nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-18180-7 OPEN

Analysis of chromatin organization and gene expression in T cells identifies functional genes for rheumatoid arthritis

Jing Yang[1,7], Amanda McGovern[2,7], Paul Martin[2,3], Kate Duffus[2], Xiangyu Ge[2], Peyman Zarrineh[1], Andrew P. Morris[2], Antony Adamson[4], Peter Fraser[5,8], Magnus Rattray[1,8✉] & Stephen Eyre[2,6,8✉]

Genome-wide association studies have identified genetic variation contributing to complex disease risk. However, assigning causal genes and mechanisms has been more challenging because disease-associated variants are often found in distal regulatory regions with cell-type specific behaviours. Here, we collect ATAC-seq, Hi-C, Capture Hi-C and nuclear RNA-seq data in stimulated CD4+ T cells over 24 h, to identify functional enhancers regulating gene expression. We characterise changes in DNA interaction and activity dynamics that correlate with changes in gene expression, and find that the strongest correlations are observed within 200 kb of promoters. Using rheumatoid arthritis as an example of T cell mediated disease, we demonstrate interactions of expression quantitative trait loci with target genes, and confirm assigned genes or show complex interactions for 20% of disease associated loci, including *FOXO1*, which we confirm using CRISPR/Cas9.

[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, Tallahassee, FL 32306, USA. [6] NIHR Manchester Biomedical Research Centre, Manchester University NHS Foundation Trust, Manchester, UK. [7] These authors contributed equally: Jing Yang, Amanda McGovern. [8] These authors jointly supervised this work: Peter Fraser, Magnus Rattray, Stephen Eyre. ✉email: magnus.rattray@manchester.ac.uk; steve.eyre@manchester.ac.uk

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, 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 either to determine physical interactions, with methods such as Hi-C[7], use quantitative trait analysis[4] or examine 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 expression quantitative trait locus (eQTL) studies to implicate the target gene. The remainder is 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 CD4+ 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 RA, and validate long range interactions with CRISPR/Cas9, to assign SNPs, genes and direction of effect to GWAS loci for this T cell-driven disease.

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 min, 20 min, 1 h, 2 h, 4 h and 24 h after stimulation with anti-CD3/anti-CD28. A total of 76,359 ATAC-seq peaks were obtained; 6287 peaks unique to unstimulated cells and 45,869 only appearing after stimulation. 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 of promoters.

Gene expression data demonstrated good correlation between replicates (Supplementary Fig. 1a–d). From an initial FACS analysis (Supplementary Table 1 and Supplementary Fig. 2a, b) and comparison of RNA-seq data, we determined the initial T-cell purity of the samples to be reproducibly high, around 94% CD4+, displaying similar amounts of CD4+ memory and CD4+ naive (Supplementary Fig. 3a). To assess the amount of stimulation, we explored markers of naivety, activation and cytokine expression and confirmed a similar degree of stimulation across all the samples (Supplementary Fig. 3b).

ATAC-seq peaks also demonstrated good quality and correlation between replicates (Supplementary Fig. 4a–e) and enrichment for both marks of enhancer activity (ChromHMM, Supplementary Fig. 5a–d) and CTCF sites across all time points (Supplementary Fig. 5e), with an increased amount of CTCF occupancy in ATAC-seq peaks at topologically associating domain (TAD) boundaries, as expected (Supplementary Fig. 5f).

Data consistency with previous studies. Comparison of ATAC-seq peaks from CD4+ baseline (unstimulated) and 48 h after stimulation peaks from a similar data set[9] revealed strong concordance (Supplementary Fig. 4f): 71% of peaks (21,549/30,403) from unstimulated CD4+ T cells overlapped in both data sets[9], while 75% of peaks (22,911/30,593) from stimulated CD4+ T cells (24 vs. 48 h post stimulation) overlapped[9]. We also observed a similar magnitude of increase in the number of ATAC-seq peaks for merged unstimulated and stimulated data.

Comparison of our CHi-C data with published data sets from the same cell type and stimulation[10], both unstimulated or stimulated for 4 h, demonstrated good consistency (Supplementary Fig. 6a). When restricting all the interactions from both unstimulated and stimulated to those that share the same baits, we found 57% of interactions (27,794/48,570) 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 (26,836/36,698) and 87% within 200 kb of promoters (8,367/9,627). This strongly suggests that the interactions between promoters and active enhancers within 200 kb are consistent, robust and reproducible between studies. We found 22,126 genes with evidence of CD4+ T-cell expression for at least one time point in our RNA-seq data. 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 Fig. 1e–i).

Chromatin conformation dynamics. It is well established that gene expression changes with time after stimulation in CD4+ T 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. 1e–i). However, it is less well established how chromatin structure changes post stimulation, in the form of A/B compartments, TADs and individual interactions, or how enhancer activity and open chromatin change over time.

Based on Hi-C matrices with resolutions of 40 kb, 1230 merged TADs were recovered across all the time points with an average size of 983.2 kb. Consistency of TADS across time points was measured by the fraction of overlapping TADS (90% reciprocal), over the number of TADs at each specified time point (Supplementary Fig. 7a). On average between time point 0 and 4 h 84% of TADS intersect. When adding the 24 h TAD data, the mean intersection reduces to 74%, illustrating more substantial dynamic changes in TADs over longer times. Figure 2b shows the stratum-adjusted correlation coefficient (SCC) between Hi-C data sets[14] and shows a slight but significant reduction in correlation as the time separation of experiments increases, consistent with our observations regarding TADs. SCC between replicates and within replicates does not show clear differences, implying that the changes in correlations are not due to batch effects.

We recovered 1136 A compartments and 1266 B compartments merged across the time course data, with the maximum compartment sizes being 39.5 and 34.5 Mb, respectively. Similar to TADs, the consistency of compartments across time points was measured by the fraction of overlapping compartments (90% reciprocal) over the number of compartments at each specified time point (Supplementary Fig. 7b, c). The same consistency was

