High-resolution genetic mapping of putative causal interactions between regions of open chromatin

Natsuhiko Kumasaka, Andrew J. Knights and Daniel J. Gaffney*

Physical interaction of regulatory elements in three-dimensional space poses a challenge for studies of disease because non-coding risk variants may be great distances from the genes they regulate. Experimental methods to capture these interactions, such as chromosome conformation capture, usually cannot assign causal direction of effect between regulatory elements, an important component of fine-mapping studies. We developed a Bayesian hierarchical approach that uses two-stage least squares and applied it to an ATAC-seq (assay for transposase-accessible chromatin using sequencing) data set from 100 individuals, to identify over 15,000 high-confidence causal interactions. Most (60%) interactions occurred over <20 kb, where chromosome conformation capture-based methods perform poorly. For a fraction of loci, we identified a single variant that alters accessibility across multiple regions, and experimentally validated the BLK locus, which is associated with multiple autoimmune diseases, using CRISPR genome editing. Our study highlights how association genetics of chromatin state is a powerful approach for identifying interactions between regulatory elements.

Three-dimensional (3D) interactions between regulatory elements are a fundamental process in gene regulation. Understanding the guiding principles that control these interactions is a major research interest in genomics[1,2]. Long-range regulation poses a challenge for studies of human disease because risk variants may be located many kilobases from the genes they regulate, making causal variant identification difficult[3,4]. Chromosome conformation capture (3C)-based techniques have enabled the generation of genome-scale maps of 3D contacts in human cells[5]. These maps have provided valuable insights into large-scale structure and organization of chromosomes[6,7], and often also provide useful information linking distal disease risk alleles with putatively regulated genes[8,9]. However, it can be hard to distinguish functional interactions, such as enhancer–promoter looping, detected using 3C-based methods from a background of random collisions[10], which are particularly pronounced over distances of less than 20 kb[11].

A complementary approach to mapping genome-wide 3D interactions is to utilize germline genetic variation. Quantitative trait locus (QTL) mapping of chromatin traits can identify genetic variants that regulate chromatin both locally and distally, sometimes over distances of hundreds of kilobases[12-14]. These distal QTLs are known to be enriched in topologically associating domains[15,16], (TADs), suggesting that regulatory regions mapped by chromatin QTLs do indeed physically interact with one another. For fine-mapping of putative causal variants identified in human disease studies, this approach has some attractive features. First, unlike 3C-based techniques, our ability to detect interactions between regulatory elements is not correlated with the distance between them. Second, QTLs identified in these studies can be naturally aligned with those from disease studies using colocalization[17]. Third, causal interactions between different regulatory elements can be potentially deduced by Mendelian randomization[18-20], where germ line genetic variants are used as instrumental variables to resolve relationships between different active regions. Here we develop a pairwise hierarchical model (PMH) that incorporates a technique from Mendelian randomization in a Bayesian framework to map causal regulatory interactions using ATAC-seq data from 100 unrelated individuals of British ancestry.

Results

The model. Associations between genotype at the same genetic variant and chromatin accessibility often appear spread across multiple independent ATAC-seq ‘peaks’ of open chromatin[21] and can arise for multiple reasons. Two or more variants in linkage disequilibrium (LD) can drive independent associations at different peaks (hereafter, ‘linkage’). Alternatively, a single variant might independently drive association signals at multiple peaks (‘pleiotropy’). Finally, individual variants may alter accessibility at one regulatory element that in turn alters accessibility elsewhere in the genome, an indication that these elements functionally interact in 3D space (‘causality’). Our PHM classifies peak pairs within 500 kb of one another into hypotheses of linkage, pleiotropy, causality, a single QTL at either of the modeled peaks, or a null hypothesis of no QTLs in either peak (Fig. 1a). To compute the pairwise likelihood (Methods) for a given peak pair j and k, we calculate Bayes factors (BF or BMF) for the association between genotype at a putative causal genetic variant and chromatin accessibility at each member of the pair (Fig. 1b). For the hypothesis of causality we compute Mendelian randomization Bayes factors (MR BF or MRBMF) for the regression of chromatin accessibility in peak j on peak k (or vice versa) using two-stage least squares (2SLS), with genotype at the given genetic variant as the instrumental variable (Fig. 1b). We compute Bayes factors for all variants in a cis window extending 500 kb 5’ and 3’, marginalizing by the appropriate prior probabilities to derive a ‘regional’ Bayes factor (BRF) (Fig. 1c). We use a ‘variant-level’ prior probability of being a causal regulatory variant within the cis window (Fig. 1d) assuming a single causal variant[22]. We also model a ‘peak-level’ prior probability on the probability of observing a chromatin accessibility QTL (caQTL), which is a function

https://doi.org/10.1038/s41588-018-0278-6
of peak height (Fig. 1c), and a ‘peak-pair-level’ prior probability that adjusts the support for pleiotropy or causality between two peaks, as a function of the distance between them (Fig. 1f). Both the peak-level and peak-pair-level priors are conceptually similar to independent hypothesis weighting. The model outputs a posterior probability that a peak pair belongs to one of the interaction categories, including the posterior probability of a causal interaction (PPC). Hereafter, PPC\(j\) denotes the posterior probability that peak \(j\) regulates, or is ‘upstream’ of, peak \(k\), while PPC\(k\) denotes the converse (i.e. downstream of) \(k\). PPC without a subscript refers to the sum of PPC\(j\) and PPC\(k\).

