstimulated cells and CTCF/BORIS motifs were detected under the 'shared' unstimulated and stimulated peaks, closely matching our findings.

Correlating chromatin dynamics with gene expression

We next went on to test whether these dynamic measures of DNA activity, interaction and expression exhibited any correlation between their time course profiles. Previous studies, using measurements of H3K27ac, Hi-C and expression across different cell types, demonstrated how subtle changes in contact frequency correlated with larger changes in active DNA and expression[17]. We wanted to determine the nature of this relationship in our data, from a single, activated cell type. We used a randomisation procedure to identify whether the number of correlations observed at a particular level could be considered significantly enriched (see Methods). We show an enrichment for extreme correlations between ATAC-seq, CHi-C and RNA-seq datasets, particularly an enrichment for high positive correlations within a distance of 200 kb around promoters (Fig. 4a), suggesting that functional, interactive correlations are most common within 'contact domains', as supported by previous findings, where the median distance between H3K27ac loop anchors and interacting otherEnds (130 kb)[5] and the median distance of cohesion constrained regulatory DNA-loops (185 kb)[7] are typically within a ~200 kb range. We also see some evidence for significant enrichment of correlations across longer distances, for example with gene expression and interaction strength more likely to be either positively or negatively correlated when associated with significant long-distance looping events.

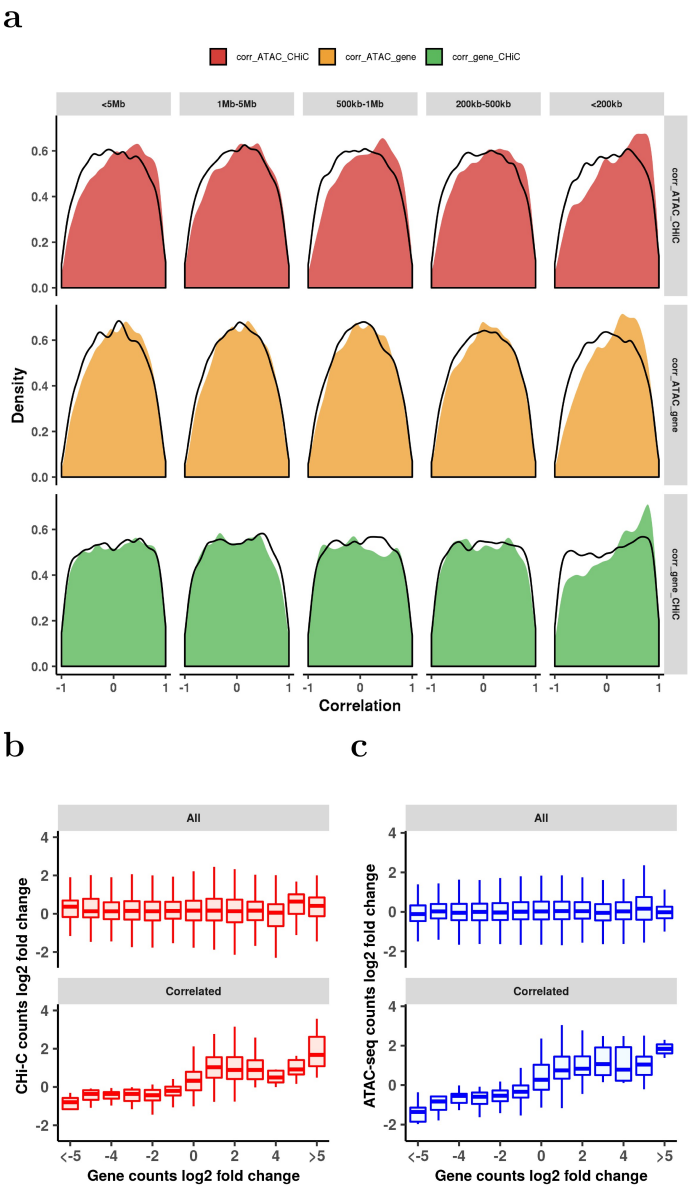


Fig 4. **Illustrations of the correlations between ATAC-seq, CHi-C and RNA-seq time course profiles**. **a**, Density plots of the Pearson correlation coefficients between ATAC-seq and CHi-C (labelled as corr_ATAC_CHiC), ATAC-seq and gene (labelled as corr_ATAC_gene) and gene and CHi-C (labelled as corr_gene_CHiC) under various distance ranges around promoters. Distances ranges include those less than 5 Mb (labelled as <5Mb), between 1 Mb and 5 Mb (labelled as 1Mb-5Mb), between 500 kb and 1 Mb (labelled as 500kb-1Mb), between 200 kb and 500 kb (labelled as 200kb-500kb) and less than 200 kb (labelled as <200kb). Black lines represent the density plots of the corresponding random background. **b**, Comparison of the log2 fold change between CHi-C and gene data for all dataset (upper) and those highly correlated ones with Pearson correlation coefficients between ATAC-seq, gene and CHi-C over 0.5 (lower). **c**, Comparison of the log2 fold change between ATAC-seq and gene data for all datasets (upper) and those highly correlated ones with Pearson correlation coefficients between ATAC-seq, gene and CHi-C over 0.5 (lower).

Boxplots of the log fold change in ATAC-seq, CHi-C and RNA-seq intensity in the highly correlated regions (Fig. 4b,c) revealed how relatively small changes in both ATAC-seq and CHi-C intensity (~2 fold change) correlated with larger changes in expression (~5 fold change). This is consistent with similar patterns observed in different cell types[17].

Previous studies have indicated how i) using eQTL data, ~50% of ATAC-seq peaks are already active/poised before influencing gene expression[18] , ii) using HiChIP data, expression can be correlated with either H3K27ac or interactions[5], and iii) empirical ranking of enhancers by CRISPR corresponds most strongly when combining terms for interaction and activity[19], all suggesting both interactions and activity have important roles in gene regulation. Examining the relationship of CHi-C interactions, DNA dynamics and expression in closer detail in our data with 200 kb distance between bait and otherEnd fragments revealed three broad patterns of dynamics associated with four clustered gene expression patterns (Supplementary Fig. 5): around 8% (469/5,939) of links were associated with dynamic ATAC-seq peaks only (Supplementary Fig. 5a,b), 32% (1,901/5,939) were associated with dynamic CHi-C interactions only (Supplementary Fig. 5c,d) and 6% (349/5,939) were associated with dynamics in both (Supplementary Fig. 5e,f). Our findings, together with previous studies, therefore suggest that both activity and interactions are independently important in gene regulation and that subtle change in interaction and ATAC-seq intensity has a larger effect on gene expression.

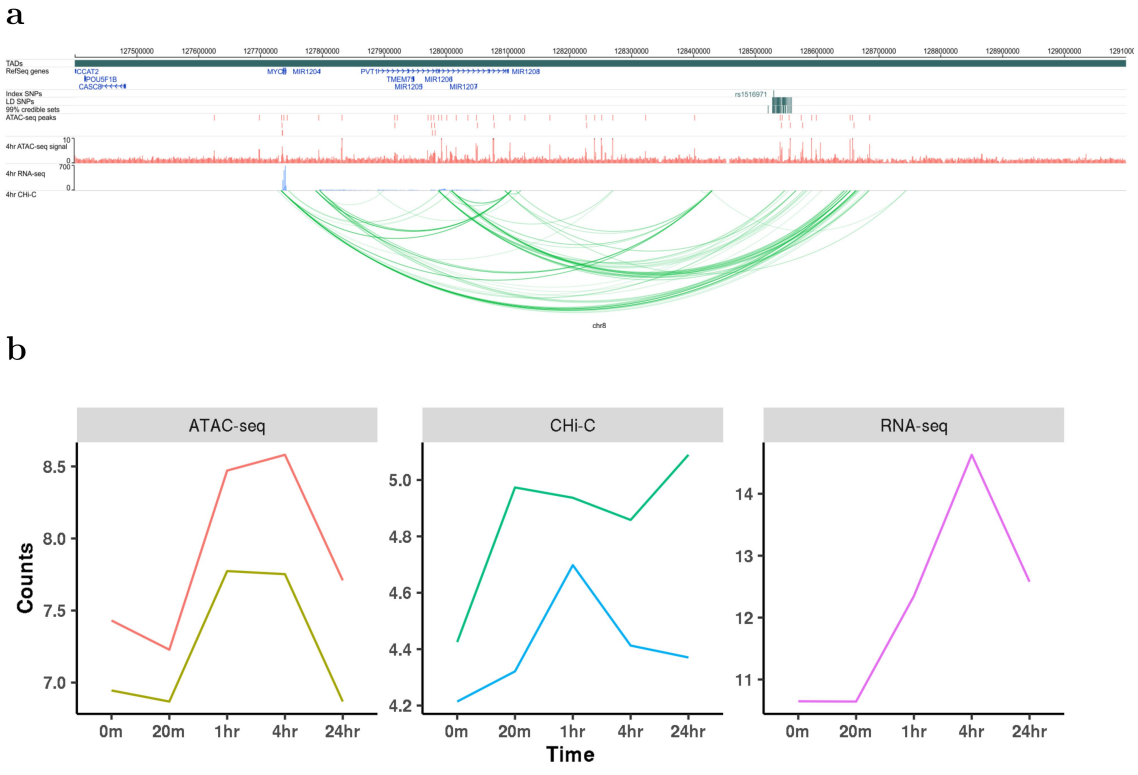


Fig. 5 **Illustration of genomic interaction activities around MYC**. **a**, Screenshot of the SNPs (dark green), ATAC-seq peaks (red), RNA-seq (blue) and CHi-C interactions (green) around MYC at time 4 hrs. **b**, Time course profiles of ATAC-seq (left), CHi-C (middle) and RNA-seq (right) of the data associated with SNPs data around MYC.

Prioritisation of causal genes in RA loci

We considered 80 loci previously associated with RA at genome-wide significance in European ancestry[20,21]. For each locus, we constructed a 99% credible set of SNPs that accounted for 99% of the probability of containing the causal variant. We find that 97% (2,131/2,192) of RA-associated variants from our 99% credible SNP set lie within A compartments across all times while 28 (1%) lie consistently within B compartments after stimulation. 15 SNPs are found in regions that change between A and B over time. Of these 15 SNPs, we identified a set of RA-associated variants on chromosome 1, proximal to the *TNFSF4* and *TNFSF18* genes, that were initially contained within an inactive B compartment at 20 mins and were then found in an A compartment at 4 hrs (Supplementary Fig. 4d).