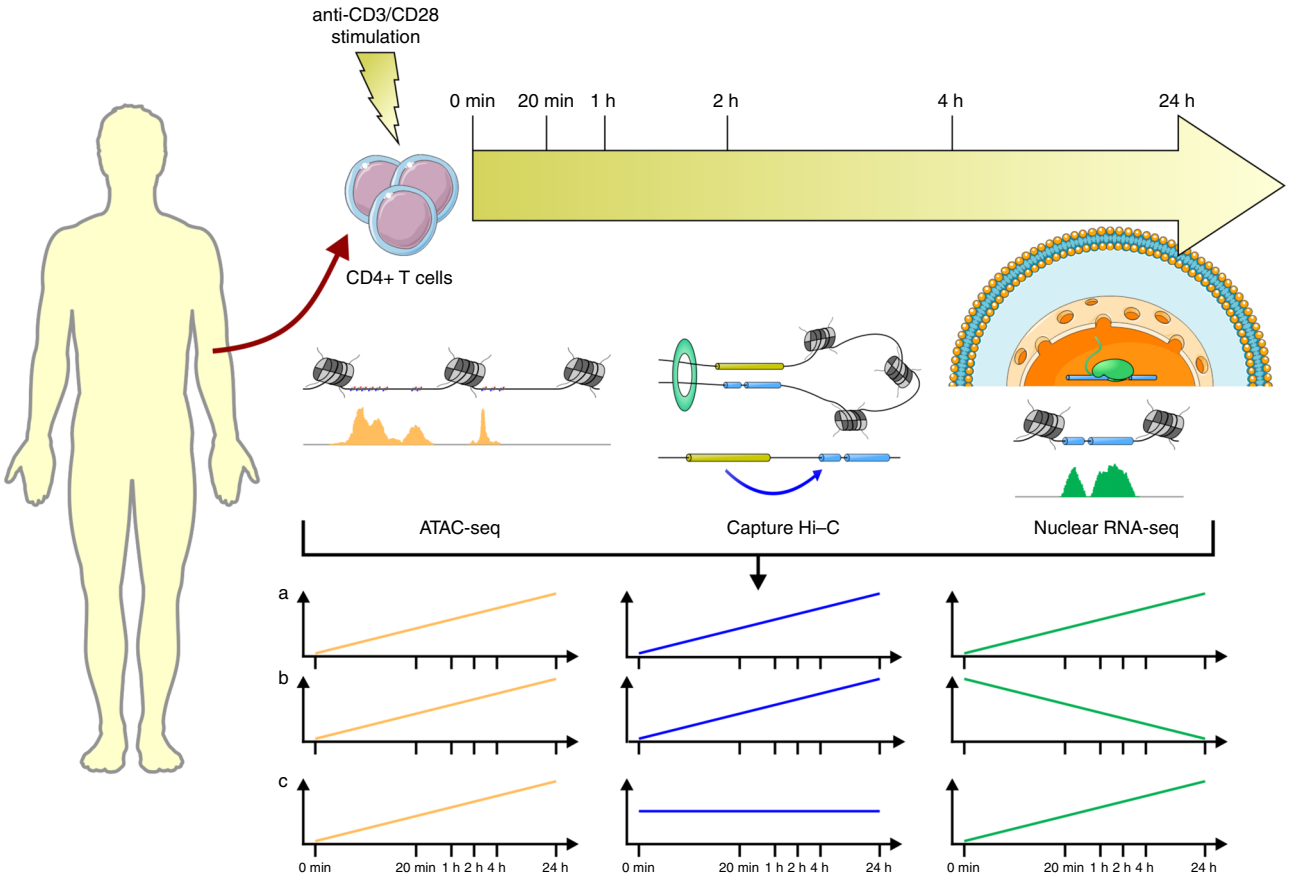

Fig. 1 Schematic of the study design. ATAC-seq, CHi-C and nuclear RNA-seq experiments were carried out for unstimulated and stimulated CD4+ T-cell samples at time 0 min, 20 min, 1 h, 2 h, 4 h and 24 h. Time course profiles were created by aligning features (ATAC-seq peaks and CHi-C interactions) across time and counting reads supporting each feature at each time point. Courtesy of Servier Medical Art licensed under a Creative Commons Attribution 3.0 Unported license. https://smart.servier.com.

observed for compartments A and B. Figure 2c shows the correlation between A/B compartment allocations, demonstrating a slight but significant reduction between experiments as time separation increases. Consistent B compartments and compartments that changed over time from A to B were found to be enriched for lamina associated domains when compared to ones from a resting and activated T-cell line[15] and genes with repressed expression (Supplementary Fig. 8).

These results are broadly consistent with other studies, demonstrating how the higher chromatin conformation states, in the form of A/B compartments and TADS, are largely invariant between cell types[11]. Here, we demonstrate similar levels of consistency in a single cell-type post stimulation (Supplementary Fig. 9a–c).

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)[16] and a χ^2 test to compare a dynamic (Gaussian process) model to a static model[17] for CHi-C interaction count 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% (4837/20,224) are within 200 kb of promoters.

Open chromatin dynamics. 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% (7852/74,583) of ATAC-seq peaks with evidence of change over 24 h ($BIC_{dynamic} < BIC_{static}$) with 2780 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. 3b). Mapping of 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, whilst 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 h. These findings match those reported in a previous study of ATAC-seq data in CD4+ T cells stimulated with anti-CD3/anti-CD28[9]. There it was demonstrated that the 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. ATAC-seq peaks overlapping H3K27ac marks and without CTCF binding are more likely to be gained over time (blue to red, Supplementary Fig. 10a), a pattern not seen in ATAC-seq peaks bound by CTCF without H3K27ac, although the distribution of interactions from these two classes of peaks (gained or lost) remains similar, with no bias seen in either category of peak (Supplementary Fig. 10b). We found evidence that the increase in ATAC-seq activity over time (e.g., Fig. 3b, cluster 1) corresponded to an increase in interactions (Supplementary Fig. 10d),

