Functional effects of variation in transcription factor binding highlight long-range gene regulation by epromoters

Joanna Mitchelmore[1], Nastasiya Grinberg[2], Chris Wallace[2,3] and Mikhail Spivakov[1,4,5,*]

[1] Nuclear Dynamics Programme, Babraham Institute, Babraham Research Campus, Cambridge CB22 3AT, UK
[2] Cambridge Institute of Therapeutic Immunology & Infectious Disease (CITIID), Cambridge Biomedical Campus, University of Cambridge, Cambridge, CB2 0AW, UK
[3] MRC Biostatistics Unit, Cambridge Biomedical Campus, University of Cambridge, Cambridge CB2 0SR, UK
[4] MRC London Institute of Medical Sciences, Du Cane Road, London W12 0NN, UK
[5] Institute of Clinical Sciences, Faculty of Medicine, Imperial College, Du Cane Road, London W12 0NN, UK

* Corresponding author. Email mikhail.spivakov@lms.mrc.ac.uk

Present address: Joanna Mitchelmore, Novartis Institutes of Biomedical Research, Fabrikstrasse 16, Novartis Campus Basel, 4056. Switzerland

ABSTRACT

Identifying DNA cis-regulatory modules (CRMs) that control the expression of specific genes is crucial for deciphering the logic of transcriptional control. Natural genetic variation can point to the possible gene regulatory function of specific sequences through their allelic associations with gene expression. However, comprehensive identification of causal regulatory sequences in brute-force association testing without incorporating prior knowledge is challenging due to limited statistical power and effects of linkage disequilibrium. Sequence variants affecting transcription factor (TF) binding at CRMs have a strong potential to influence gene regulatory function, which provides a motivation for prioritising such variants in association testing. Here, we generate an atlas of CRMs showing predicted allelic variation in TF binding affinity in human lymphoblastoid cell lines (LCLs) and test their association with the expression of their putative target genes inferred from Promoter Capture Hi-C and immediate linear proximity. We reveal over 1300 CRM TF-binding variants associated with target gene expression, the majority of them undetected with standard association testing. A large proportion of CRMs showing associations with the expression of genes they contact in 3D localise to the promoter regions of other genes, supporting the notion of 'epromoters': dual-action CRMs with promoter and distal enhancer activity.

INTRODUCTION

Identifying DNA cis-regulatory modules (CRMs) that control the expression of specific genes is crucial for deciphering the logic of transcriptional control and its aberrations. Advances of the last decade have made it possible to predict active CRMs based on chromatin features (1, 2) and detect the binding of dozens of TFs to these regions (3, 4). However, deletion of known or predicted CRMs often shows no observable phenotype, suggesting that some CRMs either lack appreciable gene regulatory function or are efficiently buffered by other sequences, at least under normal conditions (5–9). In addition, the sequence, chromatin state and genomic location of CRMs do not immediately provide information on their target genes (10). Therefore, evidence from complementary approaches is required to establish the function of specific CRMs in transcriptional control.

Natural genetic variation can theoretically provide a direct indication of gene regulatory function by revealing the allelic associations between specific variants and gene expression (11, 12). While expression quantitative trait loci (eQTLs) identified this way have provided important insights into gene control and the mechanisms of specific diseases (13, 14), a number of challenges hamper comprehensive detection of functional sequences in 'brute-force' eQTL testing (15, 16). In particular, the immense search space leads to a heavy multiple testing burden resulting in reduced sensitivity. This problem is typically mitigated in part by testing for 'cis-eQTLs' separately within a limited distance window (~100kb); this distance range is, however, an order of magnitude shorter than that of known distal CRM activity (17–19). In addition, correlation structure arising from linkage disequilibrium (LD) requires disentangling causal from spurious associations, which is particularly challenging in the likely scenario when multiple functional variants with modest effects co-exist within the same LD block (20). These challenges provide a strong motivation for incorporating prior knowledge into association testing for identifying causal regulatory variants.

The recruitment of transcription factors (TFs) to CRMs plays a key role in their regulatory function (21, 22), and mutations leading to perturbed TF binding are known to underpin developmental abnormalities and disease susceptibility (18, 23, 24). Therefore, sequence variation affecting TF binding affinity at CRMs has a strong potential to have causal influence on their regulatory function and can therefore provide insights into the logic of gene control. Variation in TF binding across multiple individuals has been assessed directly for several TFs (25–30), but high resource requirements of these analyses limit the number of TFs and individuals profiled this way. Alternatively, the effects of local sequence variation on TF binding can be predicted, at least in part, based on prior information regarding the TFs' DNA binding preferences. The representation of such preferences in the form of position weight matrices (PWMs) (31) has proven particularly useful, as it provides a quantitative measure of how much a given sequence substitution is likely to perturb TF binding consensus. Consistent with this, we and others have previously shown that the specificity of TF binding preferences to a given motif position correlates with the functional constraint of the underlying DNA sequences, both within and across species (32–34). Classic PWM-based approaches to TF binding prediction focused on identifying short sequences showing a non-random fit to the PWM model compared with background (35, 36). More recently, biophysical

modelling of TF binding affinity (37, 38) has provided a natural framework to extend this analysis by integrating over all PWM match signals within a DNA region (39, 40), including those from lower-affinity sites that are a known feature of many functional CRMs (41–43).

Long-range CRMs such as gene enhancers commonly act on their target promoters through DNA looping interactions (44, 45). Therefore, information on three-dimensional chromosomal organisation enables predicting the putative target genes of these elements (46, 47) and thus has the potential to significantly improve the functional interpretation of regulatory variation. Approaches that couple chromosome conformation capture with target sequence enrichment such as Promoter Capture Hi-C (48–50) are particularly useful in this regard, as they make it possible to detect regulatory interactions globally and at high resolution with reasonable amounts of sequencing (51–59).

Here we integrate TF binding profiles in a human lymphoblastoid cell line (LCL) (4) with patterns of natural sequence variation (60) to generate an atlas of CRMs predicted to show significant TF binding variability across LCLs derived from multiple individuals. We delineate the putative target genes of these CRMs from their interactions with gene promoters based on Promoter Capture Hi-C and linear proximity (49, 61), and test for associations between the CRMs' TF-binding affinity and target gene expression using transcriptomics data for hundreds of LCLs (62). Prioritising CRMs that show predicted variation in TF-binding affinity based on a biophysical model (39, 40) makes it feasible to perform association analysis in a manner that accounts for multiple variants affecting the binding of the same TF, as well as for multiple CRMs targeting the same gene. Using this approach, we reveal over 1300 CRM variants associated with expression of specific genes, the majority of them undetected with conventional eQTL testing at a standard FDR threshold. We find that a large proportion of CRMs showing associations with the expression of distal genes localise in the immediate vicinity of the TSSs of other genes and connect to their targets via DNA looping interactions, suggesting their role as 'epromoters': the recently identified dual-action regulatory regions with promoter and distal enhancer activity (63–65).

MATERIALS AND METHODS

CRM definition

ChIP-seq narrow peak files for 52 TFs in GM12878 were downloaded from the UCSC ENCODE portal (4). Where multiple datasets were available for the same TF, the intersect of the ChIP-seq peaks was taken for all TFs except ERG1, for which we took the union of the two datasets available, since one of them had substantially fewer peaks than the other. CRMs were defined by taking the union of the peaks for the 52 TFs with a minimum overlap of one base pair.

Detection of TF binding affinity variants

Variant calls for 359 LCLs of European ancestry (CEU, TSI, FIN, GBR, and IBS) that overlapped with the CRMs defined as above were downloaded from the 1000 Genomes Project (release Phase 3; 20130502) (60). Multi-allelic variants and variants with a minor allele frequency < 5% were removed. Unique haplotypes (i.e., unique

combinations of SNPs/indels) were identified across the 359 LCLs individuals for each CRM. The GRCh37 genomic sequence for each CRM (accessed using the Bioconductor package BSGenome (https://doi.org/10.18129/B9.bioc.BSgenome) was then patched to create the sequence for each unique haplotype.

For each TF detected as bound at a given CRM in GM12878 (based on ChIP-seq data), we computed the affinity for each haplotype and each PWM for this TF available from ENCODE (66) using the TRAP biophysical model (39), as implemented in the R package tRap (https://github.com/matthuska/tRap). Default parameters were used, with the exception of setting pseudocount to zero, since we were using frequency as opposed to count matrices. We chose TRAP over a motif hit-based approach, as it naturally incorporates the effects of multiple low affinity sites and multiple variants per CRM.

CRM binding affinities were normalised using a method proposed by Manke et al. (40), such that changes in them could be compared between different PWMs. Briefly, CRM affinities are converted to statistical scores (A) representing the probability of observing a given or higher affinity for a given TF in the background sequence (note that lower values of A therefore reflect higher affinities). Binding affinities are parameterised using the extreme value distribution whose parameters are estimated for a range of background sequences encompassing the lengths of all CRMs (40, 100, 200, 250, 300, 400, 500, 800, 1000, 2000, 3000) using the fit.gev function in the the tRap R package. CRMs not bound by a given TF are cut/extended to the required length and used as background sequences.

For all CRM-TF/PWM combinations with $A<0.1$ in the highest-affinity allele of GM12878, we computed the log-fold change in affinity between all observed haplotypes and the highest-affinity allele of GM12878 for the given PWM:

$$\text{logFCA} = \log10(A_{ALT}) - \log10(\min(A_{GM12878})),$$

where $\min(A_{GM12878})$ is the normalised affinity of the highest-affinity allele in GM12878 cells, and A_{ALT} is the normalised affinity of the alternative haplotype. For instances where A_{ALT} or $A_{GM12878}$ for a given PWM was zero, the lowest observed non-zero normalised affinity for that PWM across all CRMs was used instead. The logFCAs for multiple PWMs of the same TF were then combined by taking the median. Overall, this approach produced a single logFCA for each TF binding affinity haplotype at each CRM. We shall refer to this quantity as the "log-ratio" in the Results section.