We next investigated whether we could map RA GWAS implicated ATAC-seq peaks to genes, identify the likely causal SNPs within the peaks and determine a mechanism and direction of effect through the correlated expression data. We found that 42/100 GWAS loci contained ATAC-seq peaks with at least one associated SNP (66 associated SNPs) that interacted and correlated with the expression of 167 genes, an average of 2.5 genes per peak (Supplementary Table 6). Of these 42 loci, there are 17 where we either a) show correlated interaction of an eQTL with the target gene, b) confirm the assigned causal gene with correlated interaction of a RA ATAC-seq peak or c) show that the locus is likely to be complex and suggest novel RA causal genes:

i) Interactions confirmed with eQTLS. 29 RA loci contained 50 genes where the top eQTL variant for the gene was in the credible set of disease associated SNPs. Of these 50 genes, 41 (82%) were either located within the interval covered by the RA credible set, or interacted with the RA associated eQTL SNP (Supplementary Table 7). All eQTL SNPs fell within the same TAD as the target gene (maximum distance 500 kb), whilst of the 21 eQTL SNPs within 200 kb of the target gene, 15 (75%) demonstrated a correlated interaction. These interactions with known RA susceptibility SNPs support causal genes in 8 loci, including *CD5, PXK, TPD52, IL6ST* and *CDK6*, and also implicate a single interacting SNP/ATAC-seq peak for causality in each locus (Supplementary Table 7).

ii) Confirming assigned genes. For 7 loci, our correlated, dynamic data provides the first biological evidence for the currently assigned gene (Supplementary Table 6). These genes include *DDX6, PRKCH, RBPJ, PVT1* and *ERBB2*, with many interactions confirming genes up to 200 kb, and one over 800 kb (*PVT1*), from the associated SNP, but again interactions are always constrained within TADs. These data have the ability to limit the number of putative causal SNPs/ATAC-seq peaks for each locus. For example, there are 18 SNPs within the 99% credible SNP set for the *RBPJ* locus, but this reduces to 2 SNPs within 2 ATAC-seq peaks that interact and correlate with gene expression (Supplementary Fig 6a).

iii) Novel and complex gene regions. For a number of regions, we suggest complex or novel relationships between associated SNP regions and putative causal genes (Supplementary Table 6). On chromosome 10, an intronic region within the *ARID5B* gene, containing SNP variants associated with RA, interacts with *RTKN2,* involved in the NFKB pathway, and containing nsSNPs associated with Asian RA[22] (Supplementary Fig 6b) . Similarly on chromosome 3, a region with RA associated variants intergenic of *EOMES* interacts with *AZI2* (an activator of NFKB), suggesting the region contains an enhancer that could potentially control two important genes in the T cell immune pathway (Supplementary Fig 6c). Two regions in particular provided insight into potential novel genes for RA. Two ATAC-seq peaks, containing 5 SNPs that are associated with RA on

chromosome 8, both interact with *PVT1* and *MYC1*, situated some 450-800 kb away from these peaks (Fig. 5). Here we demonstrate a positive correlation between the ATAC-seq peak dynamics and gene expression for both *PVT1* and *MYC1* (r^2=0.67 and 0.68 respectively). Interestingly, this ATAC-seq peak region has previously been demonstrated to be a key repressor of *MYC1* gene expression, following a comprehensive CRISPRi screen in K562 erythroleukemia cells[19]. We therefore confirm this relationship between a distant regulatory region and *MYC1* expression in primary T cells, highlighting the likelihood that this gene has a role in the susceptibility to RA.

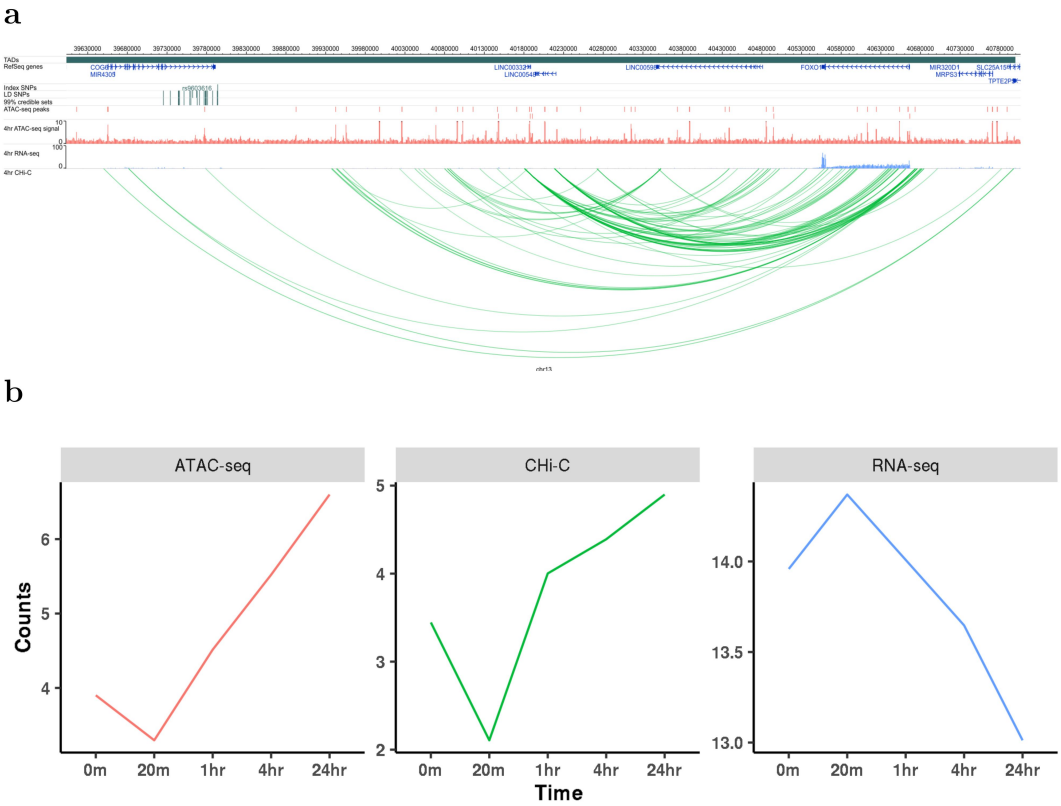


Fig. 6 **Illustration of genomic interaction activities around FOXO1**. **a**, Screenshot of the SNPs (dark green), ATAC-seq peaks (red), RNA-seq (blue) and Chi-C interactions (green) around FOXO1 at time 4hr. **b**, Time course profiles of ATAC-seq (left), Chi-C (middle) and RNA-seq (right) of the data associated with SNPs data around FOXO1.

Confirmation of correlated interaction with CRISPR/Cas9

Finally we show how an ATAC-seq peak, containing a SNP associated with RA, intronic of the *COG6* gene interacts with, and is correlated with the expression of, the *FOXO1* gene located some 900 kb away (Fig 6). We wanted to investigate whether this dynamic ATAC-seq peak is functionally interacting with the *FOXO1* promoter, a transcription factor involved in T cell development, and a gene that has previously been strongly implicated in RA through functional immune studies in patient samples[23–26] . We used CRISPRa, with dCas9-p300, and the HEK

cell line. Our results demonstrate that when we activate the *COG6* intronic enhancer with this system and targeted gRNAs, not only do we observe a consistent increase in the *COG6* mRNA expression itself, we obtain robust, reproducible up regulation of *FOXO1* gene expression (Fig. 7). Although the associated variant in this region is a strong eQTL for *COG6*, this CRISPR validation of the correlated interaction, DNA activity and expression data, alongside the previous immunological studies, imply that the associated enhancer may have diverse roles on a number of genes within this 1 Mb TAD region, and that GWAS implicated enhancers should not necessarily be assigned to single genes.

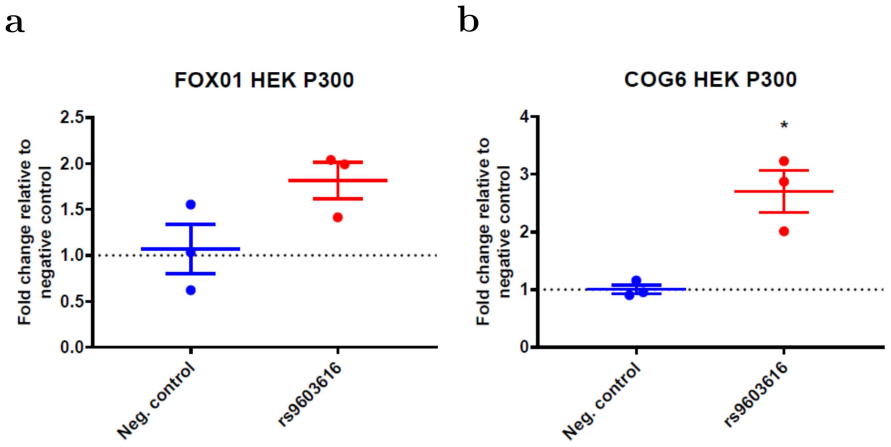


Fig. 7 CRISPR dCAS9 activation (CRISPRa) targeting of an RA associated variant. The region around an associated RA variant (rs9603616) on chromosome 13 intronic of the COG6 genewas targeted in the HEK2937T cell line, using p300 as the activator. **a**, Fold change (qPCR) affect on FOXO1 gene expression compared to negative control. **b**, Fold change (qPCR) affect on COG6 gene expression compared to negative control.

Discussion

We have generated a unique, high quality, high value resource, correlating a range of dynamic data, to inform the assignment of regulatory regions to genes. Our analysis adds to the growing evidence that the relationship between enhancers and promoters is complex, that interactions are more strongly correlated within a distance of 200 kb, and that they are mostly constrained within TADs.