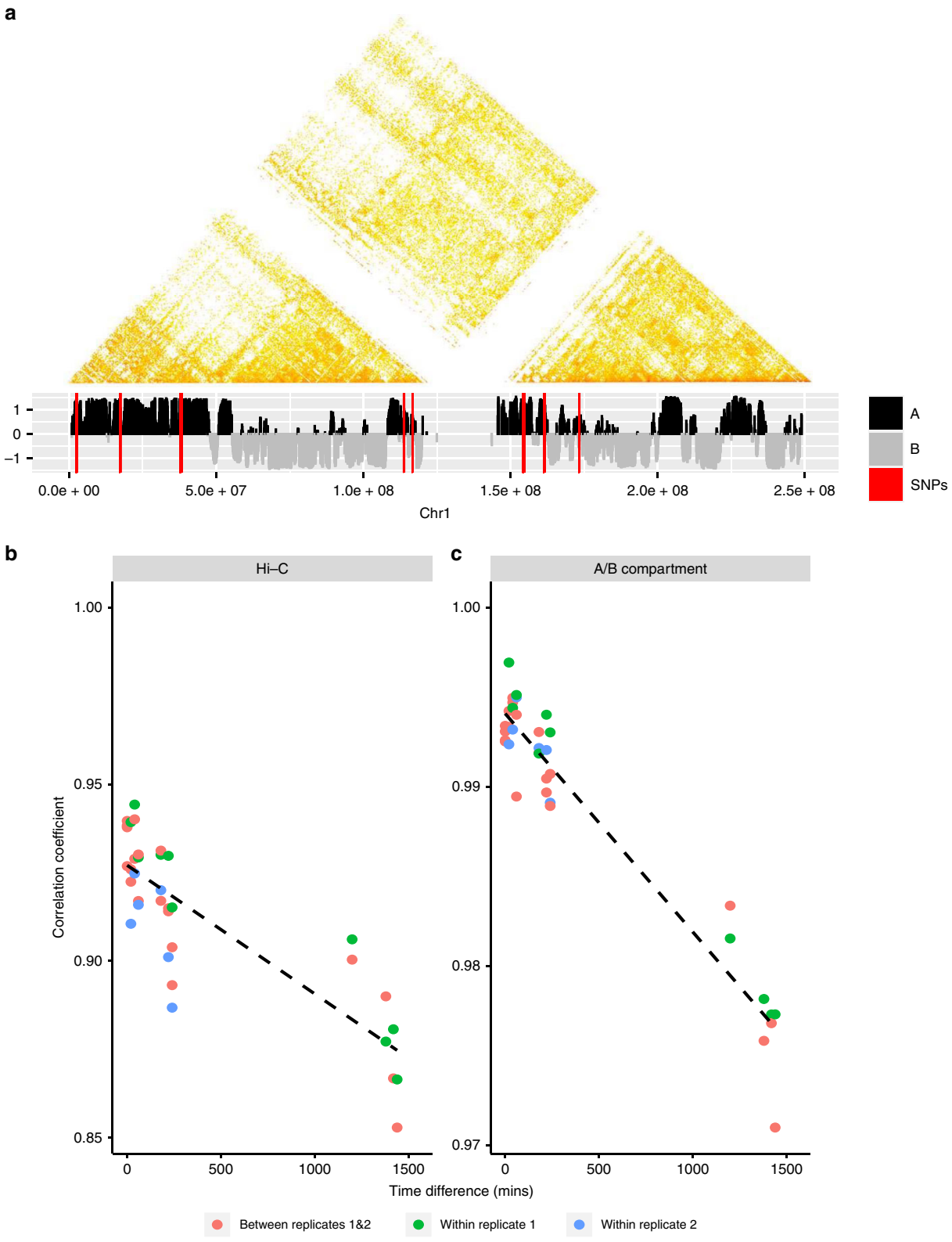

Fig. 2 Illustration of Hi-C dynamics. a Hi-C interaction matrix (100 kb resolution) of replicate 1 for chr1 at time 0 min (upper) with corresponding A/B compartments (lower), where red lines represent positions of SNPs. **b** Correlation changes between Hi-C data with respect to differences between times, where the dashed line is the fitted linear line for the correlation coefficients. **c** Correlation changes of A/B compartments with respect to the differences between times, where the dashed line shares the same information as conveyed in the plot b.

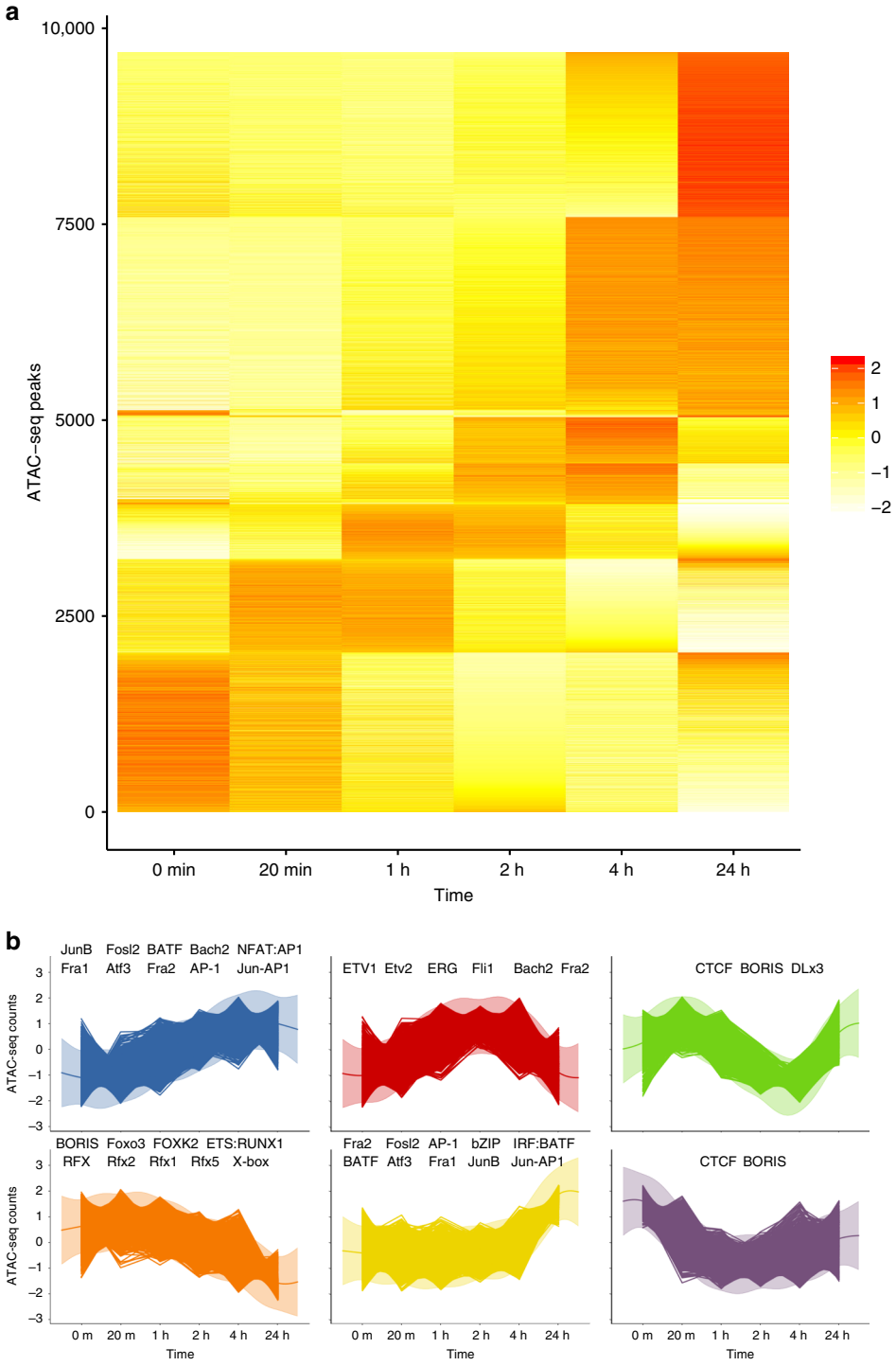

Fig. 3 Illustration of ATAC-seq time course profile dynamics. a Heatmap of ATAC-seq counts data for peaks showing evidence of temporal dynamics. **b** Clustering of ATAC-seq time course data using a Gaussian process mixture model. Significantly enriched DNA-binding MOTIFs in each peak (using static peaks are background) are labelled in each cluster.

particularly DNA locations containing the JunB transcription factor.

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[18]. We wanted to determine the nature of this relationship in our data, from a single, activated cell type. We looked for otherEnd baits that contained ATAC-seq peaks, and linked these ATAC-seq peaks with CHi-C interactions to promoter baits and associated genes. Using this approach, we formed 44,113 links. Pearson correlation coefficients for paired time course data between ATAC-seq and CHi-C, ATAC-seq and gene, and gene

and CHi-C were calculated. We used a link 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 data sets, particularly an enrichment for high positive correlations within 200 kb of promoters (Fig. 4a), suggesting that functional, interactive correlations are most common within 'contact domains'. This observation is 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.

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[18].

Previous studies have indicated how using (1) eQTL data, ~50% of ATAC-seq peaks are already active/poised before influencing gene expression[19], (2) HiChIP data, expression can be correlated with either H3K27ac or interactions[5], and (3) empirical ranking of enhancers by CRISPR corresponds most strongly when combining terms for interaction and activity[20]. Taken together, these observations suggest that 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. 11): ~8% (577/6,892) of links were associated with dynamic ATAC-seq peaks only (Supplementary Fig. 11a, b), 29% (2014/6892) were associated with dynamic CHi-C interactions only (Supplementary Fig. 11c, d) and 5% (371/6892) were associated with dynamics in both (Supplementary Fig. 11e, 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 have a larger effect on gene expression.