Mapped causal interactions. Our consensus set contained 277,128 peaks of open chromatin, corresponding to over 17 million peak pairs. We found that 14% of peak pairs showed some evidence of genetic control (Fig. 1g). Summing over the posterior probabilities, we estimated that 23,036 peak pairs (0.13%) causally interact (for example, see Supplementary Fig. 1a); 15,487 we refer to as ‘high confidence’ (PPC > 0.5) with a Bayesian false discovery probability under one of the five hypotheses in Fig. 1a (Methods). The model found it more challenging to correctly assign the direction of causal effects (Fig. 2a; 18.9% incorrect directionality on average). Under simulated linkage, the FPR of causality increased with increasing LD between the two variants (Fig. 2b), but overall was low even for variants in high LD (0.0025 for variants in \(|r| > 0.99\)).

We extended our simulations to include two causal variants in the focal peak for each scenario (Methods). Multiple causal variants did not substantially increase the FPR for any scenario (Supplementary Fig. 2a). Finally, we simulated hybrid hypotheses of linkage, pleiotropy, and causality. Here, our power to detect causality reduced to 62.9% (hybrid pleiotropy, causality \((j \rightarrow k)\) or 37.5% (hybrid linkage, causality \((j \rightarrow k)\) and the FPR also became 5.3% on average across all hypotheses (Supplementary Fig. 2a).

Model performance assessment by simulation. To test model performance, we simulated data with one causal variant per focal peak, under one of the five hypotheses in Fig. 1a (Methods). The false positive rate (FPR) of causality when linkage or pleiotropy were simulated was 0.7% or 1.5%, respectively (Fig. 2c). The model found it more challenging to correctly assign the direction of causal effects (Fig. 2d; 18.9% incorrect directionality on average). Under simulated linkage, the FPR of causality increased with increasing LD between the two variants (Fig. 2e), but overall was low even for variants in high LD (0.0025 for variants in \(|r| > 0.99\)).

We extended our simulations to include two causal variants in the focal peak for each scenario (Methods). Multiple causal variants did not substantially increase the FPR for any scenario (Supplementary Fig. 2a). Finally, we simulated hybrid hypotheses of linkage, pleiotropy, and causality. Here, our power to detect causality reduced to 62.9% (hybrid pleiotropy, causality \((j \rightarrow k)\) or 37.5% (hybrid linkage, causality \((j \rightarrow k)\) and the FPR also became 5.3% on average across all hypotheses (Supplementary Fig. 2a).
We also compared our model’s performance on simulated data with MR Steiger, an alternative approach to identifying causal interactions\(^8\). We note that MR Steiger assumes that the causal variant is known, while our model attempts to infer the causal variant from the data. Despite this, the PHM produced a lower FPR for causality when data were simulated under the linkage or pleiotropy models (Fig. 2c), but MR Steiger was better at identifying the causal direction of effect. For example, at a PPC > 0.5, the PHM correctly called 24,332 causal (j → k) peak pairs and incorrectly called 216, 73, and 3,978 linkage, pleiotropy, or causality (k → j) under the true hypothesis of causality (j → k), respectively. For the same number of true positives (same power to detect the true causal (j → k) interactions), the equivalent numbers were 2,998, 1,014, and 9 for MR Steiger. We observed that the misclassification rate of causal direction for the PHM decreased substantially when the causal variant was more strongly associated with the focal peak (less than 1% for \(BF_j > 100\)) (Supplementary Fig. 2b).

**Model performance assessment using real data.** Next, we investigated model performance on real data. Effect directions for inferred causal peak pairs were substantially more likely to be in the same direction than peaks in linkage (Fig. 2d–e), with 98.2% of peak pairs concordant in the confident causal set compared with 57.5% in the confident linkage hypothesis (posterior probability of linkage > 0.5). Using RoadMap Epigenomics data on chromatin state from 53 different cell types (Methods), we also observed that the activity of causally interacting or pleiotropic peak pairs was highly correlated across tissues than distance-matched controls, but significantly lower than matched controls for linkage peak pairs (Fig. 2f). An example of cross-tissue correlation is shown in Supplementary Fig. 2c.

Because allele-specific signals were not used in the original model, they can provide independent confirmation of a genetic effect. We used RASQUAL\(^8\) to map caQTLs using allele-specific counts only at feature SNPs in the downstream peak region. A QQ-plot of \(P\) values for the allele-specific signals was significantly skewed toward 0 compared with the distance-matched controls of null peak pairs or linkage peak pairs (posterior probability of null or linkage hypothesis > 0.5) (Fig. 2g and Supplementary Fig. 2d–e).

Finally, we examined the overlap between transcription factor binding site footprints and the lead variants detected by our PHM, and compared this with results from a hierarchical model that did not consider interactions between peaks (as used in the first stage of optimization). Compared with the hierarchical model, lead variants detected by the PHM were highly enriched in transcription factor footprints (Fig. 2h), particularly B-cell-specific transcription factor motifs (Fig. 2i), such as IRF1 and PU.1. The ratio of putative binding affinity between reference and alternative alleles at the PHM...
lead variant was also more highly correlated with allelic imbalance of ATAC-seq reads at PHM lead SNPs (Fig. 2j) than hierarchical model lead SNPs (Supplementary Fig. 2f).