Using CD4+ T-cells, stimulated with CD3/CD28, we analyse ATAC-seq, RNA-seq, Hi-C and CHi-C over 24 hours. We show that the conformation of the DNA at a higher structural level of A/B compartments[27] and TADs[28] remains relatively constant throughout the stimulation time course. In contrast, DNA interactions at the level of discrete contacts, for example between open chromatin and gene promoters, is highly dynamic post stimulation, with over 30% of individual interactions showing some degree of change over 24 hours; however only a minority of these changes are correlated with a change in expression. Our results suggest ATAC-seq peaks are associated with 2.0 CHi-C interactions on average, with each gene making contact with on average 7.7 ATAC-seq peaks. Although this correlation can occur over large distances (up to 5 Mb in our data), it is strongly enriched within 200 kb. We also demonstrate how subtle changes in both ATAC-seq and interaction intensity can have more marked effects on gene expression. Our data therefore suggests that, when assigning GWAS variants to putative causal genes that all genes within 200 kb, all within the TAD structure, and multiple causal genes, should be

considered as candidates for functional validation. In addition, as a proof of principle, we have confirmed one novel, long range (~1 Mb) gene target using CRISPR/Cas9.

We also implicate genes in RA associated loci not previously highlighted as likely causal from GWAS, most notably MYC and FOXO1. MYC is a proto-oncogene transcription factor, involved in pro-proliferative pathways, highly expressed in a wide range of cancers. It has long been known that this gene is expressed in the RA synovium[29,30], potentially playing a role in the invasiveness of these cells. More notably for our study, it has recently been demonstrated how a CD4+ T-cell subset from RA patients demonstrates higher autophagy, and that MYC is a central regulator of this pathway. Here it was suggested that autophagy could contribute to the survival of inflammatory T cells in patients, particularly a pathogenic-like lymphocyte (CPL) subset, found in inflamed joints and associated with disease activity. Similarly FOXO1 has long been established to be downregulated in both synovium and blood from RA patients, and is also correlated with disease activity[25]. FOXO1 is a transcription factor, thought to play a role in apoptosis and cell cycle regulation, where reduced expression in RA is suggested to have a role in the accumulation of fibroblasts in the disease synovium[25]. In our study, for both these genes with established biological mechanisms and expression patterns relevant to RA, we have demonstrated how genetic variants that lead to an increased risk of developing disease are physically linked and correlated with gene expression, providing evidence that these genes may be causal instigators in disease, and not simply on the pathways that are dysregulated in disease.

Our results indicate that, first, both DNA activity and interaction intensity are independently important in the regulation of genes; second, since a minority of interactions correlate with gene expression, simply assigning target genes by interactions is too simplistic. Instead, other methods, such as the simultaneous measurement of DNA activity and expression data followed by CRISPR experimental validation, are required to confidently assign genes to GWAS implicated loci. Finally, we confirm that subtle changes in interaction intensity are correlated with much larger changes to gene expression. In combination our findings have important implications for fully exploiting GWAS data, assigning causal SNPs, genes, cell types and mechanism to trait associated loci, on the pathway to translating these findings into clinical benefit.

Methods

Isolation of CD4+ T-cells and stimulation time course

Primary human CD4+ T-cells were collected from three healthy individuals with informed consent and with ethical approval (Nat Rep 99/8/84). Samples were isolated from PBMCs using an EasySep T-cell isolation kit (StemCell), plated in 6-well plates then stimulated with CD3/CD28 Dynabeads (Life Technologies) over a period of 24-hours, with samples removed at the appropriate time point and processed according to the experiment the cells would be used for. Unstimulated samples were also prepared (t=0 sample). For Hi-C experiments, CD4+ T-cells were harvested and fixed in formaldehyde, samples for RNA-seq (5 x 10^6 CD4+ T-cells) were stored in RNA CellProtect reagent before extraction, and samples for ATAC-seq (50,000 cells) were processed immediately. ATAC-seq samples from three individuals were taken at time 0 mins, 20 mins, 1 hr, 2 hrs, 4 hrs and 24 hrs. Two pooled nuclear RNA-seq replicates were taken at the same time points as ATAC-seq samples. Two pooled Hi-C and CHi-C replicates were taken at 0 min, 20 min, 1 hr and 4 hrs and one sample for Hi-C and CHi-C was taken at 24 hrs, respectively.

Library generation for CHi-C and Hi-C

To generate libraries for Hi-C experiments, 8-10 x 10^6 CD4+ T-cells were harvested at the appropriate time point and formaldehyde crosslinking carried out as described in Belton et al[31]. Cells were washed in DMEM without serum then crosslinked with 2% formaldehyde for 10 minutes at room temperature. The crosslinking reaction was quenched by adding cold 1M glycine to a final concentration of 0.125M for five minutes at room temperature, followed by 15 minutes on ice. Crosslinked cells were washed in ice cold PBS, the supernatant discarded and the pellets flash-frozen on dry ice and stored at -80^{o}C.

Hi-C libraries were prepared from fixed CD4+ T-cells from three individuals which were pooled at the lysis stage to give ~30 million cells. Cells were thawed on ice and re-suspended in 50ml freshly prepared ice-cold lysis buffer (10mM Tris-HCl pH 8, 10mM NaCl, 0.2% Igepal CA-630, one protease inhibitor cocktail tablet). Cells were lysed on ice for a total of 30 min, with 2x10 strokes of a Dounce homogeniser 5 min apart. Following lysis, the nuclei were pelleted and washed with 1.25xNEB Buffer 2 then re-suspended in 1.25xNEB Buffer 2 to make aliquots of 5-6x10^6 cells for digestion. Following lysis, libraries were digested using HindIII then prepared as described in van Berkum et al[32] with modifications described in Dryden et al[33]. Final library amplification was performed on multiple parallel reactions from libraries immobilised on Streptavidin beads using 8 cycles of PCR if the samples were to be used for CHi-C, or 6 cycles for Hi-C. Reactions were pooled post-PCR, purified using SPRI beads and the final libraries re-suspended in 30µl TLE. Library quality and quantity was assessed by Bioanalyzer and KAPA qPCR prior to sequencing on an Illumina HiSeq2500 generating 100bp paired-end reads (Babraham sequencing facility).

Solution hybridisation capture of Hi-C library

Pre-CHi-C libraries corresponding to 750ng were concentrated in a Speedvac then re-suspended in 3.4µl water. Hybridisation of SureSelect custom capture libraries to Hi-C libraries was carried out using Agilent SureSelectXT reagents and protocols. Post-capture amplification was carried out using 8 cycles of PCR from streptavidin beads in multiple parallel reactions, then pooled and purified using SPRI beads. Library quality and quantity was assessed by Bioanalyzer and KAPA qPCR prior to sequencing on an Illumina HiSeq2500 generating 100 bp paired-end reads (Babraham sequencing facility).

Defining regions of association for bait design

All independent lead disease-associated SNPs for RA were selected from both the fine-mapped Immunochip study[20] and a *trans*-ethnic GWAS meta-analysis[21]. This resulted in a total of 138 distinct variants associated with RA after exclusion of *HLA*-associated SNPs. Associated regions were defined by selecting all SNPs in LD with the lead disease-associated SNP (r^2>=0.8; 1000 Genomes phase 3 EUR samples; May 2013). In addition to the SNP associations, credible SNP set regions were defined for the Immunochip array at a 95% confidence level.

Target Enrichment Design

Capture oligos (120 bp; 25-65% GC, <3 unknown (N) bases) were designed to selected gene promoters (defined as the restriction fragments covering at least 500bp 5' of the transcription start site (TSS)) using a custom Perl script within 400 bp but as close as possible to each end of the targeted HindIII restriction fragments and submitted to the Agilent eArray software (Agilent) for manufacture. Genes were selected as follows: all genes

within 1Mb upstream and downstream of associated RA SNPs from Eyre *et al*[20] and Okada *et al*[21] as previously described; all gene promoters showing evidence of interacting with an associated region in our previous CHi-C study using GM12878 and Jurkat cell lines; all genes contained within the KEGG pathways for "NF-kappa B signalling", "Antigen processing and presentation", "Toll-like receptor signalling", "T cell receptor signalling" and "Rheumatoid arthritis"; all genes showing differential expression in CD4+ T-cells after stimulation with anti-CD3/anti-CD28; all genes from Ye *et al*[4] within the 'Early induced', 'Intermediate induced I' and 'Intermediate induced II' categories; and all genes from the Ye *et al*[4] NanoString panel. Additionally control regions targeting the HBA, HOXA and MYC loci were included for quality control purposes**.**

Library generation for RNA-seq

Nuclear RNA-seq was used to quantify nascent transcription to determine changes through time. Five million CD4+ T-cells were harvested, stored in Qiagen RNAprotect solution and the nuclear RNA isolated using a Qiagen RNeasy kit and quantified. Samples were either pooled in equal amounts (same individuals as for Hi-C to create matched samples), or processed individually to give duplicate samples. Libraries for RNA-seq were prepared using the NEB Next Ultra Directional RNA-seq reagents and protocol using 100ng of nuclear RNA as Input. Each library was sequenced on half a lane of an Illumina HiSeq2500 generating 100bp paired-end reads (Babraham sequencing facility).

Library generation for ATAC-seq

ATAC-seq libraries were generated from 50,000 CD4+ T-cells from three individual samples using the protocol detailed in Buenrostro et al[34] using the Illumina Nextera DNA Sample Preparation Kit. Each library was sequenced on half a lane of an Illumina HiSeq2500 generating 100 bp paired-end reads (Babraham sequencing facility).

Hi-C data processing

Hi-C data were mapped to GRCh38 by HiCUP[35]. The maximum and minimum di-tag lengths were set to 800 and 150, respectively. HOMER[36] Hi-C protocol was applied to Hi-C bam file and normalized Hi-C matrices were generated by analyzeHiC command from HOMER with resolution of 40,000bp (analyzeHiC –res 40000 –balance). TADs were generated by the command findTADsAndLoops.pl (-res 40000). A/B compartments were generated by runPCA.pl (-res 40000) followed by findHiCCompartments.pl with the default parameters to generate compartments A and –opp parameters to generate compartments B.