Prioritisation of causal genes in GWAS loci for autoimmune disease. We identified 312 loci which contained an autoimmune disease lead SNP that was highly correlated with a variant ($r^2 > 0.8$) located within an ATAC-seq peak, which was itself interacting and highly correlated with the expression of a gene (Supplementary Data 1). These data highlight potential causal SNP, ATAC-seq and gene relationships for autoimmune disease risk in CD4+ T cells, confirming previous reports or providing new evidence for genes such as *PRKCQ, CD44, ETS1* and *ARID5B*.

We next considered 80 of the 100 loci previously associated with RA that attain genome-wide significance in European ancestry GWAS meta-analysis[21,22]. For each locus, we constructed a 99% credible set of SNPs that accounted for 99% of the probability of containing the causal variant. We found that 97% (2131/2192) of RA-associated variants from our 99% credible SNP sets lie within A compartments across all time points while only 28 (1%) lie consistently within B compartments after stimulation. Fifteen credible set SNPs were found in regions that change between A and B over time. These included RA credible set SNPs on chromosome 1, proximal to the *TNFSF4* and *TNFSF18* genes, that were initially contained within an inactive B compartment at 20 min and were then found in an A compartment at 4 h (Supplementary Fig. 9d).

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. Genome-wide, RA credible set SNPs signals were enriched (5–30 fold) in open regions of chromatin, as expected. The strongest enrichment was observed at 4 and 24 h post stimulation (Supplementary Fig. 12a), where SNPs mapping to open regions of chromatin explained up to 30% of the heritability of RA (Supplementary Fig. 12b). We found that 43/80 GWAS loci contained 67 ATAC-seq peaks with at least one RA credible set SNP SNP (98 SNPs in total) that interacted and correlated with the expression of 168 genes, an average of 2.5 genes per peak (Supplementary Fig. 13 and Supplementary Data 2).

These data have the ability to limit the number of putative causal SNPs/ATAC-seq peaks for each locus. The 43 loci with ATAC-seq peaks that contain RA GWAS-associated SNPs have 5527 credible SNPs, which reduce down to 98 SNPs in 67 ATAC-seq peaks that interact with and are correlated with the expression of 168 genes. 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. 14a).

For six loci, interactions between 30 and 220 kb implicate either a single ATAC-seq peak (*CD5, PXK, TCTE1, CDK6, TPD52*) or two ATAC-seq peaks (*IL6ST*) that contain an RA credible set SNP, interacting with the target eQTL gene. This implicated a single likely causal SNP in three loci (*PXK, CDK6, TPD52, CD5*) and less than three likely causal SNPs in the other loci (*IL6ST, TCTE1*) (Supplementary Data 2).

For 13 loci, our correlated, dynamic data support the currently assigned gene, and provide the first biological evidence for 7 of these genes (Supplementary Data 2). These genes include *DDX6, PRKCH, RBPJ, PVT1, PRDM1 and PTPN2*, with many interactions confirming genes up to 200 kb, and one over 800 kb (*PVT1*), from the RA credible set SNPs SNP, but always constrained within TADs.

For a number of loci, our results point to complex relationships between RA credible set SNP regions and putative causal genes (Supplementary Data 2). For some regions, whilst we did not demonstrate direct interaction between an ATAC-seq peak and gene, the genomic region spanning the credible SNP set made a long distance physical interaction with genes that have supportive evidence for disease involvement. For example, on chromosome 10, an intronic region within the *ARID5B* gene, containing RA credible set SNPs, interacts with *RTKN2*, involved in the NFKB pathway, and containing nsSNPs associated with Asian RA[23] (Supplementary Fig. 14b). Similarly on chromosome 3, a region with RA credible set SNPs 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. 14c). We also provide evidence for, complex regions with direct ATAC-seq, gene promoter interactions. Two regions in particular provided insight into potential target genes for RA. Two ATAC-seq peaks, containing five SNPs that are RA credible set SNPs 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