DNase I sensitivity QTL (dsQTL) analysis

The dsQTL dataset from (67) lists significant associations between normalized DNase-seq read depth (binned in 100bp non-overlapping windows) and the genotypes of SNPs/indels within 1kb of the DHS in 70 Yoruban LCLs. We downloaded this dataset from Gene Expression Omnibus (accession number GSE31388), and converted it to GRCh37 using liftOver (68). For all CRMs with a predicted logFCA > 0 for at least one TF, the individual effect of all SNPs at the CRM on TF affinity was calculated. CRMs were then filtered for those, where the SNP causing the largest change in TF affinity ("driver SNP") had a MAF<0.05 in the 70 individuals from (67). We then counted the number of overlaps between these CRMs and the 100bp DNase HS windows (minimum

overlap 1bp), repeating this for CRMs filtered according to successively larger logFCA thresholds. To estimate expected overlap, for each threshold, we randomly sampled a control set of CRMs 1000 times, matching the sample size and "driver" SNP allele frequency distribution to the test set at a given threshold, and overlapped this set with DNase HS windows in the same way as the test set.

Linking of CRMs with target genes

Promoter Capture Hi-C data for GM12878 were obtained from Mifsud et al. (49). Significant interactions were re-called at a *HindIII* restriction fragment level using the CHiCAGO pipeline (61), with a CHiCAGO score cutoff of five (CHiCAGO scores correspond to soft-thresholded, –log weighted p-values against the background model). Baits were annotated for transcriptional start sites (TSSs) using the bioMart package in R (69) based on Ensembl TSS data for GRCh37 reference assembly. Baits containing TSSs for more than one gene were excluded (4,178 out of 22,076), leaving 17,898 baits in the analysis. CRMs were assigned to target promoters by overlapping with the promoter-interacting regions (PIRs) of significant interactions ("distal" CRMs). Restriction fragments immediately flanking the promoter fragment are excluded from Promoter Capture Hi-C analysis, creating a "blind window". Therefore, we additionally called "proximal" CRMs using a window-based approach, assigning all CRMs located within within 9kb of the midpoint of the promoter-containing fragment to the respective promoter.

Gene expression data processing

We downloaded PEER-normalised (70) gene-level RPKMs for 359 EUR LCLs profiled in the GEUVADIS project (62) from ArrayExpress (71) (accession E-GEUV-3). The data were filtered to expressed genes by removing genes with zero read counts in >50% of samples. For expression association testing by linear regression, the PEER-normalised residuals for each gene were further rank-transformed to standard normal distribution, using the rntransform function in the R package GenABEL (72).

Association between TF binding affinity variants and gene expression: thresholded approach

In this approach, we classified each predicted TF-binding affinity CRM haplotype as either "high" or "low" affinity based on a threshold. In some instances, however, using a hard threshold to classify alleles can result in alleles with very similar log-fold affinity changes being differentially classified, which can obscure true affinity-expression associations. To avoid this, we used a dynamic thresholding approach, where for each affinity variant we set the threshold logFCA_0 to 80% of the value of the 85th percentile of variants less than or equal to the hard threshold of -0.3. All alleles with logFCA <= logFCA_0 were taken as low affinity. Alleles with either logFCA > $\text{logFCA}_0/4$ (for $\text{logFCA}_0/4 > -0.3$) or logFCA > -0.3 were taken as high affinity. Note this resulted in some alleles classified as neither high nor low affinity. Individuals containing at least one unclassified allele for a given TF/CRM were excluded from the testing for the respective association (the number of individuals tested for each association is listed in Table S1).

A regression model was then fitted using TF-binding affinity CRM haplotypes as predictors of the expression level of their target genes (presented in terms of normalised PEER residuals). Suppose that a gene is targeted by K predicted TF-affinity CRM variants, denoted as $X = (X_1, X_2,\ldots, X_K,)$, which are encoded as the number of copies of the low affinity allele carried by each individual. The regression model is fitted as follows:

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K,$$

where $E[Y]$ is the expected value of the normalised PEER residuals Y. Where multiple predicted TF affinity CRM variants targeting a given gene were in perfect correlation ($|\beta|>0.99$), they were collapsed into a single predictor.

ANOVA was used to test the overall significance of each regression model, with multiple testing correction performed on the gene-level p-values by FDR estimation. For genes showing significant associations in models with multiple TF binding affinity variants as predictors, t-tests were performed to identify variants with regression coefficients significantly different from zero. Variants with unadjusted coefficient-level p-values<0.05 were taken to be significantly associated with target gene expression, conditional on significant gene-level association at 10% FDR.

Association between TF binding affinity variants and gene expression: threshold-free approach

In this approach, we performed multiple regression using PEER expression residuals for each gene as the response variable, this time using the sum of logFCA across both alleles for each individual for each TF affinity CRM variant as predictors instead of thresholded CRM haplotypes. For each gene, all distal and proximal CRMs with logFCA > 0 were included. As with the thresholded approach, ANOVA was used to test the significance of each gene model, and genes showing associations at 10% FDR were considered significant.

Due to high collinearity among the predicted affinity changes, to identify specific CRM variants significantly associated with target gene expression we used elastic net regression for each significantly associated gene (λ_2=0.5). The significance of each predictor as it entered the model was then tested using a method by Lockhart et al. (73) and implemented in the covTest R package (https://cran.r-project.org/src/contrib/Archive/covTest/covTest_1.02.tar.gz). Variants that entered the model with p<0.05 and remained in the model were taken as significant.

eQTL fine-mapping

We fine-mapped eQTL causal variants in the LCL expression data within a window of +/-200kb of each CRM using a Bayesian stochastic search method that allows for multiple causal variants, GUESSFM (https://github.com/chr1swallace/GUESSFM) (74). This requires a prior on the number of causal variants per region, which we set as Bin(n,2/n) where n is the number of variants in the fine mapping window. This setting gives a prior expectation of 2 causal variants per region but allows all values from 0 to n. We visually checked traces to ensure the MCMC samples had converged. Raw GUESSFM data have been uploaded to OSF (https://osf.io/e5vsh/).

To estimate the proportion of possibly causal eQTLs identified by GUESSFM (marginal posterior probability of inclusion [mppi]>> 0.001) among the TF-binding affinity variants showing the strongest eQTL signal per CRM ("test SNPs"), we compared it with the same proportion obtained for "random SNPs". The "random SNPs" were sampled from the same +/-200kb windows around CRMs, matching the distribution of their minor allele frequencies to that across the "test SNPs".

Causal variant colocolisation analysis

An association between an epromoter variant and the expression of both a proximal and a distal gene may indicate that this variant is causal for the expression of both genes. However, the same association may arise from distinct causal variants for each gene that are in LD with each other and are tagged by the same epromoter variant. To differentiate between these situations, we used the Bayesian colocalisation technique coloc (75). Coloc evaluates the posterior probabilities of five mutually exclusive hypotheses: no association of any variant in the region with either trait (H0), association with first trait but not the second (H1), association with second trait but not the first (H2), two separate causal variants (H3), and, finally, a unique shared causal variant (H4). Coloc assumes at most one casual variant per locus. To mitigate this limitation, where there was evidence for multiple causal variants, we tested for colocalisation between all pairs of signals for each gene, by conditioning out the other signals. Coloc has also been originally designed for testing two sets of associations measured on different individuals. Therefore, before running it on the data measured in the same individuals (i.e., the expression of the proximal and distal gene across the 359 CEU LCLs) we confirmed by simulation that for a quantitative trait the results appear robust to correlated errors (Figure S1).

RESULTS

An atlas of CRMs with predicted variation in TF binding affinity in LCLs

We used the ChIP-seq binding profiles of 52 TFs profiled by the ENCODE project (4) in GM12878 LCL to define 128,766 CRMs in these cells, merging across overlapping ChIP regions for multiple TFs (Figure 1). Just over half (55%) of CRMs defined this way were bound by more than a single TF. For 41/52 TFs with known PWMs, we then used a biophysical model (39) to estimate their binding affinity to each allele of each CRM in GM12878, pooling information across multiple PWMs for the same TF where available (see Materials and Methods). To enable the comparison of binding affinities between different TFs, we expressed them relative to the respective 'background' affinities using an approach based on the generalised extreme value distribution (40) (see Materials and Methods for details).

We next asked how natural genetic variation at CRMs affects their TF-binding affinity. For this, we took advantage of the genotypes of an additional 358 LCLs also derived from European-ancestry individuals that are available from the 1000 Genomes project (60). We then calculated a TF affinity log-ratio between each alternative haplotype and the highest-affinity haplotype of GM12878 (Figure 1; see Materials and Methods). Overall, 38,804 CRMs had one or more alternative haplotypes with predicted changes in binding affinity for at least one TFs (affinity log-ratios ranging between -12.9

and 13.17). We have made the full atlas of TF-binding CRM variants publicly available at https://osf.io/fa4u7.

CRMs showing TF-binding variation are enriched for chromatin accessibility variants

TF binding is known to associate with increased chromatin accessibility. Therefore, to validate our predicted changes in TF-binding affinity, we took advantage of a published study (67) that profiled chromatin accessibility across 70 LCLs using DNase-seq and identified ~9,000 significant associations between DNase-seq signal and genotype ("DNase I sensitivity QTLs", dsQTLs). If our predicted TF affinity variants reflected real changes in TF binding affinity, we would expect them to show enrichment at regions of differential chromatin accessibility. To verify this, we quantified enrichment of differential chromatin accessibility at sets of CRMs showing predicted TF affinity variation above successively larger thresholds. As can be seen from Figure 2, CRMs with non-zero differences in TF-binding affinity across LCLs showed a significant enrichment at differential DNAse I sensitivity regions compared with a matched random set of CRMs (permutation test $p<0.001$, see Materials and Methods for details). Moreover, this enrichment increased with the magnitude of the predicted affinity change (Figure 2). These results provide direct functional evidence that our approach adequately predicts changes in TF binding associated with genetic variation. It should be noted that the observed overlap in absolute terms is likely underestimated, due to the relatively limited DNase-seq sequencing depth and sample size in the available dataset (67).