CHi-C data processing

CHi-C data were mapped to GRCh38 by HiCUP. The maximum and minimum di-tag lengths were set to 800 and 150, respectively. CHiCAGO[37] was applied to each bam file with the CHiCAGO score set to 0. Counts data for each interaction were extracted from the .rds files generated by CHiCAGO. Time course interactions were concatenated. Those interactions with at least one time point having CHiCAGO score over 5 were kept. Bait-to-bait interactions were registered as two interactions with either side defined as 'bait' or 'otherEnd'.

ATAC-seq data processing

Individual ATAC-seq reads data were mapped to GRCh38 by Bowtie2[38] (with option -x 2000) and reads with length less than 30 were filtered by SAMtools[39]. Duplications were removed by Picard (https://broadinstitute.github.io/picard/). The three replicated bam files at each time point were merged by SAMtools. MACS2[40] was applied on each merged bam file to call peaks (with option --nomodel --extsize 200 --shift 100). Peaks generated from each time point were merged by Diffbind[41] with default parameter to form the time course profile for ATAC-seq peaks. .

RNA-seq data processing

RNA-seq data were mapped to GRCh38 by STAR[13] with default parameters. Counts data for exons and introns were generated by DEXSeq[42]. Individual counts data from each time point were combined to form the time course gene expression data. Exons and introns counts data were summed to get the gene expression data for each gene at each time point, respectively. Genes with the sum of counts data across the six time points less than 10 were removed in each replicate. Only genes that have expressions in both replicates were kept. 18,162 genes remained after this processing.

Linking CHi-C, ATAC-seq and RNA-seq time course data

CHi-C time course data were linked to RNA-seq time course data with baits design specifying the mapping between baits and genes. ATAC-seq peaks residing at an otherEnd fragment were correlated with CHi-C interactions originating from that specific otherEnd fragment to different baits. Averaged data from replicates were used in correlation analysis. Pearson correlation coefficients between the connected CHi-C, gene and ATAC-seq time course data were calculated, respectively. Background random correlation tests were carried out by randomly picking up relevant time course data within the targeted dataset without any restrictions and calculating their Pearson correlation coefficients accordingly.

Gaussian process test for dynamic time course data

Time course data were fitted by a Gaussian process regression model[16] with a Radial Basis Function (RBF) kernel plus a white noise kernel (dynamic model) and a pure white noise kernel (static model), respectively. BIC was calculated for the dynamic model and flat model, respectively.

$$BIC = kln(n) - 2\ln(\hat{L})$$

where k is the number of parameters used in the specified model, n is the sample size and $\hat{L}$ is the maximized likelihood for the model. Models with smaller BIC are favoured for each time course profile. Those with smaller BICs in dynamic models were classified as time-varying. A more stringent χ^2 test with degree of freedom (df) of 1 was also applied to the Loglikelihood Ratio (LR) statistics, with $LR = -2\ln(\hat{L}_{\mathrm{RBF}} - \hat{L}_{\mathrm{STATIC}})$, where $\hat{L}_{\mathrm{RBF}}$ and $\hat{L}_{\mathrm{STATIC}}$ are the maximized likelihoods for the Gaussian process model and a static model, respectively. A p value of 0.05 was deemed significant.

ATAC-seq data clustering and MOTIF searching

A more inclusive threshold of LR>1 was applied to ATAC-seq peaks prior to clustering, which leaves 16% (12,215/74,583) ATAC-seq dynamical peaks, among which 9,680 were outside promoter regions ([+500bp,-1000bp] around genes). These ATAC-seq peaks were clustered using a Gaussian Process mixture model[43].

MOTIFs for each cluster were searched by findMotifGenome.pl (-mask –len 5,6,7,8,9,10,11,12 –size given) from HOMER with remaining peak data being used as background data.

Construction of 99% credible SNP sets for RA loci

We considered 80 loci attaining genome-wide significance for RA in the European ancestry component of the most recently published trans-ethnic GWAS meta-analysis[21], after excluding the MHC. For each locus, we calculated the reciprocal of an approximate Bayes' factor in favour of association for each SNP by Wakefield's approach[44], given by

$$\sqrt{\frac{V}{\mathrm{V}+\omega}}\exp\left[\frac{\omega\beta^2}{2\mathrm{V}(\mathrm{V}+\omega)}\right]_L$$

where β and V denote the estimated log odds ratio (log-OR) and corresponding variance from the European ancestry component of the meta-analysis. The parameter ω denotes the prior variance in allelic effects, taken here to be 0.04. The posterior probability of causality for the SNP is then obtained by dividing the Bayes' factor by the total of Bayes' factors for all SNPs across the locus. The 99% credible set for each locus was then constructed by: (i) ranking all SNPs according to their Bayes' factor; and (ii) including ranked SNPs until their cumulative posterior probability of causality attained or exceeded 0.99.

CRISPR activation using dCas9-p300

Cell lines

HEK293T cells (clontech) were cultured in high glucose-containing Dulbecco's modified Eagle's medium (DMEM; Sigma) supplemented with 10% FBS and 1% penicillin streptomycin at 37oC/5% C02 and kept below passage 15.

Generation of the dCas9-p300 cell line and delivery of guides

HEK293T cells were first transduced lentivirally with the pLV-dCas9-p300-p2A-PuroR expression vector (addgene #83889) and were selected with 2ug/ml of puromycin and grown for a week before being banked as a cell line. A second round of lentiviral transduction was done to introduce the guide RNA (gRNA) using the vector pLK05.sgRNA.EFS.GFP (addgene #57822) and cells were doubly selected using both a maintenance selection of puromycin and sorted for the top 60% cells expressing GFP.

Guide RNAs

All guide RNAs were cloned into the guide delivery vector pLK05.sgRNA.EFS.GFP (addgene #57822) and are listed Table 1. A negative control guide Scr2 (AACAGTCGCGTTTGCGACT) is a scrambled guide sequence for comparison of gene expression that is not expected to target any known genes in the genome. A positive control guide IL1RN (CATCAAGTCAGCCATCAGC) was included, this is a guide directed to the transcription start site of the promoter of the *IL1RN* gene that has been previously shown to increase expression of the IL1RN gene substantially[45]. For the COG6/FOX01 locus three guides (TGGGGACTATCTAGCTGCT; AGGGCCTTATAATGTAGT; AGTCATCCTGGAGCACAGAGG) were pooled simultaneously in equimolar amounts to target the active enhancer marked by H3K27ac in proximity to the lead GWAS variant rs7993214.

Lentivirus production

The day before transduction HEK293T cells were seeded at a density of 1E07 per transfer vector in 15cm plates in a volume of 20ml of DMEM 10% FBS without P/S. Each of the transfer vectors, together with packaging plasmids pmDLg/pRRE (#12251) and pRSV-REV (#12253) along with envelope plasmid pMD2.G (#12259), were combined to a total of 12ug at a ratio of 4:2:1:1, respectively in 2ml of serum free DMEM w/o phenol red.

PEI 1mg/ml was batch tested and added at a ratio of 6:1 PEI: DNA. The solution was briefly vortexed and incubated at room temperature for 15 minutes. Following this the solution was added dropwise to the cells. Flasks were rocked gently in a circular motion to distribute the precipitates, and then returned to the incubator. 24 hours later fresh growth medium was added of DMEM with 10% FBS and 1% P/S. The viral supernatant was collected 72 hours after transduction, cleared by centrifugation at 1500rpm for 5 minutes at 4^{o}C and then passed through a 0.45um pore PVDF Millex-HV (Millipore). Lentivirus was aliquoted and stored at -80^{o}C for future use.

Transduction of HEK293T p300 cell line with the gRNAs

300,000 dCas9-p300-HEK 293T cells were plated onto 6 well plates in triplicate for each gRNA. 24 hours later the medium was changed to DMEM 10% FBS without penicillin streptomycin. 1ml of each gRNA generated lentivirus was added to each well of 300,000 HEK293T cells cultured in DMEM supplemented with 10% FBS in triplicate. 24 hours later the medium was changed to DMEM containing 10% FBS and 1% penicillin streptomycin. Cells were grown up for 5 days and then sorted for the top 60% of cells expressing GFP using flow cytometry.

RNA extraction and qPCR

When confluent 2E06 cells were spun down at 400xg for 5 minutes and washed in PBS. RNA was extracted using the RNeasy mini kit (Qiagen) according to manufacturer's instructions and the genomic DNA removal step was included. 100ng of RNA for each sample was used in a single RNA-to-Ct reaction (Thermofisher) to assay gene expression. Taqman assays FOX01 (Hs00231106_m1), COG6 (Hs01037401_m1) and IL1RN (hs00893626_m1) were used alongside housekeeping genes YWHAZ (Hs01122445_g1) and TBP (hs00427620_m1) for normalisation.

Data analysis

Delta-delta CT analysis was carried out using the Scr2 generated dCas9-p300 HEK293T cells as the control and normalised against the YWHAZ and TBP housekeeping genes. The data was analysed in graph pad using one-way ANOVA.

Data availability

Raw sequencing data and processed counts data for ATAC-seq, RNA-seq, CHi-C and Hi-C that support the findings of this study have been deposited in National Center for Biotechnology Information's Gene Expression Omnibus and are accessible through GEO Series accession number GSE138767 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138767) with the secure token khsjquaejhyndcp while the paper is in review status.

Code availability

Scripts to reproduce the analysis and figures in this study are available on github https://github.com/jyangUK/IntegratingATAC-RNA-HiC/ .

Acknowledgements

We acknowledge support from Versus Arthritis (grant ref 21754) and MRC (grant ref MR/N00017X/1). P.M. is funded on Versus Arthritis Foundation fellowship (grant ref 21745). K.D. is funded on a Granville Hugh King foundation fellowship (grant ref 21146). The authors would like to acknowledge the assistance given by Research IT and the use of the Computational Shared Facility at the University of Manchester, UK. We acknowledge Servier Medical ART (https://smart.servier.com) for creating Fig. 1 in the main text.

Author Contributions

M.R., S.E. and P.F. conceived the project. A.M. performed ATAC-seq and RNA-seq assays, Hi-C and CHi-C experiments. P.M. designed the Hi-C and CHi-C experiments. J.Y. analysed all the data, P.Z. and P.M. additionally analysed the data. A.A and K.D. designed and performed the CRISPR experiment. A.P.M. additionally analysed the ATAC-seq data. J.Y, A.P.M., M.R. and S.E. wrote the manuscript.