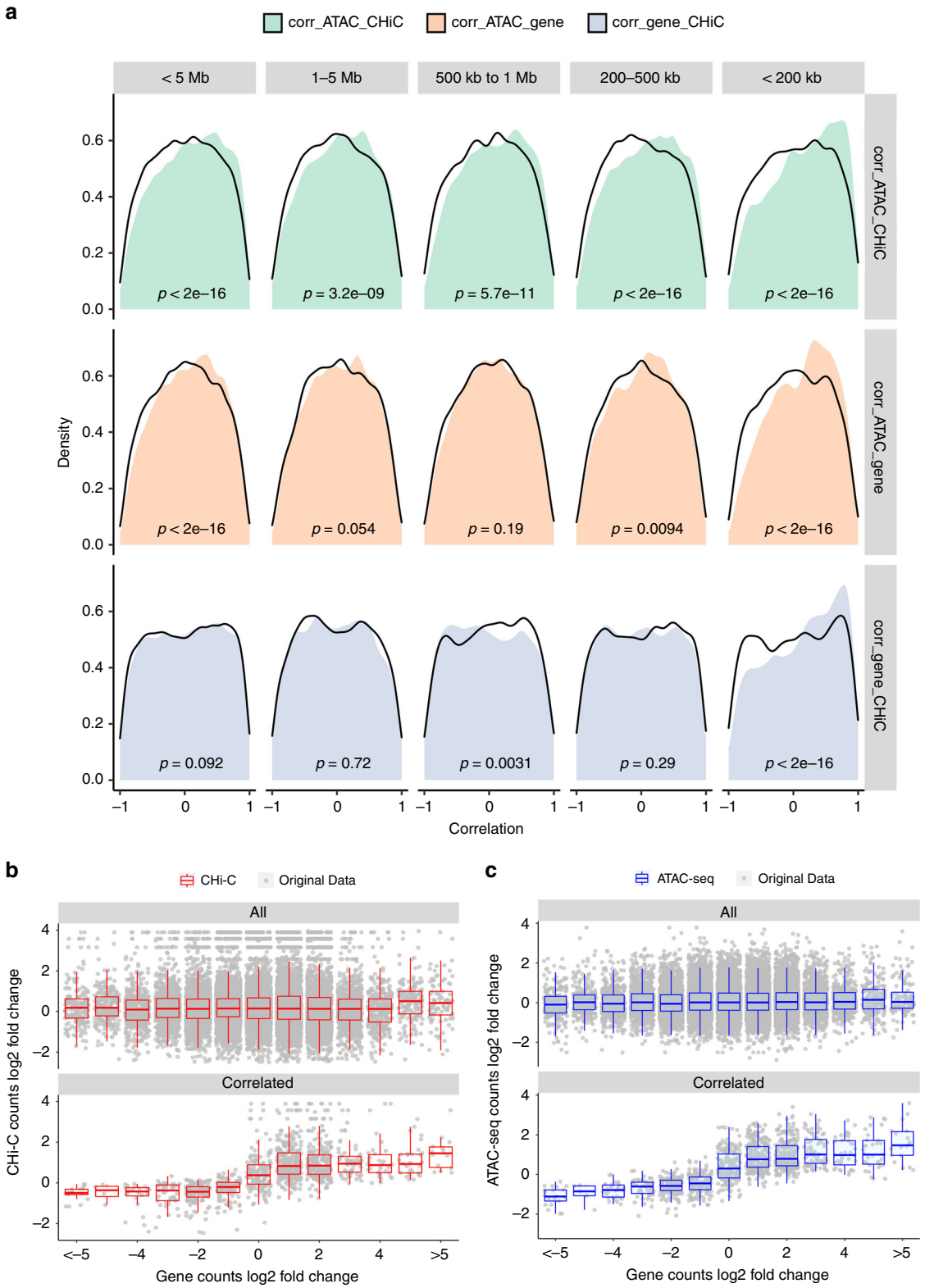

Fig. 4 Illustrations of the correlations between ATAC-seq, CHi-C and RNA-seq time course profiles. a Density plots of the Pearson correlation coefficients between ATAC-seq and CHi-C, ATAC-seq and gene and gene and CHi-C under various distance ranges around promoters. Distances ranges include those <5 Mb, between 1 and 5 Mb, between 500 kb and 1 Mb, between 200 and 500 kb and <200 kb. Black lines represent the density plots of the corresponding random background. *P* values from Wilcoxon tests are labelled in each panel with sample sizes of 27,805, 8229, 4742, 7942 and 6892 for categories < 5 Mb, 1–5 Mb, 500 kb–1 Mb, 200–500 kb and <200 kb, respectively. **b** Comparison of the log2 fold change between CHi-C and gene data for all data set (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 data sets (upper) and those highly correlated ones with Pearson correlation coefficients between ATAC-seq, gene and CHi-C over 0.5 (lower). Box plot shows the median, the interquartile range (IQR) and Turkey whiskers (±1.5 times IQR).

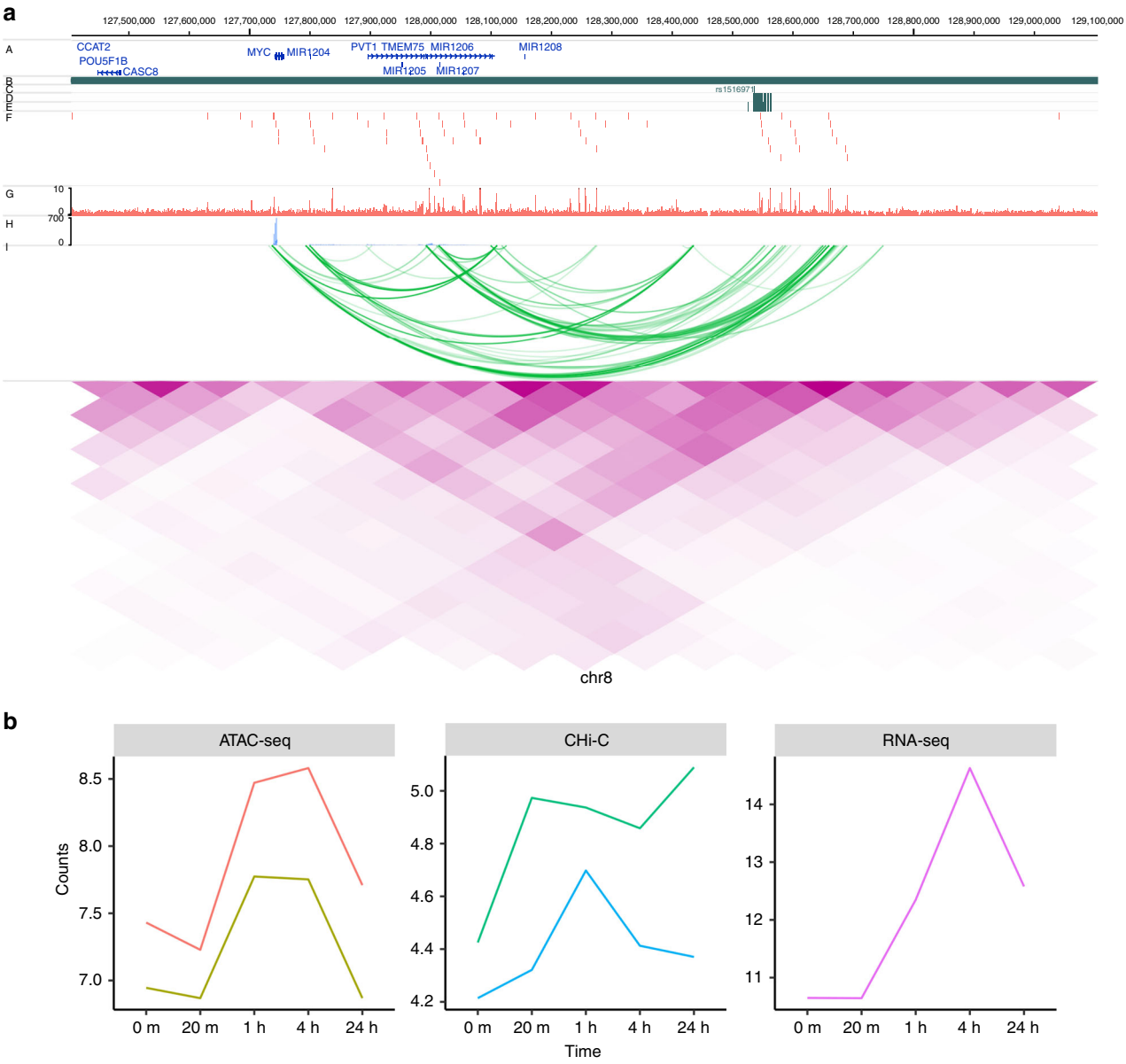

Fig. 5 Illustration of genomic interaction activities around MYC. a Screenshot of the SNPs (dark green), ATAC-seq peaks (red), RNA-seq (lightblue), CHi-C interactions (green) and heatmap from Hi-C (purple) around MYC at time 4 h. *y*-axis labels A–J represent different tracks plotted. A RefSeq genes, B TADs, C Index SNPs, D LD SNPs, E 99% credible sets, F ATAC-seq peaks, G ATAC-seq signal, H RNA-seq, I CHi-C, J Hi-C. **b** Time course profiles of ATAC-seq (left), CHi-C (middle) and RNA-seq (right) of the data associated with SNPs data around MYC.

screen in K562 erythroleukemia cells[20]. 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.

Putting this gene list through a drug target pipeline[24] demonstrated how the newly implicated or confirmed genes from this study, such as *MyC*, *GSN* and *MMP9*, are existing therapeutic targets (Supplementary Table 2).