**Comparison with 3C-based assays.** We compared causal interactions inferred from our chromatin loops inferred from Hi-C, promoter capture Hi-C (Chi-C), and H3K27ac HiChIP (Hi-C chromatin immunoprecipitation) applied to GM12878 (Fig. 3a). Seventy-four percent of causal interactions were between peaks located within the same TAD called from Hi-C, a fivefold enrichment over genomic background (Fig. 3b,c). The remaining causal interactions (26%) were primarily in non-TAD regions (Supplementary Fig. 3a–b), with peaks spanning TAD boundaries being significantly (15-fold) depleted (Supplementary Fig. 3c). Although rare, the TAD boundary-spanning interactions we did detect were as strongly supported by allele-specific accessibility analysis as those found within TADs (see the ‘Across TADs’ panel, Supplementary Fig. 3d). Effect directions of lead variants were less concordant when peak pairs spanned one or more insulators or TAD boundaries (Supplementary Fig. 3e), with average concordance of 89.0% and 86.3%, respectively \((P = 4.5 	imes 10^{-7} \text{ and } 2.9 	imes 10^{-14})\), compared with background sets \((91.8\% \text{ and } 91.6\%)\). An example of a causal interaction spanning a TAD boundary is shown in Supplementary Fig. 2d.

Causal interactions were also enriched for loops inferred from H3K27ac HiChIP and Chi-C data \((7.7\text{-}1.4\text{-fold, respectively})\), although the absolute numbers of overlaps were small \((152 \text{ and } 324; \text{respectively})\). Our model also highlighted interactions that could be missed by promoter capture-based techniques. Of the 49,579 peak regions linked to baited promoters, we estimated that there were 2,208 causal interactions between the non-promoter elements and a further 561 between two peaks located within the same Chi-C bait region (Fig. 3b–c).

Most causal interactions occur over sub-20-kb distances. We found that causal interactions inferred from the PHM occurred over much shorter distances than those captured by 3C-based techniques (Fig. 3d): 63% were less than 20 kb distant from one another, compared with 7% of Chi-C interactions (Fig. 3d). We confirmed that the distance distribution did not reflect interactions between pseudo ‘subpeaks’ that were part of the same broad peak (Supplementary Fig. 3f,g). Our results suggest that many functional 3D interactions may be below the resolution of conventional 3C-based techniques. One example is shown at the promoter region of the MAP1B gene (Fig. 3c). Here, a high-confidence \((\text{PPC}_c > 0.99)\) causal interaction occurs between a promoter and an enhancer that is less than 13 kb distal, but the contact domain inferred from Chi-C data processed by CHICAGO (Capture Hi-C Analysis of Genomic Organisation) software has weak statistical support (CHICAGO score 1.87).

**Enhancer–enhancer and promoter–enhancer interactions are common.** We next examined the functional classes to which the members of causally interacting regulatory elements belonged, using...
Fig. 4 | Comparison with genome segmentation. a, The numbers of causal peak pairs overlapping ENCODE genome segmentation. Numbers of interactions were computed as a weighted sum of PPC_{jk}. The ATAC-seq peaks are classified by seven different regulatory categories: active promoter (AP), poised promoter (PP), strong enhancer (SE), weak enhancer (WE), repressor (R), transcribed region (T), and insulator (I). Each bar indicates upstream peak category and the color code indicates downstream peak category. b, Enrichment of ENCODE segmentation category pairs for our causal interaction. Heatmap shows the odds ratios (see Methods for computation of enrichment using PPC_{jk}) for all combinations of segmentation categories at upstream and downstream peaks (among n = 17,349,412 total peak pairs). The segmentation category pairs that were above FDR 10% or supported by less than 10 causal peak pairs are masked by gray. c, The numbers of causal peak pairs that are jointly colocalized with one or more eQTLs overlapped with the ENCODE segmentation. d, Enrichment of ENCODE segmentation category pairs for our causal interactions that are jointly colocalized with one or more eQTLs (among n = 23,068 causal peak pairs) (see Methods for computation of enrichment using PPC_{jk}). e, The number of peak pairs whose upstream peak overlaps with one of the seven segmentation categories, stratified by the genotypes of GM12878 at lead QTL variant (Methods). Each genotype is labeled as a combination of decreasing ‘D’ and increasing ‘I’ alleles according to the sign of QTL signal at the lead variant. Color code is same as in a. f, An example of causal interaction from a repressed region to a weak enhancer. The normalized ATAC-seq coverage is stratified by three genotype groups at rs2046338:C >A. The yellow line shows ATAC-seq coverage of GM12878, whose genotype is AA (decreasing homozygote) at rs2046338.
the ENCODE (Encyclopedia of DNA Elements) genome segmentation annotations for lymphoblastoid cell lines (LCLs)27,28 (Methods). The most frequent class of interactions (5,061 peak pairs, 22% of all interactions) were strong enhancers that appeared to regulate other element types, including other strong enhancers (1,531 peak pairs, 6.6%), a 2.5-fold enrichment (Fig. 4a,b). The effect directions of

![Figure 5](image_url)