Competing Interests statement

The authors declare no competing interests.

References

1. Fairfax, B. P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).
2. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–43 (2015).
3. 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).
4. Ye, C. J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science (80-.).* **345**, 1254665 (2014).
5. Mumbach, M. R. *et al.* Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.* **49**, 1602–1612 (2017).
6. Simeonov, D. R. *et al.* Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549**, 111–115 (2017).
7. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–80 (2014).
8. Waszak, S. M. *et al.* Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell* **162**, 1039–1050 (2015).
9. Gate, R. E. *et al.* Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* **50**, 1140 (2018).

10. Burren, O. S. *et al.* Chromosome contacts in activated T cells identify autoimmune disease candidate genes. *Genome Biol.* **18**, 165 (2017).

11. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384.e19 (2016).

12. 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).

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

14. Yang, T. *et al.* HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.* **27**, 1939–1949 (2017).

15. Posada, D. & Buckley, T. R. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Syst. Biol.* **53**, 793–808 (2004).

16. Yang, J., Penfold, C. A., Grant, M. R. & Rattray, M. Inferring the perturbation time from biological time course data. *Bioinformatics* **32**, 2956–2964 (2016).

17. Greenwald, W. W. *et al.* Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nat. Commun.* **10**, 1054 (2019).

18. Alasoo, K. *et al.* Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* **50**, 424–431 (2018).

19. Charles P. Fulco, Mathias Munschauer, Rockwell Anyoha, Glen Munson, Sharon R. Grossman, Elizabeth M. Perez, Michael Kane, Brian Cleary, Eric S. Lander, J. M. E. *et al.* Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science (80-.).* **6056**, 1–8 (2016).

20. Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **44**, 1336 (2012).

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

22. Myouzen, K. *et al.* Functional variants in NFKBIE and RTKN2 involved in activation of the NF-κB pathway are associated with rheumatoid arthritis in Japanese. *PLoS Genet.* **8**, e1002949 (2012).

23. Ludikhuize, J. *et al.* Inhibition of forkhead box class O family member transcription factors in rheumatoid synovial tissue. *Arthritis Rheum. Off. J. Am. Coll. Rheumatol.* **56**, 2180–2191 (2007).

24. Kuo, C.-C. & Lin, S.-C. Altered FOXO1 transcript levels in peripheral blood mononuclear cells of systemic lupus erythematosus and rheumatoid arthritis patients. *Mol. Med.* **13**, 561–566 (2007).

25. Grabiec, A. M. *et al.* JNK-dependent downregulation of FoxO1 is required to promote the survival of fibroblast-like synoviocytes in rheumatoid arthritis. *Ann. Rheum. Dis.* **74**, 1763–1771 (2015).

26. Liu, Y. & Yan, X. Eriodictyol inhibits survival and inflammatory responses and promotes apoptosis in rheumatoid arthritis fibroblast-like synoviocytes through AKT/FOXO1 signaling. *J. Cell. Biochem.* (2019).

27. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-.).* **326**, 289–293 (2009).

28. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–80 (2012).

29. Qu, Z. *et al.* Local Proliferation of Fibroblast-Like Synoviocytes Contributes to Synovial Hyperplasia. *Arthritis Rheum.* **37**, 212–220 (1994).

30. Pap, T. *et al.* Cooperation of Ras-and c-Myc–dependent pathways in regulating the growth and invasiveness of synovial fibroblasts in rheumatoid arthritis. *Arthritis Rheum. Off. J. Am. Coll. Rheumatol.* **50**, 2794–2802 (2004).

31. Belton, J.-M. *et al.* Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).

32. Van Berkum, N. L. *et al.* Hi-C: a method to study the three-dimensional architecture of genomes. *JoVE (Journal Vis. Exp.* e1869 (2010).

33. 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).

34. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21–29 (2015).

35. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, (2015).

36. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

37. Cairns, J. *et al.* CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.* **17**, 127 (2016).

38. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357 (2012).

39. Wysoker, A. *et al.* The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

40. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

41. Stark, R. & Brown, G. DiffBind: differential binding analysis of ChIP-Seq peak data. *R Packag. version* **100**, 3–4 (2011).

42. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).

43. Hensman, J., Lawrence, N. D. & Rattray, M. Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC Bioinformatics* **14**, 252 (2013).

44. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81**, 208–227 (2007).

45. Hilton, I. B. *et al.* Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* **33**, 510 (2015).

Supplementary information for "Simultaneous analysis of open chromatin, promoter interactions and gene expression in stimulated T cells implicates GWAS SNPs with causal genes"

Jing Yang[1§], Amanda McGovern[2§], Paul Martin[2,3], Kate Duffus[2], Peyman Zarrineh[1], Andrew P Morris[2], Antony Adamson[4], Peter Fraser[5]*, Magnus Rattray[1]* & Stephen Eyre[2,6]*

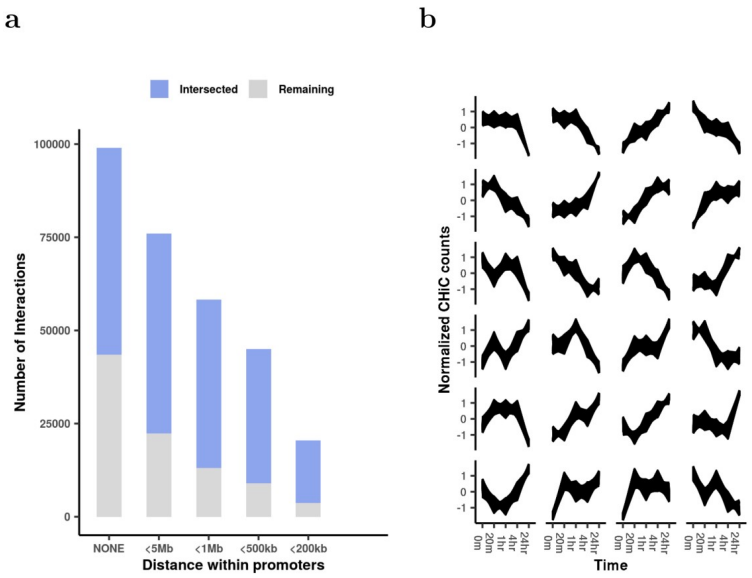
[§]Equal contribution, Joint first authors
*Equal contribution, Joint senior authors

1. Division of Informatics, Imaging & Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, M13 9PT, UK.
2. Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, Manchester Academic Health Science Centre, University of Manchester, Manchester, M13 9PT, UK.
3. Lydia Becker Institute of Immunology and Inflammation, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, M13 9PT, UK.
4. The Genome Editing Unit, Faculty of Biology, Medicine and Health, University of Manchester, Manchester M13 9PT, UK.
5. Department of Biological Science, Florida State University, 319 Stadium Drive, Tallahassee, FL 32306-4295, USA.
6. NIHR Manchester Biomedical Research Centre, Manchester University NHS Foundation Trust, Manchester, UK
   *email: Magnus Rattray (magnus.rattray@manchester.ac.uk), Stephen Eyre (steve.eyre@manchester.ac.uk)

1. Capture Hi-C

One pooled sample Capture Hi-C (CHi-C) dataset was collected and sequenced at times: 0 mins, 20 mins, 1 hr and 4 hrs. Another pooled sample CHi-C dataset was sequenced at the same times as well as an additional 24 hr time-point. The CHi-C sequence data were mapped to GRCh38 using HiCUP[1]. The maximum and minimum di-tag lengths were set to 800 and 150, respectively. CHiCAGO[2] was applied to each bam file with the CHiCAGO score set to 0. Counts data for each interaction were extracted from the .rds files generated by CHiCAGO. Interactions occurring at different time points (time 0 mins, 20 mins, 1 hr, 4 hrs and 24 hrs) were combined to create a complete set of all interactions and these were associated with counts for each time-point. Those interactions with at least one time point having CHiCAGO score over 5 were kept for further analysis. Bait-to-bait interactions are registered as two interactions with either side defined as "bait" or "otherEnd". 271,398 interactions were generated this way, among which 17,196 interactions were trans-interactions and 254,202 interactions were cis-