Confirmation of correlated interaction with CRISPR/Cas9. Finally we sought to demonstrate how an ATAC-seq peak, containing an RA credible set SNP, 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 investigated 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[25–28]. To do this we used CRISPRa, with dCas9-p300, and the HEK293T cell line. Although using a cell line, as opposed to primary immune cells, is not optimal, we used HEK293T cells as the model system since we and others have previously demonstrated how the TAD region containing both *COG6* and *FOX01* is highly conserved in a wide range of cell types, including primary cells as well as various cell lines (Jurkat, GM and HEK293T). We designed guide RNAs (gRNAs) to cover the single enhancer that contains the two ATAC-seq peaks and three RA credible set SNPs, pooling the guides in a single transfection (Supplementary Fig. 15). We demonstrated that, when we activate the *COG6* intronic enhancer with this system and targeted gRNAs, not only do we observe a consistent increase in *COG6* mRNA expression itself, we obtain robust, reproducible up regulation of *FOXO1*

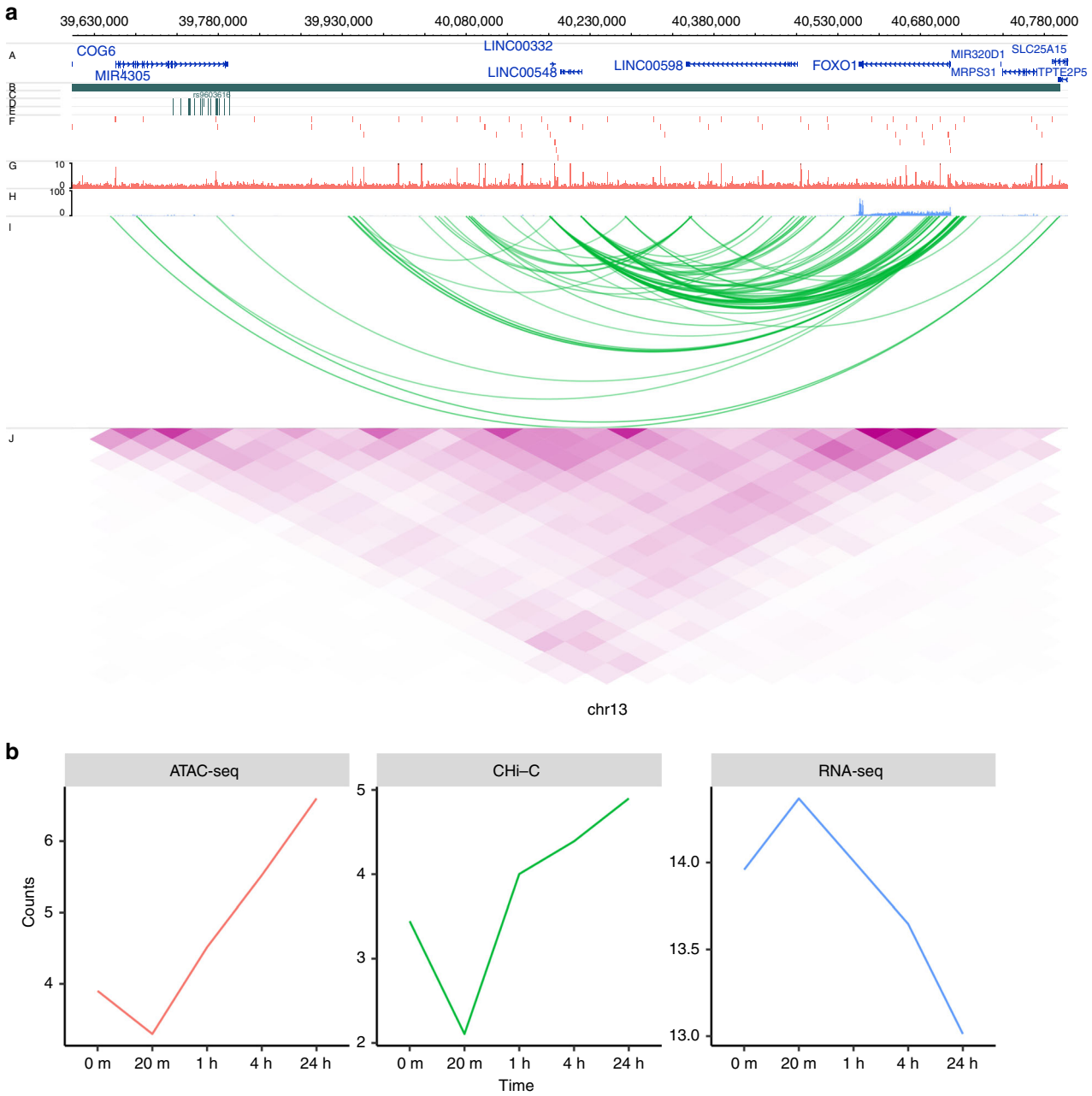

Fig. 6 Illustration of genomic interaction activities around FOXO1. a Screenshot of the SNPs (dark green), ATAC-seq peaks (red), RNA-seq (lightblue), CHi-C interactions (green) and heatmap from Hi-C (purple) around FOXO1 at time 4 h. *y*-axis labels A–J represent different tracks plotted. A RefSeq genes, B TADs, C Index SNPs, D LD SNPs, E 99% credible sets, F ATAC-seq peaks, G ATAC-seq signal, H RNA-seq, I CHi-C, J Hi-C. **b** Time course profiles of ATAC-seq (left), Chi-C (middle) and RNA-seq (right) of the data associated with SNPs data around FOXO1.

gene expression (Fig. 7). Although the credible set SNP in this region is a strong eQTL for *COG6*, this CRISPR validation of the correlated interaction, DNA activity and expression data, alongside previous immunological studies, implies 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.

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. This 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 anti-CD3/anti-CD28, we analyse ATAC-seq, RNA-seq, Hi-C and CHi-C over 24 h. Time points were selected to understand how chromatin changes in the very early stages after stimulation, and how GWAS associated variants for autoimmune diseases, in particular RA, map to these chromatin dynamics. We show that the conformation of the DNA at a higher structural level of A/B compartments[29] and TADs[30] 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, are highly

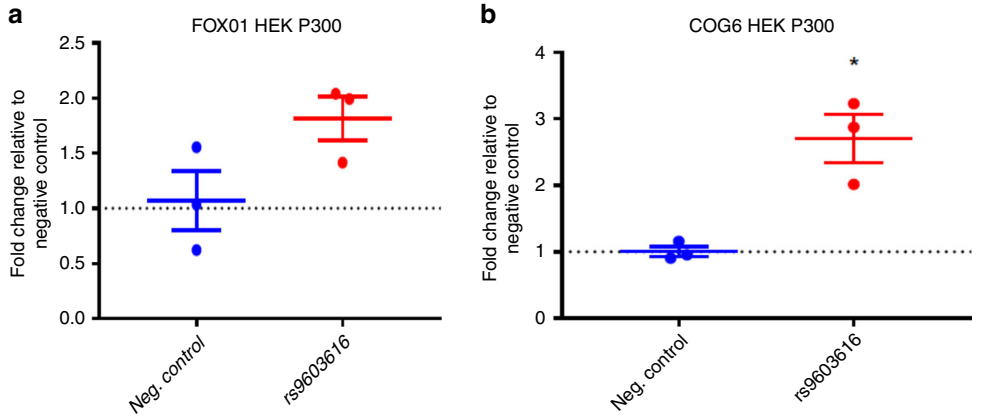

Fig. 7 CRISPR dCAS9 activation (CRISPRa) targeting of an RA-associated variant. The region around an associated RA variant (rs9603616) on chromosome 13 intronic of the COG6 gene was targeted in the HEK2937T cell line, using p300 as the activator. **a** Fold change (qPCR) effect on FOXO1 gene expression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

dynamic post stimulation, with over 30% of individual interactions showing some degree of change over 24 h; however, only a minority of these changes are correlated with a change in expression. These results suggest that ATAC-seq peaks are associated with two CHi-C interactions on average, with each gene making contact with 7.7 ATAC-seq peaks on average. 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. These data therefore suggest that, when assigning GWAS variants to putative causal genes, 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 long range (~1 Mb) gene target using CRISPR/Cas9. Marks for active enhancers (H3K27ac and H3K4me1) are highly enriched in both all ATAC-seq peaks and those interacting with a gene promoter. However, since CTCF sites are also enriched under these ATAC-seq peaks, this does not exclude the possibility that these are involved in the dynamics of the data.