Variation in TF binding affinity at CRMs associates with target gene expression

To identify quantitative associations between TF binding variation at CRMs and the expression of their target genes, we used genome-wide gene expression data from the GEUVADIS project (62) that included 358/359 of the LCLs used in our analysis (with the exception of GM12878). In contrast to traditional eQTL testing, here we devised an approach that prioritises TF-binding variants and their putative target genes *a priori* and performs testing at the CRM level. In total, we selected 3,285 CRMs with predicted variation in the binding for at least one TF (log-ratio > 0.3). We then tested the association of each CRM haplotype with the expression levels of their target genes defined on the basis of 3D interactions or close spatial proximity (within 9kb; see Materials and Methods). As evidence of 3D promoter-CRM interactions, we used high-resolution Promoter Capture Hi-C (PCHi-C) data in GM12878 cells (49, 61). The highly reduced search space has enabled testing for associations at the gene level, with all CRMs targeting the same gene and showing TF-binding variation included into the regression model (see Materials and Methods). This approach identified 245 "eGenes" with significant associations between predicted TF-binding affinity at CRMs and gene expression (16% of 1530 genes tested, at 10% FDR; Table S1). In total, 161 "proximal" (within 9 kb) and 101 "distal" TF-CRM affinity variants (with contacts detected by PCHi-C) were found to underlie these associations, corresponding to 26% and 6% of all variants tested, respectively (t-test p-value <0.05; Table S1). Figure 3 shows an example of the detected association between the expression of *KLF6* and variation in the binding affinity of BATF transcription factor at a distal CRM that is located 88 kb away from

KLF6 promoter and contacts it in 3D according to PCHi-C (gene-level FDR=1.21×10^{-2}, BATF variant p-value=5.16×10^{-4}, effect size=0.26). Individuals homozygous for the high-affinity BATF binding allele showed the lowest levels of *KLF6* expression, while those homozygous for the low-affinity BATF binding alleles showed the highest levels (Figure 3). This suggests that BATF acts as a negative regulator of *KLF6* expression, consistent with its known role as a repressor of AP-1-dependent transcriptional activity (76).

A total of 420/1530 genes (27%) were linked with multiple predicted TF-binding variants (either for different TFs bound at the same CRM or at different CRMs). For 16 of these genes, we detected significant associations between more than one such variant and the expression level. One example is the nuclear receptor gene *NR2F6* whose expression significantly associated with predicted variation in the binding affinities of SMC3 and SRF to distal CRMs located, respectively, 41 kb and 19 kb away (Figure 4; gene-level FDR = 4.06×10^{-7}, SMC3 effect size=0.26, p-value=3×10^{-4}; SRF effect size=0.61, p-value=1.19×10^{-7}).

Owing to the *a priori* prioritisation of variants for association testing in our approach (i.e., testing only variants predicted to impact TF binding), we carried out far fewer association tests than in a standard eQTL analysis, thus reducing the multiple testing burden and increasing sensitivity. We therefore asked if we were able to detect additional associations compared with those reported for a standard eQTL analysis performed by the GEUVADIS project (note that this analysis also used an additional 103 LCLs not included in our study, which were either of non-European ancestry or not genotyped in 1000Genomes). To compare our CRM-based association results to GEUVADIS eQTL SNPs, we identified the SNP causing the largest change in affinity for the respective TF at each CRM (192 eQTL SNPs in total at 5% FDR to match the FDR level used by GEUVADIS). Of these, 78 SNPs (42%) were detected as significant by GEUVADIS. Therefore, the remaining 114/192 (58%) eQTL SNPs identified in our approach corresponded to not previously reported associations.

Threshold-free testing based on TF-binding affinities reveals further expression associations

The analysis above was performed broadly within the conventional paradigm of eQTL testing, whereby expression was compared across three diploid genotypes (two homozygous and one heterozygous), except that these genotypes corresponded to cases whereby variation was predicted to appreciably disrupt TF binding based on a predefined threshold (we shall refer to this approach as “thresholded”). However, since TF-binding affinity haplotypes were defined at the CRM level, more than two alleles were commonly observed per CRM (in 12-100% cases depending on the TF). In the thresholded approach, we pooled multiple alleles into either “high-affinity” or “low-affinity” haplotypes and disregarded outliers (see Materials and Methods). We reasoned, however, that is it also possible to regress gene expression against normalised TF-binding affinities directly without thresholding and haplotype pooling, leading to increased precision and sensitivity of association testing. As expected, this threshold-free approach revealed a considerably larger number of genes significantly associated with CRM affinity variants (1033 at 10% FDR compared with 245 detected in the “thresholded” approach above).

One challenge arising in the threshold-free approach is that it leads to many more binding variants tested for each gene (both within the same CRM and across CRMs) that are often in linkage disequilibrium (LD) with each other, leading to collinearity in the regression models. Therefore, to detect significant associations at CRM level, we performed elastic net regression for each of the 895/1033 identified eGenes that were targeted by multiple CRMs with predicted TF binding variants. To ascertain the significance of regression coefficients in elastic net regression, we used a covariance test for adaptive linear models (73), identifying 1328 significant CRM-gene associations for the 895 eGenes tested (Table S2; see Materials and Methods for details). One example of a newly identified association is between a nucleotide transporter gene *SLC29A3* and the binding affinity of SIN3A at a CRM overlapping with the TSS of *SLC29A3* (gene-level FDR=1.60×10^{-4}). Five alternative SIN3A binding affinity haplotypes were observed across the 358 LCLs, with log fold-changes in affinity for SIN3A ranging from -0.037 to 0.001 (elastic net effect size=-0.14, p-value ~ 0; Figure 5A).

TF-binding affinity variants are highly enriched for fine-mapped causal eQTLs

We asked what proportion of TF-binding variants showing association in our analysis could be fine-mapped as causal purely based on the pattern of association signals in their vicinity, without *a priori* prioritisation and pooling of variants per CRM. To this end, we supplied genotype information for +/-200 kb windows around the CRMs with detected associations and the respective gene expression data to GUESSFM, a Bayesian fine-mapping approach that accounts for possible multiple causal variants per locus (74). GUESSFM identified at least one causal variant in ~38% of the analysed CRMs (1807/4718); associations in the remaining CRMs likely could not be fine-mapped due to a lack of statistical power. In ~30% (548/1807) of CRMs with successful fine-mapping, the TF binding variant showing the strongest association per CRM was ranked as possibly causal (marginal posterior probability of inclusion [mppi] >> 0.001), and in the majority of such cases (477/548) this variant was also ranked by GUESSFM among the top five highest-scoring variants in the window (Table S3 and Fig 5B and C for examples). In contrast, just 2.6% (48/1807) random variants within the same windows (matched by allele frequency) were detected as potentially causal by GUESSFM, corresponding to a very significant enrichment of fine-mapped variants for those affecting TF binding (Fisher test $p=10^{-126}$).

Many CRMs associated with distal gene expression show features of epromoters

We noted that a large number of distal CRMs showing association between TF binding affinity and target gene expression (224 CRMs, 243 TF-CRM variants; Table S4) and connecting to the distal gene promoters in 3D based on PCHi-C also mapped in close proximity (within 200 bp) of the TSS of either one or more other genes (165 and 59 CRMs, respectively, and 284 eGenes; note that the number of eGenes is greater than that of CRMs due to some CRMs mapping in close proximity of multiple TSSs). The absolute majority (87%) of these CRMs localised within chromatin segments with the characteristic features of gene promoters (Figure 6A). Taken together, this suggested that some promoter regions might act as distal regulatory regions of other genes, whose promoters they physically contact. This class of CRMs with dual

promoter and activity were independently identified in two recent studies (63, 64). We shall follow Dao et al. (63) in referring to these CRMs as 'epromoters'.

Most genes located in the immediate vicinity of the identified epromoters were appreciably expressed in LCLs (232/284, 82%). However, TF-binding variation at nearly two-thirds of epromoters whose proximal gene was expressed (139 variants, 64.7%; see Table S2) showed detectable association with a distal gene alone in independent tests (assessed with the threshold-free approach). For example, variation in ELF1 binding affinity at a CRM that shows promoter-associated chromatin marks and localises within 200 bp from the TSS of *CLOCK* gene does not affect *CLOCK* expression. Instead, it associates with expression of *SRD5A3* located 198 kb away, whose promoter it contacts in 3D as detected by PCHi-C (Figure 6B; *SRD5A3:* gene-level FDR = 3.33×10^{-21}, ELF1 elastic net p-value=0, ELF1 elastic net beta = -0.21; *CLOCK:* gene-level FDR = 0.88).

The remaining 76 TF-epromoter CRM variants showed associations between with the expression levels of both distal and proximal genes. To obtain formal evidence that these associations were indeed driven by the same variant and not by different variants in LD with each other, we used colocalisation analysis (75), while accounting for multiple independent associations (see Materials and Methods). We submitted to this analysis the most tractable subset of 7 epromoters, for which the association of the respective TF-binding variant with distal gene expression was independently confirmed by fine-mapping (GUESSFM mppi>0.001). At 6/7 analysed epromoters, we found prevailing evidence of shared association signals for both proximal and distal gene (p_{H4}>0.66; Table S5). An example of such high-confidence shared signal is variation in EGR1 binding affinity in the epromoter of lncRNA *RP11-71F7.7* that associates with the expression of both *RP11-71F7.7* and another gene, *IRF2BPL* (Fig. 6C). The promoters of these two genes, transcribed in a convergent orientation, are approximately 69 kb apart and contact each other in 3D as detected by PCHi-C.

Taken together, our findings confirm long-range transcriptional regulation by epromoters and suggest that regulatory variants within these elements may have both shared and independent effects on the expression of their proximal and distal target genes.

DISCUSSION

In this study we have generated an atlas of CRM variants predicted to affect TF binding in LCLs and established their associations with the expression of their target genes identified on the basis of 3D chromosomal interactions or immediate spatial proximity. Notably, we found that many TF binding variants showing associations with distal gene expression localise to the promoters of other genes, providing additional support for the recently characterised class of "epromoter" regulatory elements (63, 64).

We previously reported a collection of TF binding variants for ENCODE-profiled TFs based on 1000 Pilot data (179 individuals) (32). The atlas of binding variants generated in this study is based on a more than two-fold larger sample of EUR individuals from 1000 Genomes release. Importantly, in this study we also used a biophysical model (39) that aggregates TF binding affinities across the whole CRM to increase sensitivity. This in contrast to our previous work (32) and other published

resources such as Haploreg (77) and SNP2TFBS (78) that is based on detecting individual PWM motif matches within each region of interest. Our current approach, however, still relies on PWMs to estimate affinities at each sequence window. While this is a straightforward strategy that produces readily interpretable results, it has several limitations. First, conventional PWMs may not adequately describe the binding preferences of some TFs, leading to either overfitted or overly degenerate models. For such TFs, k-mer-based approaches to modelling sequence consensus (79) or models incorporating dependencies between motif positions (80–82) may be more appropriate. In addition, it was recently shown that modelling DNA shape may aid binding prediction for some TFs with seemingly degenerate sequence preferences (83). Moreover, even for TFs whose binding can be modelled by PWMs reasonably well, variants affecting their binding in vivo may be based outside of the immediate binding consensus, possibly due to cooperative effects (25, 26, 30). At the expense of interpretability, this problem can be mitigated by using machine learning models such as DeFine (84) and DeepSEA (85), with the latter approach employed in a recent study investigating impact of regulatory variants on gene expression (86). Alternatively, effects on TF binding can be obtained directly from allele-specific or multi-individual TF binding data (25–28, 30, 87).