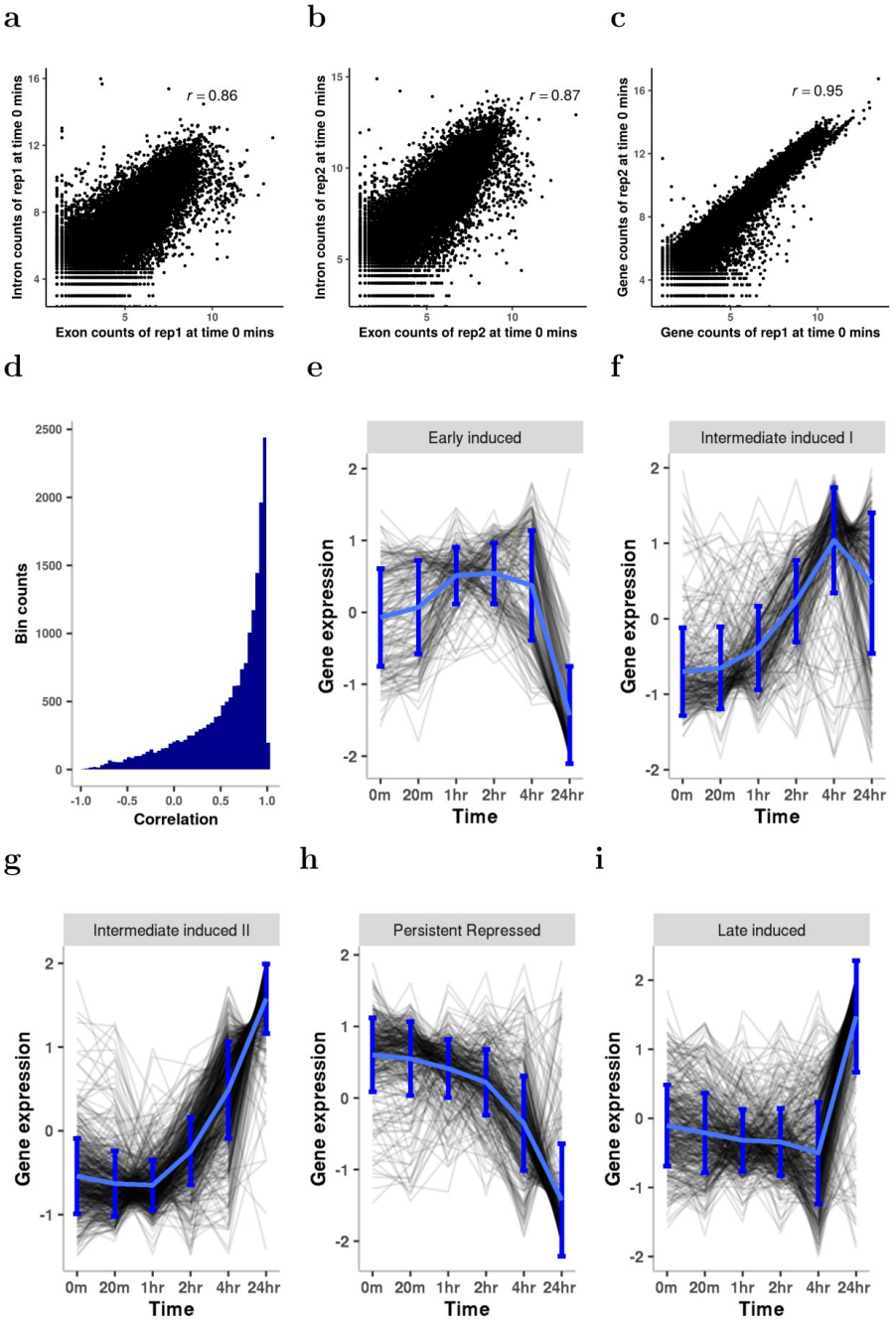
interactions. 6,888 baits and 121,656 otherEnds were involved in these interactions. CHi-C interactions occurring at either time 0 mins, time 4 hrs or both were extracted and compared to the interactions from Burren et al[3]. Comparisons for interactions originating from the two works under varied distance thresholds between bait and otherEnd fragments are shown in Supplementary Fig. 1a. It is clear that the closer the bait to otherEnd interactions are, the higher percentage of interactions are common between our experiments and experiments from Burren et al[3]. The top 24 clusters of the CHi-C interaction count time course data are illustrated in Supplementary Fig. 1b, showing varied and highly dynamic patterns of response.



Supplementary Fig. 1 **CHi-C quality check and time profile illustration. a**, Number of interactions interacted and remained between our work and work from Burren et al[3] under different distance thresholds: " NONE", "<5Mb", "<1Mb", "<500kb", "<200kb" representing range of distances to promoters considered. **b**, Largest 24 clusters of the time course profiles of CHi-C interactions within 5 Mb of promoters, using k-means clustering.

2. RNA-seq data

Two pooled sample RNA-seq time courses were collected at times 0 mins, 20 mins, 1 hr, 2 hrs, 4 hrs and 24 hrs. Reads were mapped to GRCh38 by STAR[4] with default parameters. Counts data for exons and introns were generated using DEXSeq[5]. Exon and intron counts for the same genes show good correlations (Supplementary Fig. 2a-b). Gene counts data were generated by adding up exon and intron counts data for the same genes, which also correlated well between replicates (Supplementary Fig. 2c). Individual counts data from each time point were combined to form the time course gene expression data. Genes with the sum of counts data across the six time points less than 10 were removed in each replicate. Counts data from each replicate were merged to form the time course gene expression data used for clustering and correlation with other datasets. Only genes that showed expression in both replicates were kept. 18,162 genes were remained after these processing steps. Gene expression data were normalized by DESeq2[6]. A histogram of the correlations between the two replicates across time is shown in Supplementary Fig. 2d, where 69% genes have correlations over 0.5. Gene



Supplementary Fig. 2 **Illustration of RNA-seq time profiles. a**, Scatter plot between natural logscaled exon and intron counts data of replicate 1 at time 0 mins. r is the Pearson correlation coefficient. **b**, Scatter plot between natural log scaled exon and intron counts data of replicate 2 at time 0 mins. **c**, Scatter plot between natural logscaled gene counts data of replicate 1 and replicate 2 at time 0 mins. **d**, Histogram of the correlations between the two RNA-seq replicates used in this study. **e-i**, Time course profiles of gene sets as categorized in Ye et al[7]. Grey lines represent normalized gene expression and blue lines represent the mean of the data in each dataset. Errorbars are ± std of the data in each plot.

expression data were compared to the data from Ye et al[7], where similar experiments were carried out up to time 72 hrs. Time course profiles of the gene expression of the five categorized gene sets from Ye et al[7], including 'Early induced', 'Intermediate induced I', 'Intermediate induced II', 'Persistent repressed' and "Late induced", are shown in Supplementary Fig. 2e-i. Our data show similar patterns to those in Ye et al[7].

3. ATAC-seq data

Three replicated ATAC-seq time series were collected at times 0 mins, 20 mins, 1 hr, 2 hrs, 4 hrs and 24 hrs and mapped to GRCh38 by Bowtie2[8] (with option -x 2000) with reads of length less than 30 bp filtered out using SAMtools[9]. Duplicates were removed by Picard (https://broadinstitute.github.io/picard/). The three replicated bam files at each time point were merged by SAMtools. Macs2[10] was applied on each merged bam file to call peaks (with option --nomodel --extsize 200 --shift 100) . Supplementary Table 1 is the number of peaks identified for each time point.

Supplementary Table 1. **Number of peaks called by MACS2 on each merged bam file at six time points.**

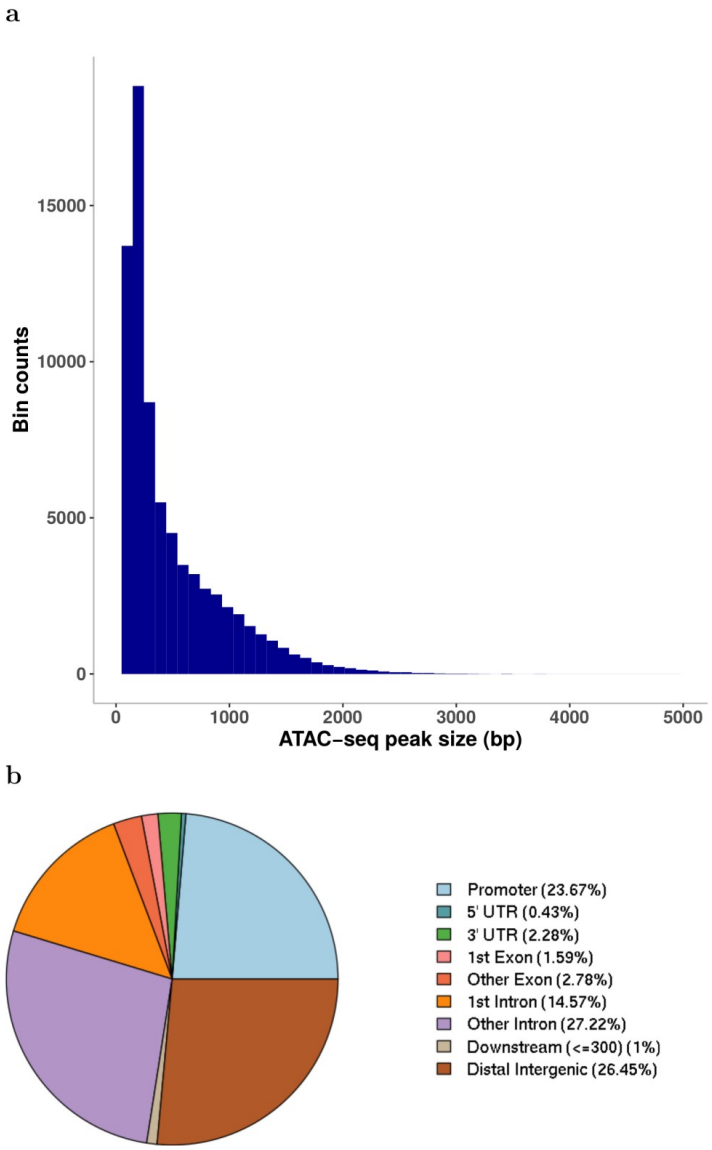
Time Points	0 mins	20 mins	1 hr	2 hrs	4 hrs	24 hrs
Number of peaks	30,403	27,793	47,823	22,664	43,537	30,593

To align ATAC-seq datasets across time, the peaks generated from each time point were merged by Diffbind[11] (with option minOverlap=1) and 76,359 peaks were obtained in the end with an average size of 488.32 bp. The distribution of the peak sizes are shown in Supplementary Fig. 3a.

ATAC-seq peaks were annotated using ChIPpeakAnno[12] (Supplementary Fig 3b). Peaks that lie in the Promoter region (22.9%) were removed in the downstream analysis in order to focus on enhancer-associated peaks. Supplementary Table 2 compares the peaks from our data and those peaks from Gate et al[13]. Due to different experiment setup, there are some discrepancies between the peaks from these two sources. However, the majority of peaks from our data are within the peaks from Gate et al[13].

Supplementary Table 2 **Comparison of peaks located in our data and peak data from Gate et al[13]**.

	0 min vs 0 min	24 hrs vs 48 hrs	merged vs merged
Number of intersecting peaks	21,549	22,911	39,021
Number of peaks only in our data	8,854	7,682	35,740
Number of peaks only in [13]	14,926	29,182	24,279



Supplementary Fig 3 **ATAC-seq peak size histogram and peak annotations. a,** Histogram of ATAC-seq peak sizes; **b**, Piechart of the annotations of the ATAC-seq peak data.

4. Hi-C data

One pooled sample Hi-C time course was collected at times 0 mins, 20 mins, 1 hr and 4 hrs. Another pooled sample Hi-C time course was collected at these times and an additional time of 24 hrs. Hi-C data were mapped to GRCh38 by HiCUP[1] and then converted to HOMER[14] format by scripts provided in HiCUP. The maximum and minimum di-tag lengths were set to 800 and 150, respectively. HOMER was applied the mapped Hi-C data (analyzeHiC –res 40000 -balance) and 1,230 distinct TADs were discovered across all 5 time points. Supplementary Table 3 shows percentages of the interactions of TADs across different time points within and between replicates with reciprocal 90% region overlap.