**Fig. 5 | Fine-mapping eQTLs using mapped causal interactions as an annotation.** a, Distribution of the number of variants in the 90% credible set, across all protein-coding genes with more than 1 colocalized ATAC-seq peaks (N genes = 1,207) over 9 different annotation combinations. FLAT, non-informative prior; ATAC, inside/outside an ATAC-seq peak; HiChiP, HiChiP anchor regions; Chi-C, Chi-C contact domains; VL, variant location. In the boxplots, the box represents the interquartile range (IQR), the black line is the median, and the whiskers are 1.5 times the IQR above or below the first and third quartiles, with data points outside the whiskers shown by open circles. b, The number of emVars overlapping lead eQTL variants detected by the eQTL hierarchical model with various annotations. c, An example of a fine-mapped region with more than 100 significant variants in almost perfect LD. The top panel shows negative log₁₀ Bayes factors of eQTL for GPATCH2L gene using GEUVADIS RNA-seq data. Each point is colored by the degree of LD index (r² value) with the index variant (rs147768071:AGTTTT > A). The SNP (rs74067641:T > C) in the master regulatory peak shows the highest PPA with ATAC + PMR + VL annotation.
the lead variant between strong enhancer–strong enhancer interac-
tions were significantly more concordant compared with the back-
ground (Supplementary Fig. 4a), a 95.0% concordance (P = 0.0043) compared with the complement set (82.2%), suggesting that those regions may work in a coordinated manner.

When we focused only on variants that also altered gene expres-
sion, using 4,670 interacting peak pairs that jointly colocalized with an expression QTL (eQTL) from the GEUVADIS (Genetic European Variation in DiSetase) data set (Methods), we found that these were enriched (2.4-fold, P = 6.4 × 10−19) for strong enhancer to active promoter interactions (Fig. 4c,d). However, expression-associated variants were also enriched for interactions from active promoters to strong enhancers (2.2-fold) or between pairs of strong enhancers (2.2-fold enrichment) (Fig. 4d). One hypothesis is that many of these interactions are mediated by transcriptionally induced changes in chromatin accessibility over the gene body, cre-
ating apparent interactions between a single upstream functional element and chromatin peaks throughout the transcribed region. Consistent with this idea, peaks downstream of an active promoter were significantly enriched in the gene body (2.3-fold enrichment, P = 8.1 × 10−34; Supplementary Fig. 4b) compared with peaks to the 5′ of the promoter. This hypothesis is also consistent with the obser-
vation that chromatin accessibility over the gene body is highly correlated with gene expression level (Supplementary Fig. 4c). A striking example of this potential phenomenon is found at the MB21D2 locus (Supplementary Fig. 1c).

Genetically driven changes in the reference epigenome. We found a surprisingly large number of interactions (4,134 peak pairs) origin-
ating from within repressed regions (Fig. 4a). Preliminary analy-
sis suggested that these might arise due to genotype effects on the reference epigenome annotation derived from a single individual (GM12878). To test this, we stratified all upstream peaks in causally interacting pairs based on whether their lead caQTL genotype in GM12878 was an increasing homozygote, decreasing homozygote, or heterozygote (Methods). Upstream repressed regions were highly enriched (3.1-fold) for decreasing homozygotes compared with increasing homozygotes (Fig. 4c), suggesting that in these cases a strong caQTL almost completely removes a region of open chroma-
tin in GM12878, an example of which is shown in Fig. 4f. We found that 1.4% of repressed regions overlapped a caQTL where GM12878 was a decreasing homozygote. This estimate is also likely to be a lower bound due to incomplete power to detect caQTLs.

Causal interactions improve fine-mapping. Next, we examined whether the information on causal direction of variant effects could be used to improve fine-mapping accuracy, using gene expression as a model quantitative trait. For each peak within a 1-Mb cis window around a gene transcription start site (TSS), we first computed the probability of master regulator (PMR) for each peak (Methods). We then used a hierarchical model24 to compute the posterior probabilities of association (PPA) for eQTL variants with PMR and the following four other annotations: (1) inside or outside an ATAC-
seq peak; (2) eQTL variant location, relative to an ATAC-seq peak coverage; (3) promoter CHI-C contacts; and (4) HiChIP loops from promoter regions (see Methods for details). Genome-wide, the best performing annotation was the combination of PMR with ATAC-
seq peak status and variant location, which reduced the 90% credible set of eQTL variants by 65%, from 17 to 6 variants on average, compared with 11 variants for CHI-C, 10 for ATAC-seq peaks, and 8 for CHI-C combined with ATAC-seq peaks (Fig. 5a). The effect of adding information on the causal direction, by prioritizing the most upstream variant via the PMR, significantly reduced the cred-
ible set size compared with the ATAC-seq peak annotation alone (P < 10−49, paired t-test). We then compared our results with data from the massively parallel reporter assay (MPRA) performed in LCLs35 (Methods). We found the highest overlap (21.6% or 182 expression-modulating variants (emVars)) for the combined PMR, ATAC-seq peak, and variant location annotations (Fig. 5b). We applied this approach to a challenging locus, where a strong eQTL for the GPATCH2L gene is associated with more than 100 candidate regulatory variants in almost perfect LD (Fig. 5c). With no annota-
tion information, the 90% credible set size at this locus is large, at 65 variants. Although different annotations produce varying effects, our model proposes an SNP (rs74067641:T > C) as the likely causal variant with the highest PPA = 0.42. This variant is located within a predicted master regulatory peak located furthest upstream in the regulatory cascade (Supplementary Fig. 5a–c). We note that reduc-
tion in credible set size is an imperfect measure of fine-mapping accuracy in cases where multiple causal variants are segregating.