To identify the putative target genes of remote regulatory variants, we have capitalised on Promoter Capture Hi-C (PCHi-C) - a high-resolution technique for global chromosomal interaction profiling (48–50). Our approach is based on the classic model of long-range transcriptional control that necessitates physical contacts between enhancer elements and their target promoters through DNA looping (44, 45), which has been validated by a number of approaches, most recently by in vivo imaging (88–90). However, recent evidence suggests that at least some regulatory regions may exert action on their target promoters without coming into proximity with them (91, 92). Alternative methods of assigning the target genes of regulatory regions, such as those based on correlated chromatin activity of promoters and enhancers (20, 93–96) may account for the effects of these regions. Emerging high-throughput functional screens of cis-regulatory activity (8, 9, 64, 97) are proving to be instrumental for direct functional validation of enhancer targets identified with either approach.

In our study, we restricted the analysis to CRMs with predicted variation in TF binding and their putative proximal and distal target genes, which has considerably reduced multiple testing burden and made it possible to combine the effects of multiple polymorphisms within the same CRM based on TF binding models. This analysis framework has yielded highly sensitive and interpretable associations for pre-selected loci. However, owing to its selective nature, it is by no means a substitute to conventional association testing. Theoretically, the effects of nucleotide variants on TF binding can also be incorporated as a prior in global association analyses such as fgwas (98), and have already been used in eQTL fine-mapping (99). We have previously confirmed that promoter-interacting regions are strongly enriched for eQTLs of the physically connected distal genes (51). An optimal statistical framework for incorporating 3D interaction data into eQTL testing is, however, yet to be established.

Our finding that polymorphic TFBSs at distal CRMs show gene expression associations less frequently compared with proximal regions is consistent with the high degree of redundancy of long-range regulatory elements (5–7, 100, 101). Predicting the

extent of buffering of regulatory variation for a given CRM with a reasonable precision is an important problem that is currently highly challenging due to the sheer number of parameters and the relatively small sample sizes of multi-individual expression datasets. Profiling gene expression in the emerging much larger genotype panels such as UK10K (102) may provide opportunities for addressing this question.

We observed that a large proportion of CRMs showing associations with the expression of physically connected distal genes localised in the promoter regions of other genes. This provides additional evidence to the recently characterised class of "epromoters": elements with a dual proximal and distal activity that were discovered on the large scale using high-throughput reporter and CRISPR knockout screens (63–65). Empirically, chromosomal interactions between epromoter CRMs and their distal targets fall into the category of promoter-promoter interactions. Until recently, these interactions have been viewed primarily in the context of coordinated gene activation or repression (103–105), such as that observed in Hox and histone clusters (103, 106). That some promoter-promoter contacts reflect epromoter-distal target gene relationships suggests that these contacts may show functionally and possibly even structurally distinct properties.

We show that TF-binding variation at epromoters may or may not co-associate with the expression of both proximal and distal genes at the same time. Shared association is consistent with the findings from massively parallel reporter assays that the same sequences are often involved in mediating both promoter and enhancer activity in vitro (107). It is possible that some non-shared effects observed in our study in vivo are underpinned by the role of the affected TFs in mediating long-range contacts. Additionally, epromoter elements may show different degrees of redundancy with respect to the proximal and distal target gene.

Overall, our analysis demonstrates the potential of model-based prioritisation and pooling of variants a priori of testing for increasing the sensitivity of identifying individual associations and revealing their shared biological properties.

DATA AVAILABILITY

The list of the detected TF affinity CRM variants, the full data on CRM variant – gene expression associations and the raw output of GUESSFM fine-mapping have been uploaded to Open Science Framework (https://osf.io/fa4u7/).

SUPPLEMENTARY DATA

Figure S1. Suitability of the association signal colocalisation algorithm for analysing pairs of signals within the same dataset.

Table S1. Significant associations between TF-binding affinity CRM variants and target gene expression identified with the thresholded approach.

Table S2. Significant associations between TF-binding affinity CRM variants and target gene expression identified with the threshold-free approach.

Table S3. Posterior probabilities of the prioritised TF-binding affinity eQTLs being causal estimated by GUESSFM fine-mapping algorithm.

Table S4. A list of epromoters with detected TF-binding affinity variation.

Table S5. Association signal colocalisation analysis for the proximal and distal target genes of selected epromoters.

ACKNOWLEDGEMENTS

The authors wish to thank Paula Freire-Pritchett, Hashem Koohy, Jonathan Cairns and Simon Andrews for advice and technical assistance, and all members of MS lab for helpful discussions.

FUNDING

JM was supported by BBSRC DTP studentship, NG and CW are supported by the Wellcome Trust (WT107881) and CW by the MRC (MC_UU_00002/4). MS acknowledges core support from BBSRC and MRC.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

1. Suryamohan,K. and Halfon,M.S. (2015) Identifying transcriptional cis-regulatory modules in animal genomes. *Wiley Interdiscip. Rev. Dev. Biol.*, **4**, 59–84.

2. Whitaker,J.W., Nguyen,T.T., Zhu,Y., Wildberg,A. and Wang,W. (2015) Computational schemes for the prediction and annotation of enhancers from epigenomic assays. *Methods*, **72**, 86–94.

3. Mei,S., Qin,Q., Wu,Q., Sun,H., Zheng,R., Zang,C., Zhu,M., Wu,J., Shi,X., Taing,L., *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.*, **45**, D658–D662.

4. The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

5. Spivakov,M. (2014) Spurious transcription factor binding: non-functional or genetically redundant? *Bioessays*, **36**, 798–806.

6. Osterwalder,M., Barozzi,I., Tissières,V., Fukuda-Yuzawa,Y., Mannion,B.J., Afzal,S.Y., Lee,E.A., Zhu,Y., Plajzer-Frick,I., Pickle,C.S., *et al.* (2018) Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*, **554**, 239–243.

7. Frankel,N., Davis,G.K., Vargas,D., Wang,S., Payre,F. and Stern,D.L. (2010) Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, **466**, 490–493.

8. Diao,Y., Li,B., Meng,Z., Jung,I., Lee,A.Y., Dixon,J., Maliskova,L., Guan,K.-L., Shen,Y. and Ren,B. (2016) A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. *Genome Res.*, **26**, 397–405.

9. Fulco,C.P., Munschauer,M., Anyoha,R., Munson,G., Grossman,S.R., Perez,E.M., Kane,M., Cleary,B., Lander,E.S. and Engreitz,J.M. (2016) Systematic mapping of functional enhancer-promoter connections with CRISPR

interference. *Science*, **354**, 769–773.

10. Yao,L., Berman,B.P. and Farnham,P.J. (2015) Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. *Crit. Rev. Biochem. Mol. Biol.*, **50**, 550–573.

11. Gilad,Y., Rifkin,S.A. and Pritchard,J.K. (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.*, **24**, 408–415.

12. Majewski,J. and Pastinen,T. (2011) The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.*, **27**, 72–79.

13. Stranger,B.E. and Raj,T. (2013) Genetics of human gene expression. *Curr. Opin. Genet. Dev.*, **23**, 627–634.

14. Lappalainen,T. (2015) Functional genomics bridges the gap between quantitative genetics and molecular biology. *Genome Res.*, **25**, 1427–1431.

15. Lu Tian, Tian,L., Quitadamo,A., Lin,F. and Shi,X. (2014) Methods for population-based eQTL analysis in human genetics. *Tsinghua Sci. Technol.*, **19**, 624–634.

16. Battle,A. and Montgomery,S.B. (2014) Determining causality and consequence of expression quantitative trait loci. *Hum. Genet.*, **133**, 727–735.

17. Yashiro-Ohtani,Y., Wang,H., Zang,C., Arnett,K.L., Bailis,W., Ho,Y., Knoechel,B., Lanauze,C., Louis,L., Forsyth,K.S., *et al.* (2014) Long-range enhancer activity determines Myc sensitivity to Notch inhibitors in T cell leukemia. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, E4946–E4953.

18. Lettice,L.A., Heaney,S.J.H., Purdie,L.A., Li,L., de Beer,P., Oostra,B.A., Goode,D., Elgar,G., Hill,R.E. and de Graaff,E. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.*, **12**, 1725–1735.

19. Zhou,H.Y., Katsman,Y., Dhaliwal,N.K., Davidson,S., Macpherson,N.N., Sakthidevi,M., Collura,F. and Mitchell,J.A. (2014) A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes Dev.*, **28**, 2699–2711.

20. Corradin,O., Saiakhova,A., Akhtar-Zaidi,B., Myeroff,L., Willis,J., Cowper-Sal lari,R., Lupien,M., Markowitz,S. and Scacheri,P.C. (2014) Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.*, **24**, 1–13.

21. Spitz,F. and Furlong,E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, **13**, 613–626.

22. Long,H.K., Prescott,S.L. and Wysocka,J. (2016) Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell*, **167**, 1170–1187.

23. Gonen,N., Futtner,C.R., Wood,S., Garcia-Moreno,S.A., Salamone,I.M., Samson,S.C., Sekido,R., Poulat,F., Maatouk,D.M. and Lovell-Badge,R. (2018) Sex reversal following deletion of a single distal enhancer of. *Science*, **360**, 1469–1473.

24. Miguel-Escalada,I., Pasquali,L. and Ferrer,J. (2015) Transcriptional enhancers: functional insights and role in human disease. *Curr. Opin. Genet. Dev.*, **33**, 71–76.

25. Gallone,G., Haerty,W., Disanto,G., Ramagopalan,S.V., Ponting,C.P. and Berlanga-Taylor,A.J. (2017) Identification of genetic variants affecting vitamin D receptor binding and associations with autoimmune disease. *Hum. Mol. Genet.*, **26**, 2164–2176.

26. Ding,Z., Ni,Y., Timmer,S.W., Lee,B.-K., Battenhouse,A., Louzada,S., Yang,F., Dunham,I., Crawford,G.E., Lieb,J.D., *et al.* (2014) Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLoS Genet.*, **10**, e1004798.

27. Kasowski,M., Kyriazopoulou-Panagiotopoulou,S., Grubert,F., Zaugg,J.B., Kundaje,A., Liu,Y., Boyle,A.P., Zhang,Q.C., Zakharia,F., Spacek,D.V., *et al.* (2013) Extensive variation in chromatin states across humans. *Science*, **342**, 750–752.