Supplementary Table 3 **Percentages of the intersection of TADS across and within replicates.** Data shown at (Tm,Tn) represents the percentage of the number of intersected TADs between data taken at time Tm and time Tn over the number of TADs in the data at Tn.

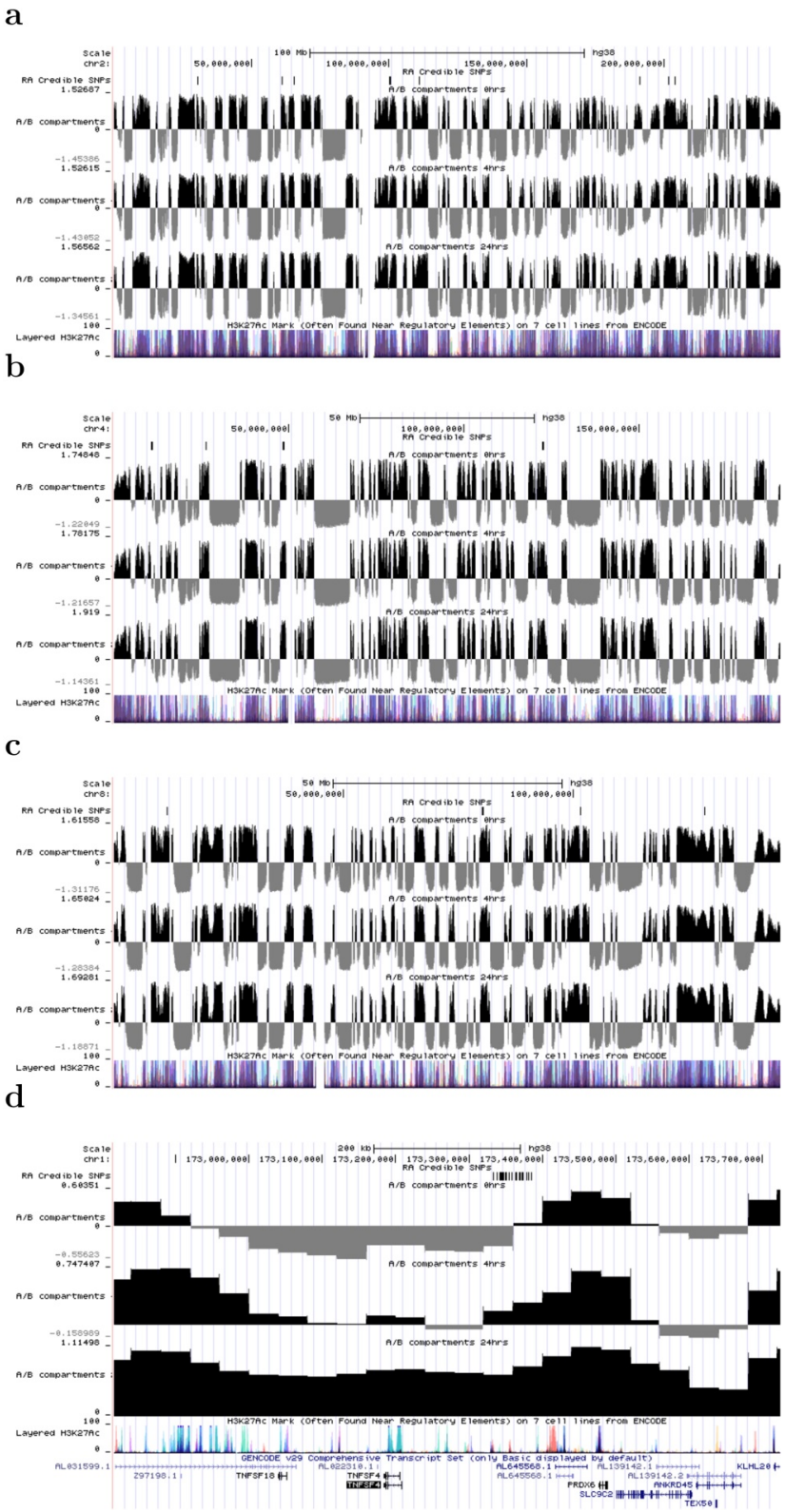
	T01	T02	T201	T202	T1H1	T1H2	T4H1	T4H2	T24H1
T01	1	84.2%	85.8%	85.5%	83.0%	84.4%	80.8%	80.5%	74.0%
T02	83.1%	1	81.8%	85.0%	79.3%	83.0%	78.1%	79.2%	72.9%
T201	84.0%	81.2%	1	87.3%	86.4%	84.1%	81.9%	79.4%	72.9%
T202	82.3%	83.0%	85.4%	1	81.5%	84.6%	79.2%	78.5%	72.2%
T1H1	81.5%	77.9%	86.7%	83.5%	1	82.5%	81.7%	77.7%	71.6%
T1H2	82.3%	81.2%	83.3%	85.8%	82.2%	1	81.2%	82.6%	73.1%
T4H1	79.0%	76.7%	81.4%	79.9%	81.1%	80.9%	1	85.4%	75.5%
T4H2	78.3%	77.2%	78.3%	79.5%	76.9%	81.8%	84.9%	1	75.7%
T24H1	69.2%	69.5%	69.9%	70.6%	68.6%	71.2%	74.2%	73.9%	1

5. A/B compartments

A/B compartments were found by HOMER with command runPCA.pl (with option –res 40000) followed by findHICCompartments.pl to find A compartments or findHICCompartments.pl –opp to find B compartments, respectively. 1,136 A compartments and 1,266 B compartments were discovered. Supplementary Tables 4 and 5 shows percentages of the A and B compartment across different time points within and between replicates with reciprocal 90% coverage. Chromosome positions of A/B compartments were converted to bed files and visualised in the UCSC browser[15] (Supplementary Figure 4).

Supplementary Table 4 **Percentages of the intersection of compartment As across and within replicates.** Data shown at (Tm,Tn) represents the percentage of the intersected compartment As between data taken at time Tm and time Tn over the compartment As in the data at Tn.

	T01	T02	T201	T202	T1H1	T1H2	T4H1	T4H2	T24H1
T01	1	97.3%	98.5%	97.7%	98.5%	97.8%	98.2%	97.3%	97.7%
T02	96.8%	1	97.5%	97.2%	96.8%	97.5%	97.0%	97.1%	97.4%
T201	97.5%	96.9%	1	96.6%	98.1%	97.6%	98.2%	96.4%	97.5%
T202	97.6%	97.6%	97.5%	1	98.0%	98.2%	97.8%	97.9%	97.2%
T1H1	97.4%	96.4%	97.8%	97.0%	1	97.7%	98.0%	97.2%	97.0%
T1H2	95.2%	95.7%	96.3%	95.8%	96.5%	1	97.1%	96.7%	96.8%
T4H1	95.5%	95.4%	96.8%	95.6%	96.7%	97.1%	1	96.3%	97.8%
T4H2	95.5%	95.8%	95.6%	96.3%	96.8%	97.4%	97.3%	1	97.6%
T24H1	89.7%	89.5%	90.2%	89.7%	89.6%	90.6%	91.8%	90.8%	1



Supplementary Fig 4 **UCSC genome browser plots of chromatin structure, overlaid with epigenomic features and RA associated variants. a-c**, Exemplar plots from chromosomes 2, 4 and 8 of A/B compartments (A

compartments black; B compartments grey) called in at 3 time points (0, 4hrs and 24hrs) in duplicate T cells, compared to public layered H3K27ac. **d**, A/B compartments on Chr1 around the TNFSF18 and TNFSF4 gene region. RA associated variants go from a largely inactive B region at 0hrs, through to an active A compartment at 24hrs.

Supplementary Table 5 **Percentages of the intersection of compartment Bs across and within replicates**. Data shown at (Tm,Tn) represents the percentage of the intersected compartment Bs between data taken at time Tm and time Tn over the compartment Bs in the data at Tn.