Causally interacting caQTLs are enriched in autoimmune genome-wide association study (GWAS) hits and eQTLs. We per-
formed an enrichment analysis of causally interacting caQTL peaks for disease GWAS hits. We colocalized our caQTLs with 10 GWASs whose genome-wide summary statistics were available26–29. caQTLs detected in LCLs strongly colocalized with autoimmune diseases, including rheumatoid arthritis (140 colocalized caQTL GWAS loci) and systemic lupus erythematosus (SLE; 96 loci) (Fig. 6a). Using rheumatoid arthritis as an example trait, we found that causally interacting loci were significantly more likely to colocalize (1.8-fold, P = 1.4 × 10−7) with risk loci than non-interacting caQTLs (Fig. 6b). Interacting peaks that also colocalized with an eQTL were further enriched (2.9-fold, P = 1.7 × 10−7). This suggests that causal interactions were more often involved in a gene regulatory cascade leading to downstream consequences.
CRISPR validation of a putative causal variant at the BLK locus.
Finally, we applied our method in an attempt to fine-map a challenging GWAS locus with contradictory evidence for multiple causal variants in previous studies. The BLK/FAM167A locus on 8p21 has a strong eQTL (GEUVADIS $P < 10^{-26}$ and $10^{-46}$ for BLK and FAM167A genes, respectively) in LCLs (Fig. 6c) that colocalizes
well with genome-wide significant associations for SLE and rheumatoid arthritis (Supplementary Fig. 6a,b). Previous attempts to fine-map this locus have been hampered by multiple genetic variants in tight LD (Supplementary Fig. 6c,d). Two SNP variants, rs1382568:A > C,G and rs922483:C > T, located near the promoter of the BLK gene, have previously been reported as putative causal variants of SLE that alter BLK expression in various B- and T-cell lines\(^a\). However, MPRA studies have pinpointed a different deletion variant (rs5889371:AG > A) that might also potentially alter BLK expression in LCLs\(^b\). Two of the previously reported variants (rs5889371 and rs1382568) are located in regions of low chromatin accessibility (Supplementary Fig. 6e–g) and are less likely to causally influence BLK expression in LCLs.

We detected a single base pair insertion variant (rs558245864:C > CG) located in a strong cQTL peak 14 kb upstream of the BLK promoter that interacted with 15 flanking peaks including several promoter peaks (Fig. 6c). The insertion variant showed the highest posterior probability (PPA = 0.59) of any putative causal eQTL variant for the BLK gene (Fig. 6c). This variant is located at the middle of a canonical CTCF binding motif, with an extra ‘G’ nucleotide decreasing the predicted CTCF binding affinity almost to background (Fig. 6d). The direction of binding affinity change was consistent with the cQTL signal. This variant was also a CTCF ChIP-seq (chromatin immunoprecipitation) sequencing (CTCF) (Fig. 6d), with 99.7% probability of colocalization between the CTCF and BLK for this peak (Methods). We used CRISPR-Cas9 genome engineering to generate two different heterozygous deletion lines from a parental line that was homozygous for the high CTCF binding allele (Methods). These deletions overlapped the CTCF binding site: the 6-bp deletion disrupts the right-hand side of the binding site and the 18-bp deletion removes almost the entire motif (Fig. 6d). ATAC-seq and RNA sequencing (RNA-seq) in the deletion lines showed a significant downregulation of chromatin accessibility at the focal peak compared with the parental line (P = 0.0005) (Fig. 6c), and a concomitant downregulation of BLK expression (P = 0.0095) (Fig. 6f). We observed decreases in accessibility at some neighboring peaks around the BLK promoter region (Supplementary Fig. 6h,i). We also observed an increase in accessibility around the FAM167A promoter region (Supplementary Fig. 6h,i) and in FAM167A expression (Supplementary Fig. 6i), although this was not significant (P = 0.18).

**Discussion**

We have presented a novel approach to detect interactions between regulatory elements, which uses principles of Mendelian randomization embedded within a Bayesian hierarchical model. We show that the majority of causal interactions within 500 kb occur over short distances (< 20 kb), typically a region of low sensitivity for 3C-based techniques. Many of the interactions we detect are between enhancers, which we assemble into hierarchies of interacting regulatory elements. We demonstrate that our model can be used to identify hierarchies of regulatory elements within a region and prioritize putative causal variants, validating a single locus using CRISPR-Cas9 editing.