28. Kilpinen,H., Waszak,S.M., Gschwind,A.R., Raghav,S.K., Witwicki,R.M., Orioli,A., Migliavacca,E., Wiederkehr,M., Gutierrez-Arcelus,M., Panousis,N.I., *et al.* (2013) Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*, **342**, 744–747.

29. Maurano,M.T., Wang,H., Kutyavin,T. and Stamatoyannopoulos,J.A. (2012) Widespread site-dependent buffering of human regulatory polymorphism. *PLoS Genet.*, **8**, e1002599.

30. Kasowski,M., Grubert,F., Heffelfinger,C., Hariharan,M., Asabere,A., Waszak,S.M., Habegger,L., Rozowsky,J., Shi,M., Urban,A.E., *et al.* (2010) Variation in transcription factor binding among humans. *Science*, **328**, 232–235.

31. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

32. Spivakov,M., Akhtar,J., Kheradpour,P., Beal,K., Girardot,C., Koscielny,G., Herrero,J., Kellis,M., Furlong,E.E.M. and Birney,E. (2012) Analysis of variation at transcription factor binding sites in Drosophila and humans. *Genome Biol.*, **13**, R49.

33. Kim,J., He,X. and Sinha,S. (2009) Evolution of regulatory sequences in 12 Drosophila species. *PLoS Genet.*, **5**, e1000330.

34. Chen,K., van Nimwegen,E., Rajewsky,N. and Siegal,M.L. (2010) Correlating gene expression variation with cis-regulatory polymorphism in Saccharomyces cerevisiae. *Genome Biol. Evol.*, **2**, 697–707.

35. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

36. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

37. Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–750.

38. Ruan,S. and Stormo,G.D. (2017) Inherent limitations of probabilistic models for protein-DNA binding specificity. *PLoS Comput. Biol.*, **13**, e1005638.

39. Roider,H.G., Kanhere,A., Manke,T. and Vingron,M. (2007) Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, **23**, 134–141.

40. Manke,T., Roider,H.G. and Vingron,M. (2008) Statistical modeling of transcription factor binding affinities predicts regulatory interactions. *PLoS Comput. Biol.*, **4**, e1000039.

41. Ramos,A.I. and Barolo,S. (2013) Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **368**, 20130018.

42. Farley,E.K., Olson,K.M., Zhang,W., Brandt,A.J., Rokhsar,D.S. and Levine,M.S. (2015) Suboptimization of developmental enhancers. *Science*, **350**, 325–328.

43. He,X., Duque,T.S.P.C. and Sinha,S. (2012) Evolutionary origins of transcription factor binding site clusters. *Mol. Biol. Evol.*, **29**, 1059–1070.

44. Krivega,I. and Dean,A. (2012) Enhancer and promoter interactions-long distance calls. *Curr. Opin. Genet. Dev.*, **22**, 79–85.

45. Ong,C.-T. and Corces,V.G. (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.*, **12**, 283–293.

46. Maeso,I., Acemel,R.D. and Gómez-Skarmeta,J.L. (2017) Cis-regulatory landscapes in development and evolution. *Curr. Opin. Genet. Dev.*, **43**, 17–22.

47. Schmitt,A.D., Hu,M. and Ren,B. (2016) Genome-wide mapping and analysis of chromosome architecture. *Nat. Rev. Mol. Cell Biol.*, **17**, 743–755.

48. Schoenfelder,S., Furlan-Magaril,M., Mifsud,B., Tavares-Cadete,F., Sugar,R., Javierre,B.-M., Nagano,T., Katsman,Y., Sakthidevi,M., Wingett,S.W., *et al.* (2015) The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.*, **25**, 582–597.

49. Mifsud,B., Tavares-Cadete,F., Young,A.N., Sugar,R., Schoenfelder,S., Ferreira,L., Wingett,S.W., Andrews,S., Grey,W., Ewels,P.A., *et al.* (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.*, **47**, 598–606.

50. Sahlén,P., Abdullayev,I., Ramsköld,D., Matskova,L., Rilakovic,N., Lötstedt,B., Albert,T.J., Lundeberg,J. and Sandberg,R. (2015) Genome-wide mapping of promoter-anchored interactions with close to single-enhancer

resolution. *Genome Biol.*, **16**, 156.

51. Javierre,B.M., Burren,O.S., Wilder,S.P., Kreuzhuber,R., Hill,S.M., Sewitz,S., Cairns,J., Wingett,S.W., Várnai,C., Thiecke,M.J., *et al.* (2016) Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, **167**, 1369–1384.e19.

52. Choy,M.-K., Javierre,B.M., Williams,S.G., Baross,S.L., Liu,Y., Wingett,S.W., Akbarov,A., Wallace,C., Freire-Pritchett,P., Rugg-Gunn,P.J., *et al.* (2018) Promoter interactome of human embryonic stem cell-derived cardiomyocytes connects GWAS regions to cardiac gene networks. *Nat. Commun.*, **9**, 2526.

53. Burren,O.S., Rubio García,A., Javierre,B.-M., Rainbow,D.B., Cairns,J., Cooper,N.J., Lambourne,J.J., Schofield,E., Castro Dopico,X., Ferreira,R.C., *et al.* (2017) Chromosome contacts in activated T cells identify autoimmune disease candidate genes. *Genome Biol.*, **18**, 165.

54. Petersen,R., Lambourne,J.J., Javierre,B.M., Grassi,L., Kreuzhuber,R., Ruklisa,D., Rosa,I.M., Tomé,A.R., Elding,H., van Geffen,J.P., *et al.* (2017) Platelet function is modified by common sequence variation in megakaryocyte super enhancers. *Nat. Commun.*, **8**, 16058.

55. Jäger,R., Migliorini,G., Henrion,M., Kandaswamy,R., Speedy,H.E., Heindl,A., Whiffin,N., Carnicer,M.J., Broome,L., Dryden,N., *et al.* (2015) Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat. Commun.*, **6**, 6178.

56. Dryden,N.H., Broome,L.R., Dudbridge,F., Johnson,N., Orr,N., Schoenfelder,S., Nagano,T., Andrews,S., Wingett,S., Kozarewa,I., *et al.* (2014) Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res.*, **24**, 1854–1868.

57. McGovern,A., Schoenfelder,S., Martin,P., Massey,J., Duffus,K., Plant,D., Yarwood,A., Pratt,A.G., Anderson,A.E., Isaacs,J.D., *et al.* (2016) Capture Hi-C identifies a novel causal gene, IL20RA, in the pan-autoimmune genetic susceptibility region 6q23. *Genome Biol.*, **17**, 212.

58. Martin,P., McGovern,A., Massey,J., Schoenfelder,S., Duffus,K., Yarwood,A., Barton,A., Worthington,J., Fraser,P., Eyre,S., *et al.* (2016) Identifying Causal Genes at the Multiple Sclerosis Associated Region 6q23 Using Capture Hi-C. *PLoS One*, **11**, e0166923.

59. Baxter,J.S., Leavy,O.C., Dryden,N.H., Maguire,S., Johnson,N., Fedele,V., Simigdala,N., Martin,L.-A., Andrews,S., Wingett,S.W., *et al.* (2018) Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. *Nat. Commun.*, **9**, 1028.

60. The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

61. Cairns,J., Freire-Pritchett,P., Wingett,S.W., Várnai,C., Dimond,A., Plagnol,V., Zerbino,D., Schoenfelder,S., Javierre,B.-M., Osborne,C., *et al.* (2016) CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.*, **17**, 127.

62. Lappalainen,T., Sammeth,M., Friedländer,M.R., 't Hoen,P.A.C., Monlong,J., Rivas,M.A., Gonzàlez-Porta,M., Kurbatova,N., Griebel,T., Ferreira,P.G., *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.

63. Dao,L.T.M., Galindo-Albarrán,A.O., Castro-Mondragon,J.A., Andrieu-Soler,C., Medina-Rivera,A., Souaid,C., Charbonnier,G., Griffon,A., Vanhille,L., Stephen,T., *et al.* (2017) Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat. Genet.*, **49**, 1073–1081.

64. Diao,Y., Fang,R., Li,B., Meng,Z., Yu,J., Qiu,Y., Lin,K.C., Huang,H., Liu,T., Marina,R.J., *et al.* (2017) A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods*, **14**, 629–635.

65. Dao,L.T.M. and Spicuglia,S. (2018) Transcriptional regulation by promoters with enhancer function. *Transcription*, **9**, 307–314.

66. Kheradpour,P. and Kellis,M. (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.*, **42**, 2976–2987.

67. Degner,J.F., Pai,A.A., Pique-Regi,R., Veyrieras,J.-B., Gaffney,D.J., Pickrell,J.K., De Leon,S., Michelini,K., Lewellen,N., Crawford,G.E., *et al.* (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, **482**, 390–394.

68. Hinrichs,A.S. (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.*, **34**, D590–D598.

69. Durinck,S., Spellman,P.T., Birney,E. and Huber,W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, **4**, 1184–1191.

70. Stegle,O., Parts,L., Durbin,R. and Winn,J. (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.*, **6**, e1000770.

71. Kolesnikov,N., Hastings,E., Keays,M., Melnichuk,O., Tang,Y.A., Williams,E., Dylag,M., Kurbatova,N., Brandizi,M., Burdett,T., *et al.* (2015) ArrayExpress update--simplifying data submissions. *Nucleic Acids Res.*, **43**, D1113–6.

72. Aulchenko,Y.S., Ripke,S., Isaacs,A. and van Duijn,C.M. (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, **23**, 1294–1296.

73. Lockhart,R., Taylor,J., Tibshirani,R.J. and Tibshirani,R. (2014) A SIGNIFICANCE TEST FOR THE LASSO. *Ann. Stat.*, **42**, 413–468.

74. Wallace,C., Cutler,A.J., Pontikos,N., Pekalski,M.L., Burren,O.S., Cooper,J.D., García,A.R., Ferreira,R.C., Guo,H., Walker,N.M., *et al.* (2015) Dissection of a Complex Disease Susceptibility Region Using a Bayesian Stochastic Search Approach to Fine Mapping. *PLoS Genet.*, **11**, e1005272.

75. Giambartolomei,C., Vukcevic,D., Schadt,E.E., Franke,L., Hingorani,A.D., Wallace,C. and Plagnol,V. (2014) Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.*, **10**, e1004383.