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[31,32], potentially playing a role in the invasiveness of these cells. More notably for this 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[27]. 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[27]. In this 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: (1) both DNA activity and interaction intensity are independently important in the regulation of genes; and (2) 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 these 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 isolated from peripheral blood mononuclear cells (PBMCs) and collected from three healthy individuals with informed consent and ethical approval (Nat Rep 99/8/84). PBMCs were initially isolated using Ficoll density gradient centrifugation. CD4+ T cells were isolated from 100 million PBMCs by negative selection using the EasySep Human T-cell isolation kit (StemCell Technologies, catalogue #17952) according to the manufacturer's instructions. Flow cytometry analysis was used to confirm a high level of purity of CD4+ T cells (Supplementary Fig. 2a).

For flow cytometry, samples were collected on a BD LSR Fortessa X-20, using SYTOX red (Thermo Fisher Scientific) to discriminate against dead cells and monoclonal antibodies against CD3 and CD4 to establish purity (CD3, OKT3 dilution 1/50; CD4, OKT4, dilution 1/100; Biolegend). The gating strategy for a representative sample is shown (Supplementary Fig. 2b). Acquisition data were analysed using FlowJo 10.6.1 (BD).

Isolated CD4+ T cells were plated in six-well plates then stimulated with anti-CD3/anti-CD28 Dynabeads (Life Technologies) over a period of 24 h, 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 (8–10 million) were harvested and fixed in formaldehyde, samples for RNA-seq (5 million 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 min, 20 min, 1 h, 2 h, 4 h and 24 h. 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 h and 4 h and one sample for Hi-C and CHi-C was taken at 24 h, respectively.

Library generation for CHi-C and Hi-C. To generate libraries for Hi-C experiments, 8–10 million CD4+ T cells were harvested at the appropriate time point and formaldehyde crosslinking carried out as described in Belton et al.[33]. Cells were washed in Dulbecco's modified Eagle's medium (DMEM) without serum then crosslinked with 2% formaldehyde for 10 min at room temperature. The crosslinking reaction was quenched by adding cold 1 M glycine to a final concentration of 0.125 M for 5 min at room temperature, followed by 15 min 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 °C.

Hi-C libraries were prepared from fixed CD4+ T cells from three individuals that were pooled at the lysis stage to give ~30 million cells. Cells were thawed on ice and re-suspended in 50 ml freshly prepared ice-cold lysis buffer (10 mM Tris-HCl pH 8, 10 mM NaCl, 0.2% Igepal CA-630, one protease inhibitor cocktail tablet). Cells were lysed on ice for a total of 30 min, with 2 × 10 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 5–6 million cells for digestion. Following lysis, libraries were digested using HindIII then prepared as described in van Berkum et al.[34] with modifications described in Dryden et al.[35]. Final library amplification was performed on multiple parallel reactions from libraries immobilised on Streptavidin beads using eight cycles of PCR using CHiC_TruPE_PCR primers (Supplementary Table 3) if the samples were to be used for CHi-C, or 6 cycles using the TruSeq_Universal and Indexed primers (Supplementary Table 3) for Hi-C. Reactions were pooled post-PCR, purified using SPRI beads and the final libraries re-suspended in 30 µl TLE. Library quality and quantity were assessed by Bioanalyzer and KAPA qPCR prior to sequencing on an Illumina HiSeq2500 generating 100 bp paired-end reads (Babraham sequencing facility).

Solution hybridisation capture of Hi-C library. Pre-CHi-C libraries corresponding to 750 ng were concentrated in a Speedvac then re-suspended in 3.4 µl 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 eight cycles of PCR using the TruSeq_Universal and Indexed primers (Supplementary Table 3) 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[21] and a *trans*-ethnic GWAS meta-analysis[22]. 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 ($r^2 \geq 0.8$; 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 500 bp 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 1 Mb upstream and downstream of associated RA SNPs from Eyre et al.[21] and Okada et al.[22] 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. Briefly, cells were thawed and centrifuged at 5000 × *g* for 5 min then the pellet re-suspended in 1 ml cold buffer RLN. RNA isolation was continued from this point using Qiagen RNeasy kit reagents and protocol. 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 100 ng of nuclear RNA as Input. Library quality and quantity were assessed by Bioanalyzer and KAPA qPCR prior to sequencing. Each library was sequenced on half a lane of an Illumina HiSeq2500 generating 100 bp 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.[36]. Briefly, cells were pelleted at 500 × *g* for 5 min at 4 °C, the supernatant was removed and the cells were lysed in 50 µl cold lysis buffer (10 mM Tris-HCl pH7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Igepal CA-630). Immediately following the lysis reaction, the samples were centrifuged at 500 × *g* for 10 min at 4 °C, the supernatant was discarded and the cells were kept on ice.

The transposition reaction was carried out using Illumina Nextera DNA Sample Preparation Kit reagents. Cells were re-suspended in 10 µl 2x reaction buffer (TD), 2.5 µl transposase (TDE1) and 7.5 µl nuclease-free water and then incubated at 37 °C for 30–45 min in a thermomixer. Samples were purified using Qiagen MinElute columns following the manufacturer's protocol, eluting in 10 µl. PCR was performed for ten cycles using Nextera dual indexing primers under the following conditions: 72 °C 5 min; 98 °C 30 s; ten cycles of 98 °C 10 s, 63 °C 30 s, 72 °C for 1 min. PCR products were purified using Qiagen MinElute columns following the manufacturer's protocol, eluting in 20 µl. Library quality and quantity were assessed by Bioanalyzer and KAPA qPCR prior to sequencing. Each library was sequenced on half a lane of an Illumina HiSeq2500 generating 100 bp paired-end reads (Babraham sequencing facility, Cambridge).

Hi-C data processing. Hi-C data were mapped to GRCh38 by HiCUP[37]. The maximum and minimum di-tag lengths were set to 800 and 150, respectively. HOMER[38] Hi-C protocol was applied to Hi-C bam file and normalised Hi-C matrices were generated by analyzeHiC command from HOMER with resolution of 40,000 bp (analyzeHiC –res 40,000 –balance). TADs were generated by the command findTADsAndLoops.pl (–res 40,000). A/B compartments were generated by runPCA.pl (–res 40,000) 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[39] 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[40] (with option –x 2000) and reads with mapping quality lower than 30 were filtered by SAMtools[41]. Duplications were removed by Picard (https://broadinstitute.github.io/picard/). The three replicated bam files at each time point were merged by SAMtools. MACS2[42] 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[43] 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[44]. 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 <10 were removed in each replicate. Only genes that have expressions in both replicates were kept. In all, 22,126 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 data set 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[17] with a radial basis function 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 (see the following equation):

$$\mathrm{BIC} = k\ln(n) - 2\ln(\hat{L}), \quad (1)$$

where k is the number of parameters used in the specified model, n is the sample size and $\hat{L}$ is the maximised 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 of 1 was also applied to the log-likelihood ratio (LR) statistics, with $\mathrm{LR} = -2\ln(\hat{L}_{\mathrm{RBF}} - \hat{L}_{\mathrm{STATIC}})$, where $\hat{L}_{\mathrm{RBF}}$ and $\hat{L}_{\mathrm{STATIC}}$ are the maximised 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 9680 were outside promoter regions ([+500 bp,−1000 bp] around genes). These ATAC-seq peaks were clustered using a Gaussian process mixture model[45]. MOTIFs for each cluster were searched by findMotifGenome.pl (–mask –len 5,6,7,8,9,10,11,12 –size given) from HOMER with static 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[22], 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[46], given as follows:

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

where β and V denote the estimated log odds ratio 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: (1) ranking all SNPs according to their Bayes' factor; and (2) including ranked SNPs until their cumulative posterior probability of causality attained or exceeded 0.99.