The low frequency of long-range interactions we observed agrees with previous estimates from eQTL studies\(^3,4,6\). One question is, given that most regulatory interactions detected using 3C-based methods occur over distances of 100 kb and above (Fig. 3d), why have large numbers of genetic variants operating at these distances not also been detected? Although QTL studies typically test variants in a restricted cis window of 1 Mb\(^5,6\), this does not completely explain the lack of signal: the number of eQTL associations detected decreases dramatically by approximately 20 kb distant from the gene TSS\(^5,6\). A possible explanation is that there may be an underlying relationship between interaction distance and cellular frequency, such that long-range interactions occur in a relatively small number of cells in the population\(^9\). This is consistent with the negative correlation between read coverage and distance in CHI-C data (Supplementary Fig. 7a). It seems plausible that 3C-based methods could be more sensitive to rare, long-range regulatory interactions while variants residing in these elements have relatively weak effects\(^9\), requiring large sample sizes to detect when averaged across the entire cell population. An alternative hypothesis is that short-range interactions may not be driven by chromatin looping, but instead reflect transcriptional activity and the movement of polymerase across the sequence (Supplementary Fig. 4b,c).

Our study also identified the genomic architecture of causal interactions between regulatory elements. In particular, we detected frequent interactions between annotated enhancer elements, many of which we hypothesize are mediated by an intermediate eQTL that alters chromatin accessibility globally across the gene body. Nonetheless, the enrichment of these interactions in gene bodies was modest, and we also found many examples of interactions that were not colocalized with eQTLs, and were located far from annotated genes (an example is shown in Supplementary Fig. 7b). In a small number of cases (18 DAGs) we also found strong evidence (PPC > 0.5 for each enhancer pair) that these occurred between multiple enhancers upstream of a promoter (that is, strong enhancer → strong enhancer → active promoter). It is possible that some of these represent enhancer ‘seeding’ events, where individual enhancers drive progressive activation of additional nearby elements\(^6\).

One of the limitations of our method is that regulatory elements lacking a common genetic variant that perturbs their function will be missed. Additionally, interactions between genotype and regulatory elements further downstream appear to become harder to detect, perhaps due to additional biological noise. One example of this is the systematically lower genetic effect sizes (14% decreasing) we found at downstream promoters (Supplementary Fig. 7c,d).

Our approach allows for a natural prioritization of variants in disease-associated loci. Although overlapping of those variants with open chromatin can reduce credible sets, this frequently leaves many loci with tens of variants to characterize by direct experimental follow up. Assignment of the direction of effect between different peaks allowed us to identify smaller sets of plausible candidate variants by identifying ‘master regulatory’ regions. Although we have focused on ATAC-seq data, we believe our model can be readily extended to other types of chromatin-based assay, in particular ChIP-seq for histone modifications\(^13,15,17\). Some limitations of this approach might include a greater difficulty in assigning causal variants based on their location within a ChIP-seq peak, which will typically be in a nucleosome-depleted region and therefore low-read coverage\(^1\) (for an example, see Supplementary Fig. 7c). However, we anticipate that, applied to existing data sets from primary cells, such as that generated by the BLUEPRINT initiative\(^4\), our approach will be a valuable tool in dissecting the molecular architecture of specific GWAS loci.

**URLs**

PHM, https://github.com/natsuhiko/PHM; RASQUAL, https://github.com/natsuhiko/rasqual; 1000 Genomes Phase III integrated variant set, http://www.internationalgenome.org/data; Combined Segway and ChromHMM results from ENCODE Project website, http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgSegmentation/; Roadmap Epigenomics Project, http://www.roaddmapepigenomics.org/; CHICAGO 1.1.8, http://regulatorygenomicsgroup.org/chicago; Juicer Tools 0.7.5, https://github.com/theidenlab/juicer/wiki/Juicer-Tools-QuickStart; Beagles 4.0, https://faculty.washington.edu/browning/beagles/b4_0.html; DESeq, https://bioconductor.org/packages/release/bioc/html/DESeq.html; bwa 0.7.4, https://sourceforge.net/projects/bio-bwa/files/; Bowtie2 2.2.4, http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml; TopHat2 2.0.13, http://ccb.jhu.edu/software/tophat/index.shtml; skewer 0.1.127, https://github.com/reliproc/skewer.
**Methods**

**ATAC-seq in LCLs.** We collected 100 LCL samples of British ancestry (1000 Genomes Project, GBR cohort) from Cordell. ATAC-seq library preparation was performed for each line (except for the 24 lines we previously performed \(^2\)) as previously described \(^2\). We performed 75-bp paired-end sequencing in 4.4 billion sequence fragments on a HiSeq 2500 (Illumina). Although data from the 24 lines have been previously sequenced \(^3\), we performed additional sequencing to increase the coverage. We called 277,128 chromatin accessibility peaks on autosomes from the aggregated data (Supplementary Note, Section 1), from which we mapped caQTLs. We also performed an additional ATAC-seq experiment in GM12878 that was not used for QTL mapping, but was used to assess genotypic effects on the reference epigenome.

**Sequencing data preprocessing.** All sequence data sets were aligned to human genome assembly GRCh37. We performed adapter trimming for our ATAC-seq data using skewer \(^4\) (version 0.1.127; see URLs) before alignment. FASTQ files of GEUVADIS RNA-seq data \(^1\) (N=372) were downloaded from ArrayExpress (Accession E-GEUV-3) and ChIP-seq data for CTCF binding \(^2\) (N=50) were downloaded from the European Nucleotide Archive (accession ERP000168). Our ATAC-seq data and the CTCF ChiP-seq data were aligned using bwa 0.7.4 \(^7\). RNA-seq data were aligned using Bowtie2 \(^2\) (version 2.2.4; see URLs) and reads mapped to splice junctions using TopHat \(^8\) (version 2.0.13; see URLs), using ENSEMBL human gene assembly 69 as the reference transcriptome. Following alignment, we performed peak calling in the CTCF ChiP-seq and ATAC-seq data by pooling all samples. Fragment counts of ATAC-seq, CTCF ChiP-seq, and RNA-seq for each feature (a called peak or a union of exons for each gene) were normalized to uniquely mapped FPKMs (per kilobase of feature per million reads mapped). Batch effects were adjusted by GC contents and principal components. See Sections 2.1–2.5 of the Supplementary Note for more detail.