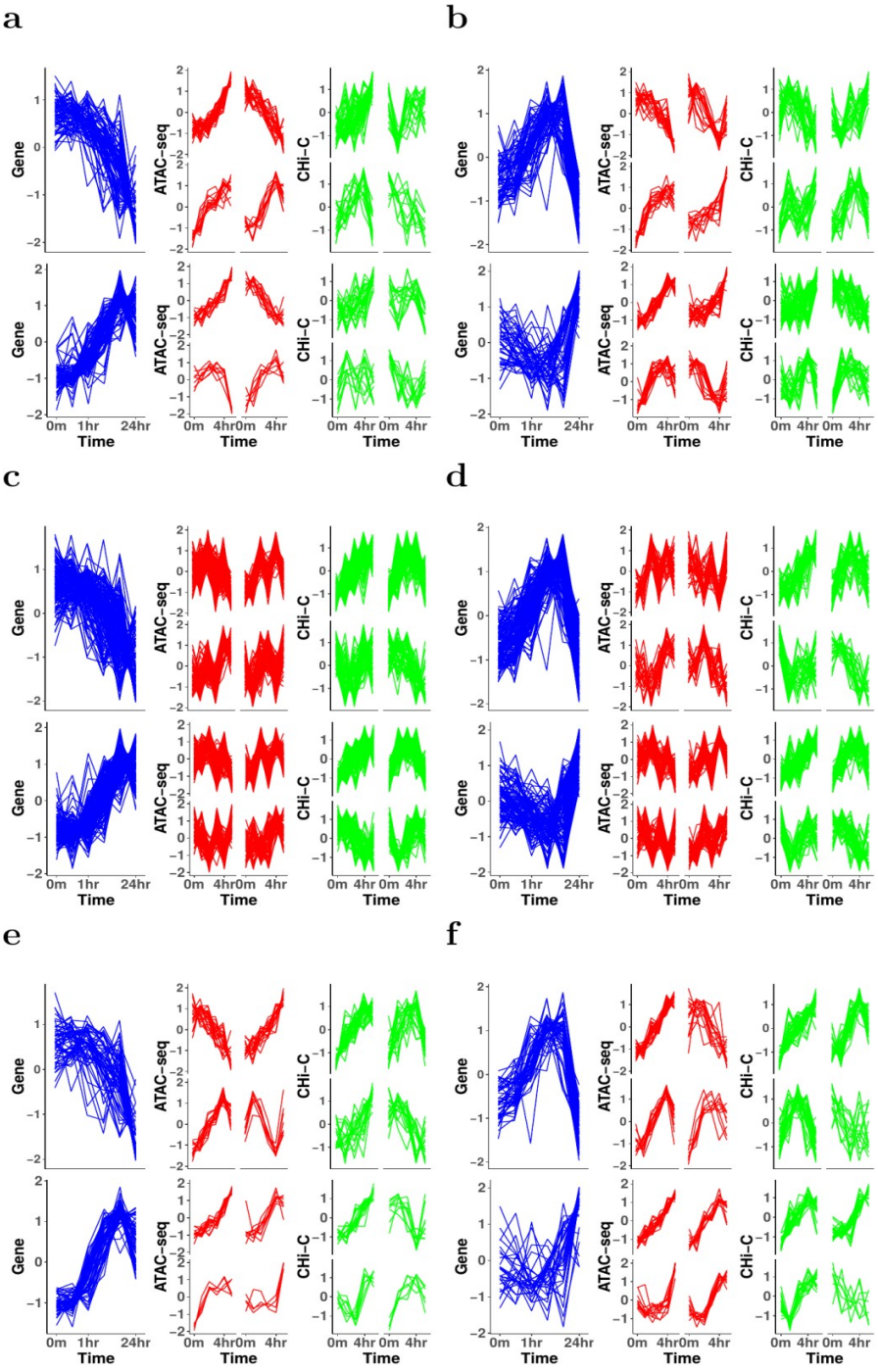
	T01	T02	T201	T202	T1H1	T1H2	T4H1	T4H2	T24H1
T01	1	95.9%	96.8%	97.8%	97.6%	96.6%	94.9%	96.2%	92.4%
T02	94.2%	1	93.6%	96.0%	94.9%	95.9%	92.6%	94.2%	89.8%
T201	97.6%	96.2%	1	98.0%	97.6%	97.0%	95.7%	96.5%	93.1%
T202	95.5%	96.1%	94.9%	1	95.9%	96.2%	93.3%	95.5%	90.6%
T1H1	96.9%	96.1%	95.8%	97.4%	1	97.2%	95.0%	96.5%	91.4%
T1H2	94.7%	96.1%	94.4%	96.3%	96.3%	1	93.7%	95.9%	90.6%
T4H1	96.0%	94.7%	96.0%	95.9%	97.1%	96.2%	1	97.2%	94.2%
T4H2	94.7%	94.8%	94.1%	96.0%	96.2%	96.4%	94.6%	1	92.1%
T24H1	92.6%	92.0%	92.3%	92.7%	92.8%	92.9%	93.8%	94.0%	1

6. ATAC-seq data clustering and MOTIF search

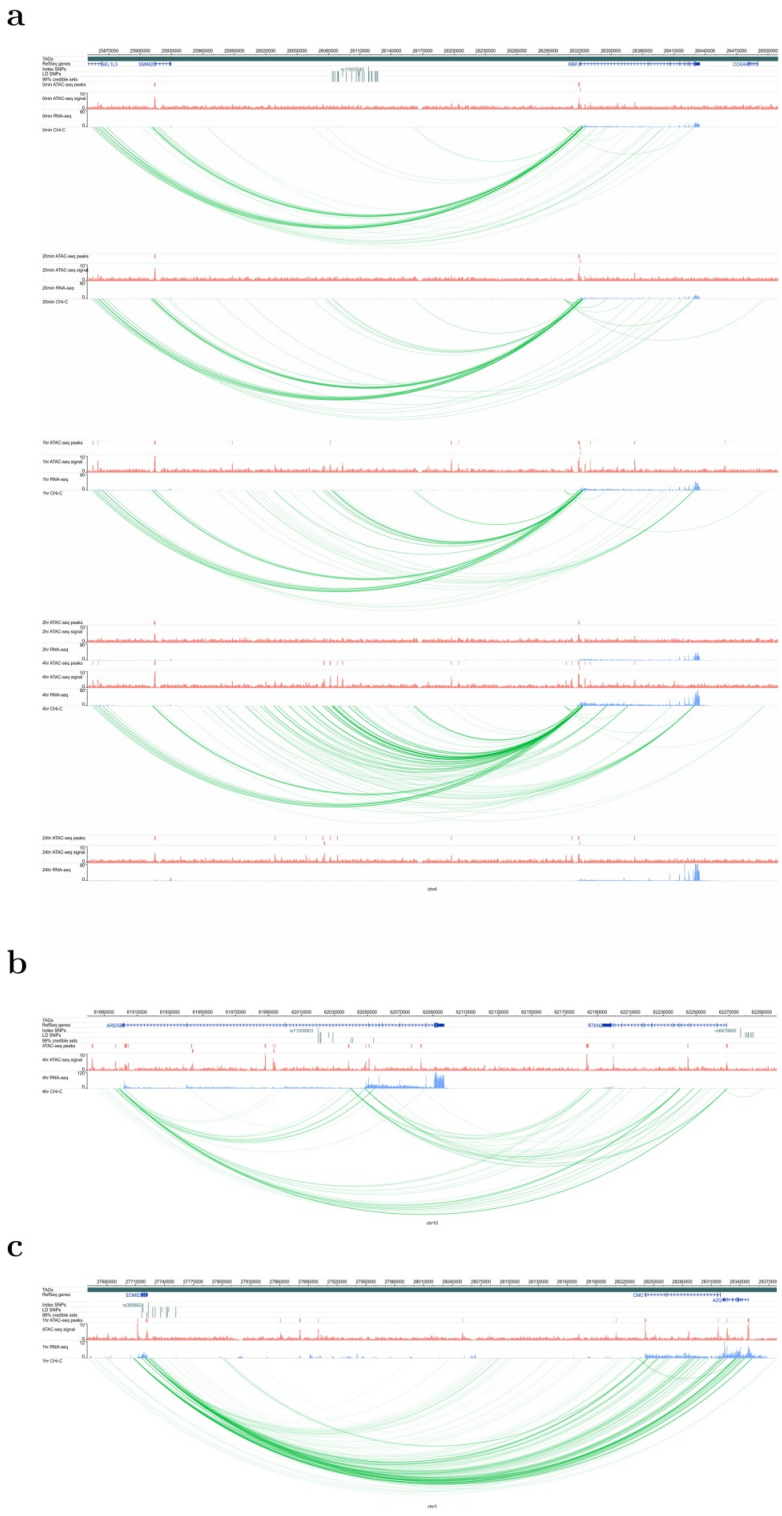
ATAC-seq data residing outside promoter regions, with loglikelihood ratio (LR, see main paper Methods) between a dynamic and static model over 1, were clustered using a hierarchical Gaussian Process mixture model[16]. MOTIFs for these clusters were searched by findMotifsGenome.pl from HOMER (-len 5,6,7,8,9,10,11,12,13 –size given) with ATAC-seq data with $BIC_{\mathrm{RBF}} > BIC_{\mathrm{NOISE}}$ (static peaks) as background data.

7. Correlations between ATAC-seq, CHi-C and gene

For ATAC-seq peaks sitting inside otherEnd fragments, the correlations between ATAC-seq time course, CHi-C time course and RNA-seq time course related to interacted baits are examined. Clusters between genes, CHi-C and ATAC-seq within the distance range of 200 kb of promoters under dynamical or stationary scenarios are shown in Supplementary Fig. 5a-f, respectively.



Supplementary Fig. 5 **Clusters of gene expression, ATAC-seq counts and CHi-C data within 200 kb of promoters**. **a,b**: Clusters with only ATAC-seq peaks are dynamical; **c,d**, Clusters with only CHi-C data are dynamical; **e,f**: Clusters with both ATAC-seq and CHi-C data are dynamical.



Supplementary Fig. 6 **Illustration of genomic interaction activities around three RA associated loci**. **Screenshots of the SNPs (dark green), ATAC-seq peaks (red), RNA-seq (blue) and CHi-C interactions (green). a**, RBPJ loci, demonstrating how both interactions between the associated SNPs/gene promoter, and ATAC-seq intensity

increase in magnitude over time (0, 20mins, 1hr, 2hr, 24hr, top to bottom) **b**, ARID5B_RTKN2 loci, demonstrating strong interactions between the region intronic of ARID5B and RTKN2. **c**, EOMES_AZI2 loci demonstrating strong interactions between the region intronic of EOMES and AZI2.

Supplementary Table 6 **ALL RA loci.** Table of all ATAC-seq peaks containing a SNP in the 99% credible set for RA, the promoters they interact with and the correlation between ATAC-seq activity, interaction strength and gene expression. Included as excel spreadsheet.

Supplementary Table 7 **RA loci with eQTL evidence.** Table of all ATAC-seq peaks containing a SNP in the 99% credible set for RA, the promoters they interact with and the eQTL evidence for the interaction. Included as excel spreadsheet.

References

1. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, (2015).
2. Cairns, J. *et al.* CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.* **17**, 127 (2016).
3. Burren, O. S. *et al.* Chromosome contacts in activated T cells identify autoimmune disease candidate genes. *Genome Biol.* **18**, 165 (2017).
4. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
5. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).
6. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
7. Ye, C. J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science (80-.).* **345**, 1254665 (2014).
8. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357 (2012).
9. Wysoker, A. *et al.* The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
10. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
11. Stark, R. & Brown, G. DiffBind: differential binding analysis of ChIP-Seq peak data. *R Packag. version* **100**, 3–4 (2011).
12. Zhu, L. J. Integrative analysis of ChIP-chip and ChIP-seq dataset. in *Tiling Arrays* 105–124 (Springer, 2013).
13. Gate, R. E. *et al.* Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* **50**, 1140 (2018).
14. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
15. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
16. Hensman, J., Lawrence, N. D. & Rattray, M. Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC Bioinformatics* **14**, 252 (2013).