Overlap between ATAC-seq peaks and autoimmune GWAS loci. Autoimmune GWAS loci (defined by NCIT:C2889, OBI:OBI 1110054, OMIM:613551, OMIM: 615952, OMIM: 617006, SNOMEDCT:85828009) were downloaded from the GWAS catalogue (https://www.ebi.ac.uk/gwas/efotraits/EFO_0005140 Accessed July 2018) and SNPs with $r^2 \geq 0.8$ identified using PLINK v1.07. These were then overlapped with ATAC-seq peaks linked to CHi-C genes, using bedtools v2.21.0, to identify regions of open chromatin containing an autoimmune GWAS SNP SNP or highly correlated variant.

RA heritability and enrichment. We performed RA SNP enrichment and heritability estimates in the ATAC peaks identified throughout the time course. This was performed using partitioned heritability analysis from the LD score regression software[47] and European ancestry data from the latest RA GWAS meta-analysis[21]. Briefly, the heritability of RA and SNP enrichments is computed in partitioned sections of the genome, in this case, the ATAC-seq peaks at each time point.

CRISPR activation using dCas9-p300. *Cell lines*: HEK293T cells (Clontech) were cultured in high glucose-containing Dulbecco's modified Eagle's medium (Sigma) supplemented with 10% FBS and 1% penicillin/streptomycin at 37 °C/5% CO_2 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 2 μg mL^{-1} of puromycin and cultured for a week before being banked as a cell line. A second round of lentiviral transduction was performed to introduce the 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 gRNAs were cloned into the guide delivery vector pLK05.sgRNA. EFS.GFP (Addgene #57822). A negative control guide Scr2 (AACAGTCGCGTT TGCGACT) 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 TSS of the promoter of the *IL1RN* gene that has been previously shown to increase expression of the IL1RN gene substantially[48]. 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 10 million cells per transfer vector in 15 cm plates in a volume of 20 ml of DMEM 10% FBS without penicillin/streptomycin. 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 12 μg at a ratio of 4:2:1:1, respectively, in 2 ml of serum free DMEM without phenol red.

PEI 1 mg mL^{-1} 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 min. 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. Twenty-four hours later fresh DMEM supplemented with 10% FBS and 1% penicillin/streptomycin was added. The viral supernatant was collected 72 h after transduction, cleared by centrifugation at 500 × *g* for 5 min at 4 °C then passed through a 0.45 μm pore PVDF Millex-HV (Millipore). Lentivirus was aliquoted and stored at −80 °C for future use.

Transduction of HEK293T p300 cell line with the gRNAs. dCas9-p300-HEK293T cells were plated at 3×10^5 cells mL^{-1} onto six-well plates in triplicate for each gRNA. Twenty-four hours later the medium was replaced with DMEM 10% FBS without penicillin/streptomycin. 1 ml of each gRNA generated lentivirus was added to each well. Twenty-four hours later the medium was changed to DMEM containing 10% FBS and 1% penicillin/streptomycin. Cells were cultured for 5 days and then sorted for the top 60% of cells expressing GFP using flow cytometry.

RNA extraction and qPCR. When confluent 2×10^6 cells were centrifuged at 400 × *g* for 5 min 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. 100 ng of RNA for each sample was used in a single RNA-to-Ct reaction (Thermo Fisher) to assay gene expression. Taqman assays for 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 were analysed in graph pad using one-way ANOVA.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

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. The full data are accessible using the following link: http://epigenomegateway.wustl.edu/legacy/?genome=hg38&datahub=http://bartzabel.ls.manchester.ac.uk/worthingtonlab/functional_genomics/GSE138767/GSE138767.json. All other relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding author upon reasonable request. The source data underlying Figs. 2–4, 5b, 6b, 7 and Supplementary Figs. 1–8, 10–12 are openly available in repository Zenodo, with https://doi.org/10.5281/zenodo.3899030.

Code availability

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

Received: 22 October 2019; Accepted: 6 August 2020;
Published online: 02 September 2020

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**, 337–343 (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* **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**, 1665–1680 (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**, 1140–1150 (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. Robson, M. I. et al. Constrained release of lamina-associated enhancers and genes from the nuclear envelope during T-cell activation facilitates their association in chromosome compartments. *Genome Res.* **27**, 1126–1138 (2017).
16. 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).
17. Yang, J., Penfold, C. A., Grant, M. R. & Rattray, M. Inferring the perturbation time from biological time course data. *Bioinformatics* **32**, 2956–2964 (2016).
18. 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).
19. 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).
20. Fulco, C. P. et al. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **6056**, 1–8 (2016).
21. Eyre, S. et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **44**, 1336–1340 (2012).
22. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
23. 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).
24. Martin, P. et al. Chromatin interactions reveal novel gene targets for drug repositioning in rheumatic diseases. *Ann. Rheum. Dis.* **78**, 1127–1134 (2019).
25. Ludikhuize, J. et al. Inhibition of forkhead box class O family member transcription factors in rheumatoid synovial tissue. *Arthritis Rheum.* **56**, 2180–2191 (2007).
26. 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).
27. 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).
28. 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.* **120**, 14628–14635 (2019).
29. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
30. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
31. Qu, Z. et al. Local proliferation of fibroblast-like synoviocytes contributes to synovial hyperplasia. *Arthritis Rheum.* **37**, 212–220 (1994).
32. 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.* **50**, 2794–2802 (2004).
33. Belton, J.-M. et al. Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
34. Van Berkum, N. L. et al. Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* **39**, e1869 (2010).
35. 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).
36. 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).
37. Wingett, S. et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310 (2015).
38. 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).
39. Cairns, J. et al. CHiCAGO: robust detection of DNA looping interactions in capture Hi-C data. *Genome Biol.* **17**, 127 (2016).
40. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
41. Wysoker, A. et al. The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
42. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
43. Stark, R. & Brown, G. *DiffBind: dIfferential Binding Analysis of ChIP-Seq Peak Data*, Vol. 100, 3–4 (R Package Version, 2011).
44. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).
45. Hensman, J., Lawrence, N. D. & Rattray, M. Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC Bioinform.* **14**, 252 (2013).
46. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81**, 208–227 (2007).
47. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
48. Hilton, I. B. et al. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* **33**, 510 (2015).

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.

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. X.G. 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 paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-18180-7.

Correspondence and requests for materials should be addressed to M.R. or S.E.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.