**SNP genotype data.** We downloaded VCF files from the 1000 Genomes Phase III integrated variant set from the project website. For the ATAC-seq, RNA-seq, and CTCF ChiP-seq samples that did not overlap with the 1000 Genomes Phase III samples, we extracted genotype data from the 1000 Genomes Phase I data or 1000 Genomes high-density SNP chip data (performed on the Illumina Omni platform). We then performed whole-genome imputation for the extracted genotype data using the Beagle software \(^9\) (version 4.0; see URLs). See Section 2.6 in the Supplementary Note for details.

**Genomic annotations.** To compute ATAC-seq peak height, we pooled ATAC-seq data for the 100 samples. The peak height was defined as the highest value of the coverage depth within each peak region. Peak height was quantile normalized across all peaks. The relative coverage at each variant location was calculated by the absolute coverage depth divided by the peak height inside the peak. This value was used as the variant location prior probability for both caQTL mapping and eQTL mapping. Peak distance was calculated based on the mid-point of a peak region. We also used various external genomic annotations for comparison. The Hi-C contact map and Hi-C loops for GM12878 were obtained from Rao et al. (2004) \(^10\). The JuiceBox output was processed by HiCUPS \(^2\) implemented in the Juice Tools (version 0.7.5; see URLs) with default parameter settings to obtain the HiChIP loops. The integrated genomic segmentation annotation \(^11\) combining Segway \(^2\) and ChromHMM \(^2\) results was downloaded from the ENCODE Project website (see URLs). Each ATAC-seq peak was labeled by one of the seven different segmentation categories at the peak mid-point. See Section 2.7 in the Supplementary Note for details.

**Roadmap Epigenomics Project data analysis.** We downloaded all DNase-seq (DNase I hypersensitive sites sequencing) data from 53 cell types from the project website (see URLs). We counted the number of reads that mapped to the 277,128 annotated peaks from our ATAC-seq data. This count matrix was normalized in the same way as our ATAC-seq data (see Sections 2.3–2.4 of the Supplementary Note). We computed Spearman’s rank correlation between all peak pairs within 500–kb distance of one another.

**PHM.** There are three key features of the model. First, support for the hypothesis of a causal relationship between two peaks is computed using 2LSL \(^2\). Second, we use a hierarchical model \(^2\) in which prior probabilities depend on genomic annotations at multiple model levels. Third, the model is empirical, such that the prior probabilities are learned as the penalized likelihood is maximized across all peak pairs simultaneously.

The PHM is a product of finite mixture probabilities over all j–k peak pairs in 500 kb if i < j < k ≤ j; \(j = 277,126\). The finite mixture model comprises the regional Bayes factor (RBF) \(^{4,11}\) to observe chromatin accessibility \(y_j\) and \(y_k\) at peak \(j\) and \(k\) across 100 samples under the different interaction hypotheses \(h\) (Fig. 1a). The pairwise likelihood is given by

\[
L_j(\Phi) \propto \prod_{1 \leq i < j < k \leq 277,126} \sum_{h=1}^{4} \Phi^{(h)} \text{RBF}^{(h)}_{jk}
\]

where \(\Phi^{(h)}\) denotes the mixture probability that j–k peak pair is in no caQTLs, \(\Phi^{(h)}\) denotes the mixture probability for the alternative hypothesis \(h\), and \(H_i\) is the set of alternative hypotheses, so that \(\Phi^{(h)} + \sum_{h \neq h} \Phi^{(h)} = 1\). RBF is obtained from the joint segmentation model \(P(y_j, y_k|h)\), which comprises two independent segmentation models that also depend on the hypothesis \(h\). For the causality hypotheses, \(H_i\) for the causal interaction from peak \(j\) to \(k\) and \(H_{j-k}\) for peak \(k\) to \(j\), we used 2SLS to estimate the causal effect between peaks with each genetic variant in the cis window as the instrumental variable (Fig. 1c).

We note that our model is not strictly Bayesian because we do not perform any Bayesian inference on the model parameters. Instead, to reduce the computational complexity, we employed a two-stage optimization of the likelihood using the EM (expectation-maximization) algorithm. In the first stage we estimated hyperparameters for the variant-level and peak-level prior probabilities. We used the standard hierarchical model \(^2\) to learn these prior probabilities by temporally assuming that peaks were independent. In the second stage we estimated hyperparameters in the peak-pair level prior regarding \(\Phi^{(h)}\). We used the expectation-maximization algorithm to iteratively estimate hyperparameters while updating the following posterior probabilities

\[
Z_{jk}^{(h)} = \Phi^{(h)} \text{RBF}^{(h)}_{jk} + \sum_{h \neq h} \Phi^{(h)} \text{RBF}^{(h)}_{jk}
\]

in the E-step. Because all model distributions belong to exponential family, we can utilize the penalized iteratively reweighted least square method \(^2\) in the M-step, which does not require calculation of the gradient and Hessian of the log likelihood. All subsequent analyses were performed based on the posterior probabilities \(Z_{jk}^{(h)}\) without any threshold. Note that the PPC is denoted by PPC\(_j\) and PPC\(_k\) (corresponding to \(Z_{jk}^{(0)}\) and \(Z_{kj}^{(0)}\), respectively) in the main text. Mathematical rationale and implementation of the PHM are fully described in Sections 3.1–3.5 of the Supplementary Note. A software package (PHM) that computes the Bayes factors and RBFs and maximizes the pairwise likelihood is available from GitHub (see URLs).