76. Williams,K.L., Nanda,I., Lyons,G.E., Kuo,C.T., Schmid,M., Leiden,J.M., Kaplan,M.H. and Taparowsky,E.J. (2001) Characterization of murine BATF: a negative regulator of activator protein-1 activity in the thymus. *European Journal of Immunology*, **31**, 1620–1627.

77. Ward,L.D. and Kellis,M. (2016) HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.*, **44**, D877–81.

78. Kumar,S., Ambrosini,G. and Bucher,P. (2017) SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.*, **45**, D139–D144.

79. Turatsinze,J.-V., Thomas-Chollier,M., Defrance,M. and van Helden,J. (2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat. Protoc.*, **3**, 1578–1588.

80. Keilwagen,J. and Grau,J. (2015) Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Res.*, **43**, e119.

81. Kulakovskiy,I., Levitsky,V., Oshchepkov,D., Bryzgalov,L., Vorontsov,I. and Makeev,V. (2013) From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *J. Bioinform. Comput. Biol.*, **11**, 1340004.

82. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.

83. Rossi,M.J., Lai,W.K.M. and Pugh,B.F. (2018) Genome-wide determinants of sequence-specific DNA binding of general regulatory factors. *Genome Res.*, **28**, 497–508.

84. Wang,M., Tai,C., E,W. and Wei,L. (2018) DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. *Nucleic Acids Res.*, **46**, e69–e69.

85. Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, **12**, 931–934.

86. Shi,W., Fornes,O. and Wasserman,W.W. (2018) Gene expression models based on transcription factor binding events confer insight into functional cis-regulatory variants. *Bioinformatics*, 10.1093/bioinformatics/bty992.

87. Shi,W., Fornes,O., Mathelier,A. and Wasserman,W.W. (2016) Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Res.*, **44**, 10106–10116.

88. Chen,H., Levo,M., Barinov,L., Fujioka,M., Jaynes,J.B. and Gregor,T. (2018) Dynamic interplay between enhancer–promoter topology and gene activity. *Nat. Genet.*, **50**, 1296–1303.

89. Bartman,C.R., Hsu,S.C., Hsiung,C.C.-S., Raj,A. and Blobel,G.A. (2016) Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping. *Mol. Cell*, **62**, 237–247.

90. Fukaya,T., Lim,B. and Levine,M. (2016) Enhancer Control of Transcriptional Bursting. *Cell*, **166**, 358–368.

91. Benabdallah,N.S., Williamson,I., Illingworth,R.S., Boyle,S., Grimes,G.R., Therizols,P. and Bickmore,W. (2017) PARP mediated chromatin unfolding is coupled to long-range enhancer activation. *Genomics*.

92. Alexander,J.M., Guan,J., Huang,B., Lomvardas,S. and Weiner,O.D. (2018) Live-Cell Imaging Reveals Enhancer-dependent Sox2 Transcription in the Absence of Enhancer Proximity. *Genetics*.

93. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B., *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.

94. Malin,J., Aniba,M.R. and Hannenhalli,S. (2013) Enhancer networks revealed by correlated DNAse hypersensitivity states of enhancers. *Nucleic Acids Res.*, **41**, 6828–6838.

95. Pliner,H.A., Packer,J.S., McFaline-Figueroa,J.L., Cusanovich,D.A., Daza,R.M., Aghamirzaie,D., Srivatsan,S., Qiu,X., Jackson,D., Minkina,A., *et al.* (2018) Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell*, **71**, 858–871.e8.

96. Sheffield,N.C., Thurman,R.E., Song,L., Safi,A., Stamatoyannopoulos,J.A., Lenhard,B., Crawford,G.E. and Furey,T.S. (2013) Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res.*, **23**, 777–788.

97. Klein,J.C., Chen,W., Gasperini,M. and Shendure,J. (2018) Identifying Novel Enhancer Elements with CRISPR-Based Screens. *ACS Chem. Biol.*, **13**, 326–332.

98. Pickrell,J.K. (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.*, **94**, 559–573.

99. Wen,X., Luca,F. and Pique-Regi,R. (2015) Cross-population joint analysis of eQTLs: fine mapping and functional annotation. *PLoS Genet.*, **11**, e1005176.

100. Cannavò,E., Khoueiry,P., Garfield,D.A., Geeleher,P., Zichner,T., Gustafson,E.H., Ciglar,L., Korbel,J.O. and Furlong,E.E.M. (2016) Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Curr. Biol.*, **26**, 38–51.

101. Barolo,S. (2012) Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy. *Bioessays*, **34**, 135–141.

102. UK10K Consortium, Walter,K., Min,J.L., Huang,J., Crooks,L., Memari,Y., McCarthy,S., Perry,J.R.B., Xu,C., Futema,M., *et al.* (2015) The UK10K project identifies rare variants in health and disease. *Nature*, **526**, 82–90.

103. Schoenfelder,S., Sugar,R., Dimond,A., Javierre,B.-M., Armstrong,H., Mifsud,B., Dimitrova,E., Matheson,L., Tavares-Cadete,F., Furlan-Magaril,M., *et al.* (2015) Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat. Genet.*, **47**, 1179–1186.

104. Joshi,O., Wang,S.-Y., Kuznetsova,T., Atlasi,Y., Peng,T., Fabre,P.J., Habibi,E., Shaik,J., Saeed,S., Handoko,L., *et al.* (2015) Dynamic Reorganization of Extremely Long-Range Promoter-Promoter Interactions between Two States of Pluripotency. *Cell Stem Cell*, **17**, 748–757.

105. Li,G., Ruan,X., Auerbach,R.K., Sandhu,K.S., Zheng,M., Wang,P., Poh,H.M., Goh,Y., Lim,J., Zhang,J., *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**, 84–98.

106. Wang,Q., Sawyer,I.A., Sung,M.-H., Sturgill,D., Shevtsov,S.P., Pegoraro,G., Hakim,O., Baek,S., Hager,G.L. and Dundr,M. (2016) Cajal bodies are linked to genome conformation. *Nat. Commun.*, **7**, 10966.

107. Nguyen,T.A., Jones,R.D., Snavely,A.R., Pfenning,A.R., Kirchner,R., Hemberg,M. and Gray,J.M. (2016) High-throughput functional comparison of promoter and enhancer activities. *Genome Res.*, **26**, 1023–1033.

108. Ernst,J. and Kellis,M. (2017) Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.*, **12**, 2478–2492.

FIGURES AND LEGENDS

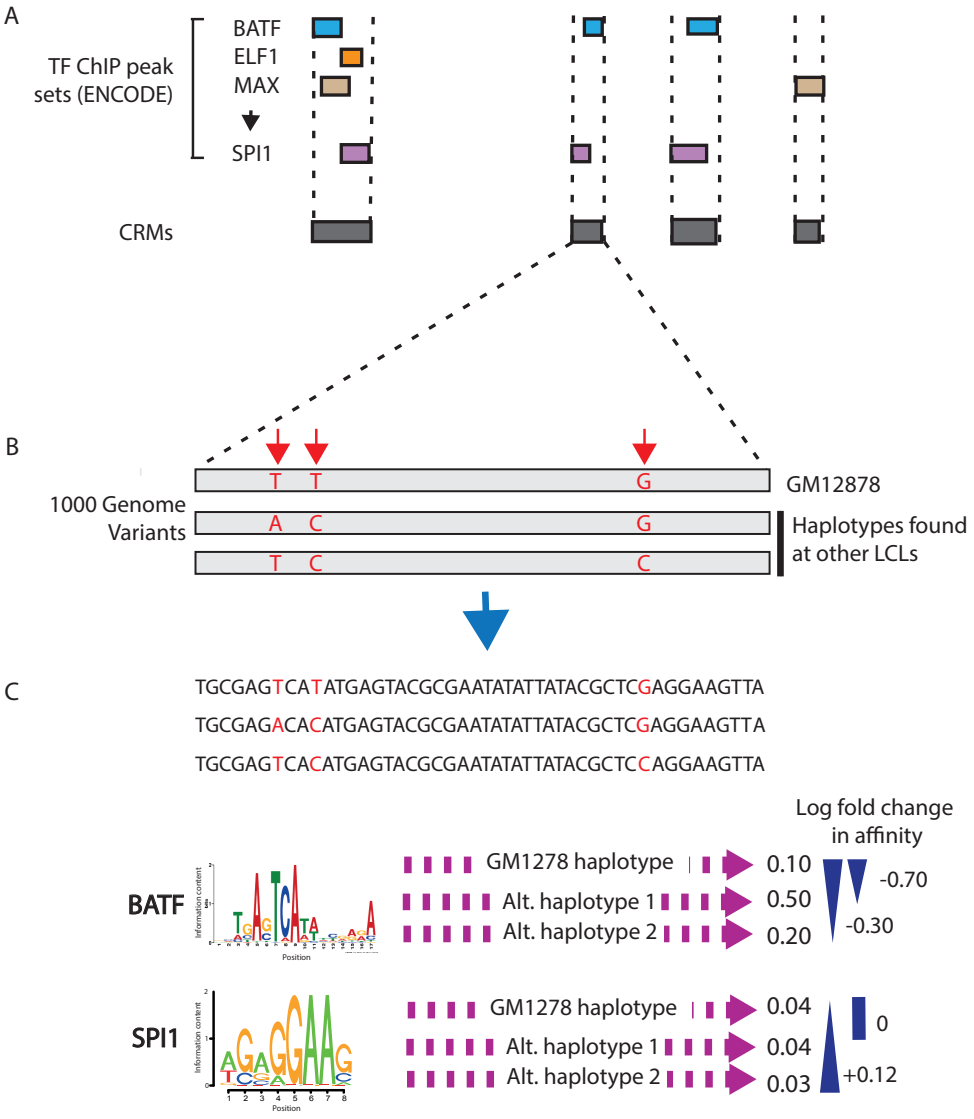

Figure 1. Definition of TF binding affinity variants. (A) TF ChIP-Seq data for 52 TFs profiled by the ENCODE project (4) in GM12878 are used to define CRMs, merging across overlapping ChIP regions for multiple TFs. Here ChIP seq regions for four TFs are depicted by different coloured rectangles. (B) Using variants from the 1000 Genomes Project, unique haplotypes for each CRM across the 359 LCLs are identified, including the haplotype/s of GM12878. (C) For each TF bound at the CRM in GM12878, ENCODE PWMs for the given TF are used to predict the normalised binding affinity for each unique haplotype (note that lower affinity values reflect higher affinities). The log fold affinity change between the highest affinity haplotype of GM12878 and each alternative haplotype is computed for each PWM and for each haplotype/TF the median log fold affinity change across all PWMs belonging to the given TF is taken (here only one PWM per TF is depicted).

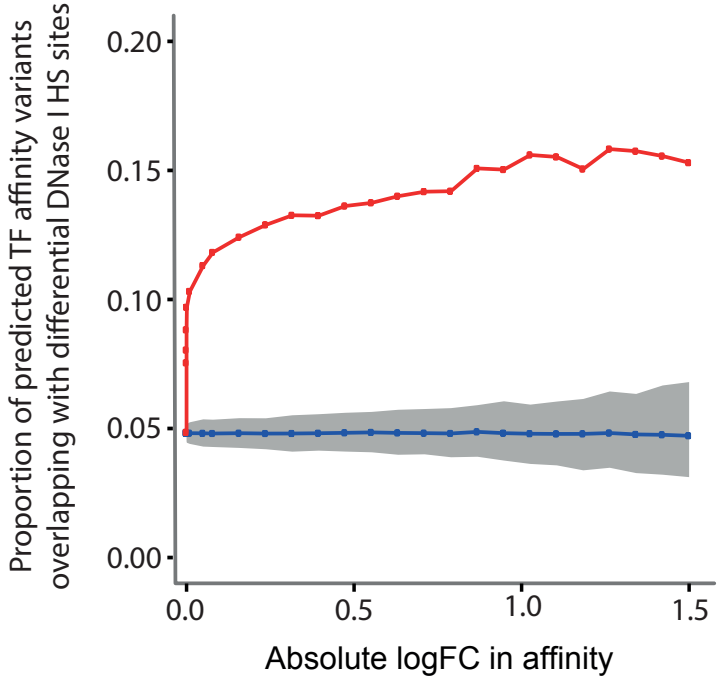

Figure 2. TF binding affinity variants are enriched at regions showing variation in DNase I hypersensitivity. The proportion of CRMs with log fold affinity changes over a range of thresholds that overlap with differential DNase I hypersensitivity (HS) sites (identified as dsQTLs in ref. (67)) are depicted by red squares. CRMs were filtered to those where the SNP driving the affinity change has a MAF > 5% in the 70 YRI individuals. The mean proportion of randomly sampled CRMs that overlap with differential DNase I HS sites across 1000 permutations are shown in blue, with the grey ribbon showing the range where 90% of the permuted values (number of overlaps) lie. For each threshold the control sets of CRMs were matched in sample size and "driving" SNP allele frequency distribution to the number and allele frequency distribution respectively of the predicted affinity variants over the corresponding threshold.

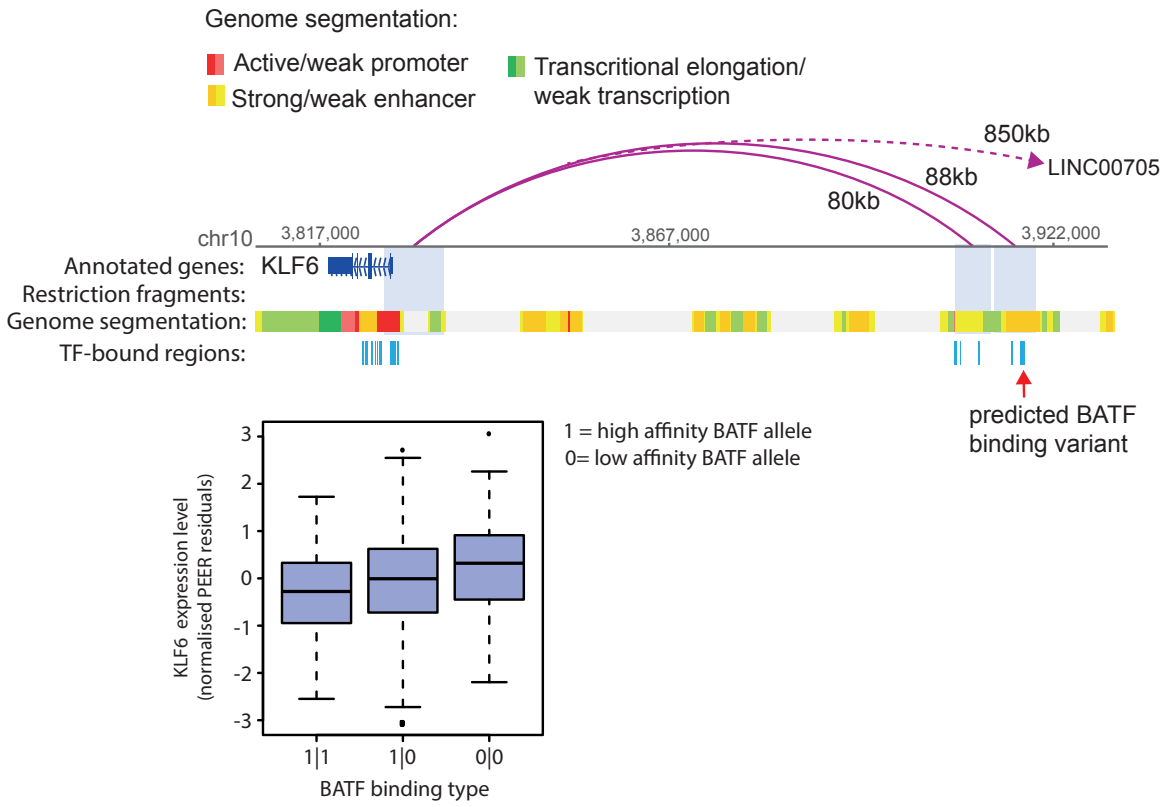

Figure 3. Example of association between a TF-binding affinity CRM variant and gene expression. (A) Genome browser representation of the distal interactions (pink arches) of *KLF6* promoter in the LCL GM12878, as detected by Promoter Capture Hi-C (49). Two out of the three fragments interacting with KLF6 are shown; the third fragment, which is located 850kb away from the KLF6 promoter and contains the gene LINC00705, was omitted due to space constraints. Genome segmentation tracks for GM12878 are shown (108). CRMs at the two distally interacting fragments and TSS-proximal window are depicted in azure blue. The far-right CRM, which interacts with the KLF6 promoter 88kb away, is predicted to impact BATF binding affinity across the 359 LCLs. (B). Boxplot showing the association in LCLs between mRNA levels (as measured with RNA-seq by the GEUVADIS consortium) and predicted BATF affinity CRM haplotype. KLF6 expression is significantly associated with BATF binding type (gene level FDR adjusted p-value=1.21×10^{-2}, BATF variant p-value= 5.16×10^{-4}, effect size=0.26).

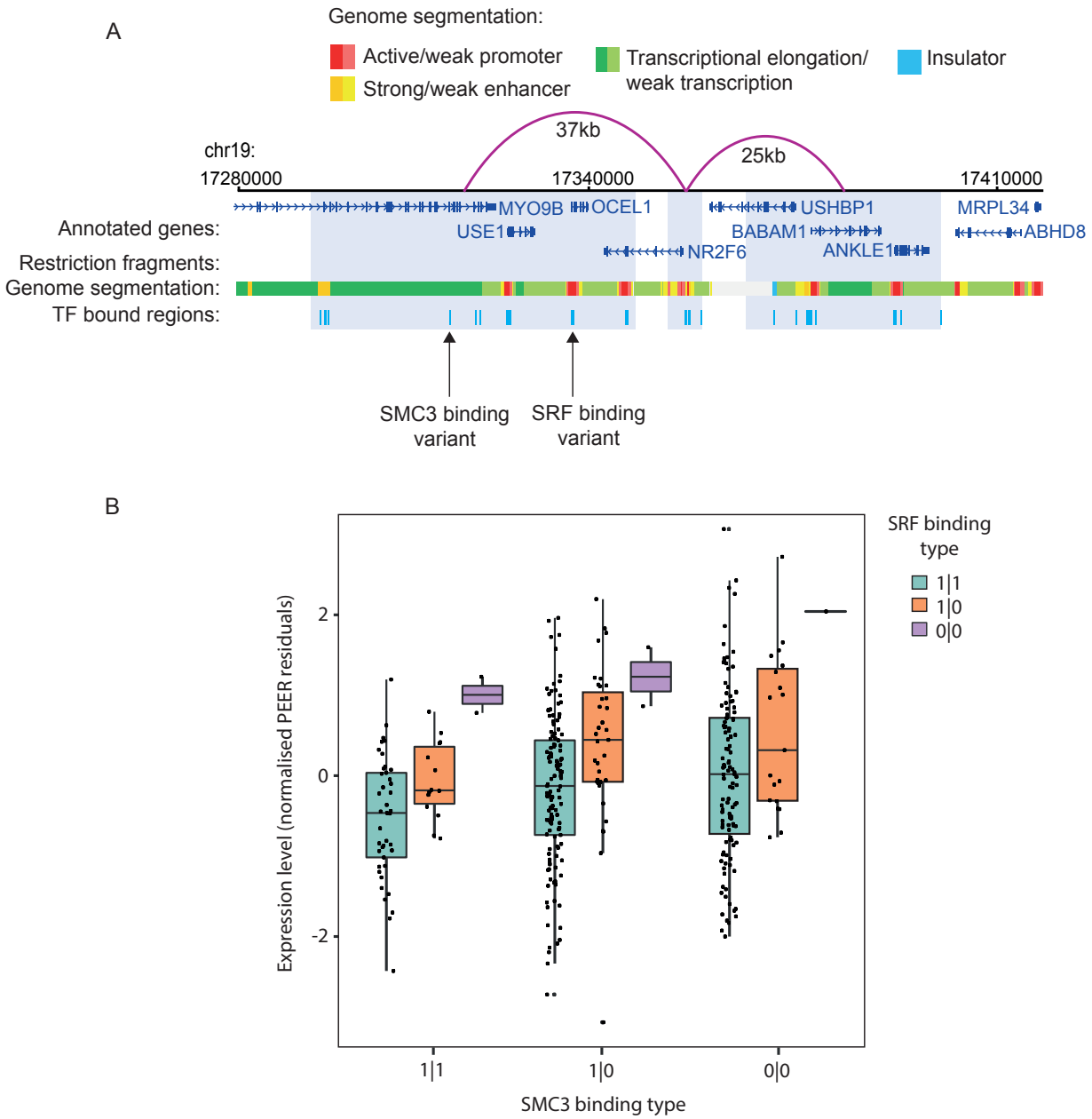

Figure 4. Example of a multi-variant association between TF binding affinity at CRMs and their target gene expression. (A) Genome browser representation of NR2F6 promoter distal interactions (represented by pink arches) as detected by promoter capture Hi-C (49) in LCL GM12878. The genome segmentation track for GM12878 based on chromHMM (108) is also shown. CRMs at the distally interacting fragments (pale blue) and NR2F6 TSS-proximal window are depicted in azure blue. The distal fragment downstream of NR2F6 contains two predicted affinity variant CRMs: one 44kb away from the NR2F6 promoter and the other 19kb away, predicted to impact SMC3 and SRF binding affinity respectively across the 359 LCLs. (B) Association between NR2F6 mRNA levels and predicted SMC3 and SRF affinity haplotypes.

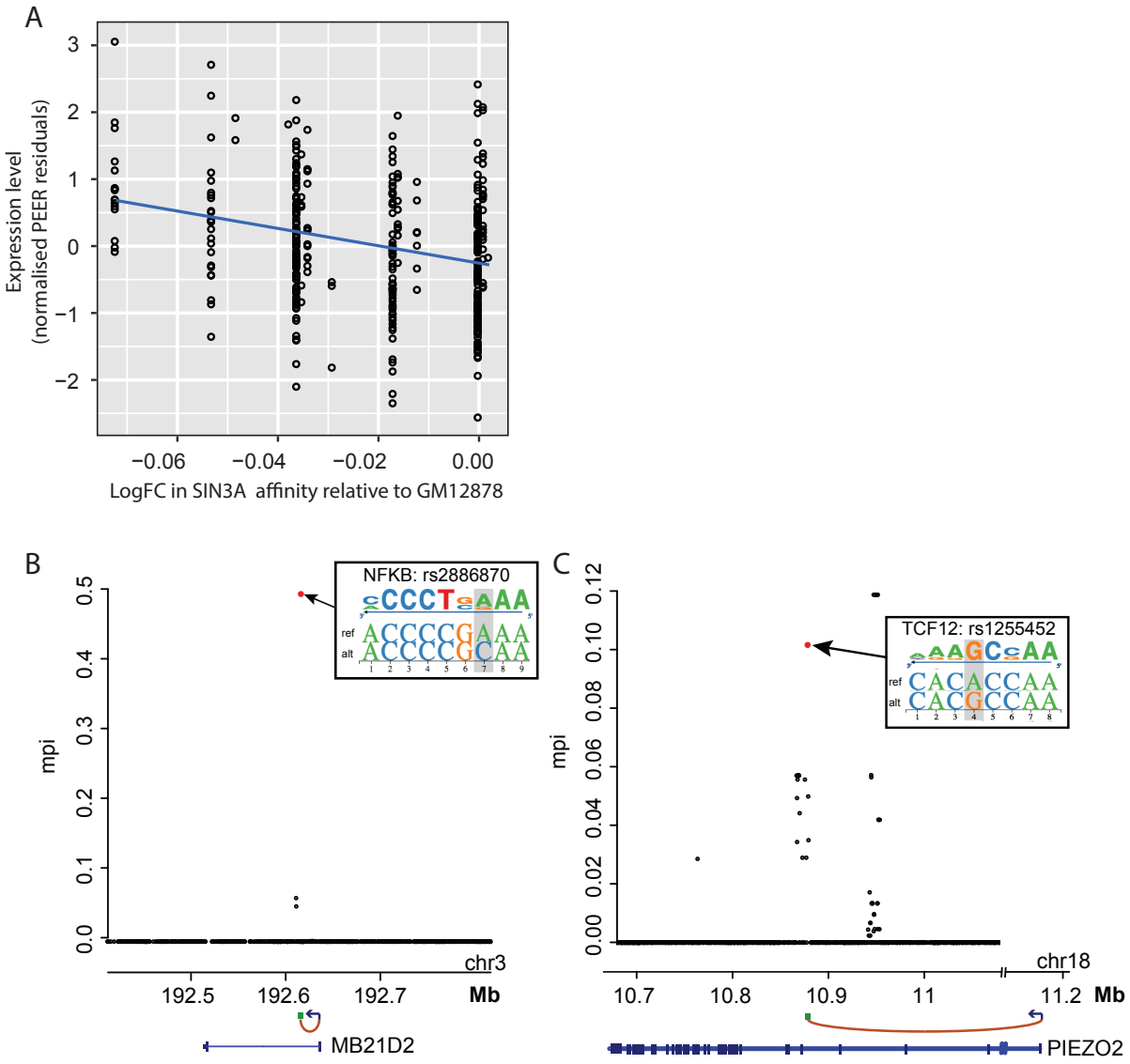

Figure 5. Threshold-free approach for detecting TF binding affinity variant associations with gene expression and validation using GUESSFM. (A) Association between log fold affinity change in CRM affinity for SIN3A relative to the highest affinity allele of GM12878 and mRNA level (normalised PEER residuals) of the connected gene, *SLC23A* (gene level FDR adjusted p-value=1.60x10^{-4}, beta=–0.14*).* (B) Examples of loci, whereby the SNP predicted to have the strongest impact on a CRM's binding affinity for a given TF has been fine-mapped as a potentially causal variant driving the locus's association with the expression of a physically connected target gene (GUESSFM marginal posterior probability of inclusion [mppi]>>0.001). Left panel: eGene: *MB21D2;* eQTL rs2886870, predicted to affect NFKB binding affinity. Right panel: eGene: *PIEZO2;* eQTL rs1255452, predicted to affect TCF12 binding affinity. See insets for the effects of the SNPs on the respective TF PWMs.

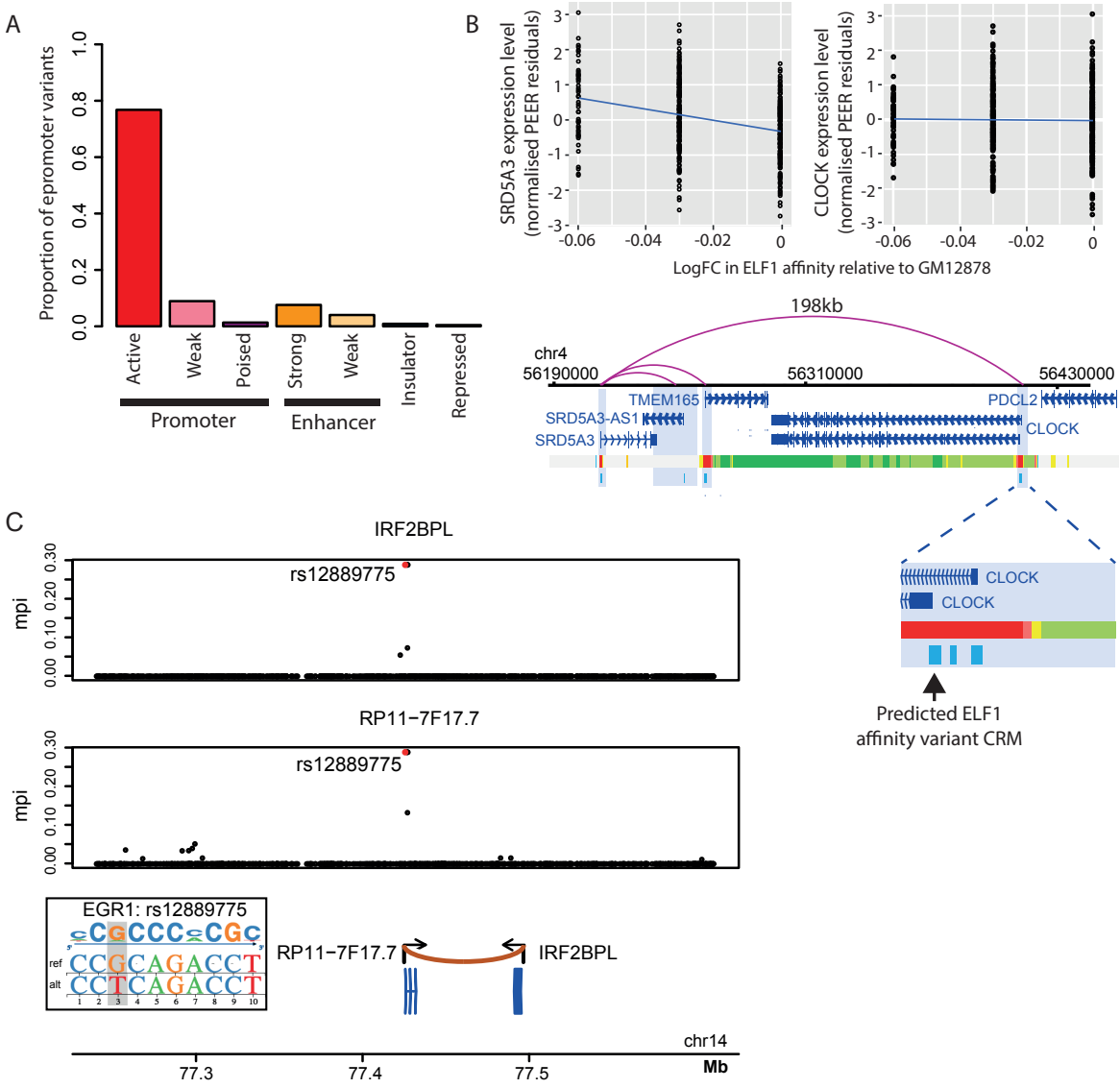

Figure 6. TF-binding affinity variants highlight transcriptional regulatory effects of epromoters. (A) Barplot showing the proportion of distal CRMs showing association between TF binding affinity and target gene expression that map in close proximity (within 200 bp) of another genes TSS overlapping each genome segmentation category (108) for GM12878. (B) Genome browser representation of the distal interactions detected by promoter capture Hi-C (49) for SRD5A3, with CRMs identified at each fragment as well as the proximal window depicted in light blue. The genome segmentation track for GM12878 based on chromHMM (108) is also shown. Enlarged view of an interacting fragment containing three CRMs, one of which harbours variants predicted to impact ELF1 binding affinity and overlaps with the promoter of CLOCK. (B) The association between logFC in CRM affinity for ELF1 relative to the highest affinity allele of GM12878 and mRNA level (normalised PEER residuals) of SRD5A3 and CLOCK. (C) Colocalisation analysis showing shared association between epromoter-located SNP rs12889775 and the expression of both its distal and proximal genes (*IRF2BPL,* top; and lncRNA *RP11−7F17.7,* bottom, respectively); Posterior probability of shared association estimated by the coloc software p_{H4}=0.997. This SNP is predicted to affect the epromoter's binding affinity for EGR1 (see inset).