**Mapping multi-way interactions.** Multi-way interactions were also constructed from PPC\(_j\) and PPC\(_k\) by finding a DAG among more than two peaks. We first used only confident causal interactions with PPC\(_j\) > 0.5, then found the most likely parent for each peak, and finally solved the cyclic graphs by discarding an interaction with the lowest PPC\(_j\). See Section 3.6 of the Supplementary Note for details.

**Detection of lead caQTL variant.** Within each cis-regulatory window (500 kb on either side of a peak), we calculated a posterior probability of each variant being the causal caQTL and obtained the maximum a posteriori variant as the lead variant. We used the pairwise likelihood to solve the problem that multiple caQTL variants are associated with chromatin accessibility due to strong LD. The central assumption here is that variants predicted by our model to be upstream in the regulatory cascade are more likely to be causal. See Section 3.7 in the Supplementary Note for details.

**Effect size calculation.** For j–k peak pairs, we identified single lead variants under the hypothesis of causality or pleiotropy and two causal variants for the linkage hypothesis based on the variant-level posterior probability (see Section 3.8 of the Supplementary Note for more detail). We computed effect sizes of the lead variant(s) against the two peaks (j and k) using simple linear regression. Under the linkage hypothesis, if the genotypes of the two causal variants were negatively correlated (LD index \(r < 0\), the effect size of peak k was multiplied by \(-1\) to align the effect direction.

**Probability of master regulator.** We defined the master regulatory peak as a peak with more than one interacting downstream peak and no interacting upstream peaks. We computed the product of the following two posterior probabilities: the probability that the peak regulates at least one other peak in the cis window, and the probability the peak is not regulated by any other peak within the cis window, which we referred to as the PMR. See Section 3.9 of the Supplementary Note for details.

**Hierarchical model for eQTL fine-mapping.** The standard hierarchical model \(^2\) was applied to the GEUVADIS RNA-seq data (372 European samples) with various combinations of the following 5 annotations: (1) inside or outside an ATAC-seq peak (referred to as ATAC); (2) eQTL variant location, relative to an ATAC-seq peak (referred to as variant location); (3) promoter CH-C contacts; (4) HiChIP loops from baited promoter regions (HiChIP); and (5) PMR value at each ATAC-seq peak. The variant-level prior was learned and the PPA was calculated for each
variant in a 1-Mb cis window centered at the TSS. For the eQTL fine-mapping of BLK/FAM167A locus, we incorporated all of the genomic annotations used in the caQTL mapping in conjunction with the colocalization probability of caQTL and eQTL as the weight of the prior probability. See Section 3.10 of the Supplementary Note for details.

Colocalization with eQTLs. The PHM can be utilized to colocalize caQTLs with other cellular QTLs, such as eQTLs. The reduced model without causality hypothesis \( H_{xj} \) and \( H_{jk} \) was applied to colocalize caQTL-eQTL pairs as well as CTCF binding QTL-caQTL pairs. We assumed a non-informative prior probability for the three different levels of hierarchy and estimated the posterior probability of pleiotropy between caQTL and eQTL/CTCF binding QTLs as the colocalization signal. Joint colocalization probability between eQTL and a peak pair is also calculated from the result. See Section 3.11 of the Supplementary Note for details.

Colocalization with GWAS summary data. We downloaded the following ten GWAS summary statistics (see Section 2.9 of the Supplementary Note for details): rheumatoid arthritis, schizophrenia, SLE, Crohn’s disease, ulcerative colitis, inflammatory bowel diseases, type 2 diabetes, Alzheimer’s disease, atopic dermatitis, and coronary artery disease. The asymptotic Bayes factors were calculated and colocalized with caQTLs using the same model as was used in colocalization with eQTLs. Posterior probability of each caQTL peak colocalized with a GWAS trait was calculated and used for the subsequent enrichment analysis. See Section 3.12 in the Supplementary Note for details.

Allele-specific accessible chromatin. We used the lead caQTL variant for each peak identified by PHM as the putative causal variant. We confirmed allelic imbalance at feature SNPs inside the downstream peak in the confident set of 15,487 causal interactions. If there was a true causal interaction, allelic imbalance was observed for individuals who were heterozygous for the lead variant. To assess statistical significance of allelic imbalance we used several methods (see URLs) with the ‘as-only’ option to map caQTLs using allele-specific counts at feature SNPs.

Overlap of lead SNPs with TFBs. In the high-confidence set of 15,487 mapped causal interactions, we detected the lead variant for each downstream peak using the hierarchical model and PHM (see Supplementary Note Section 3.8) and selected 1,577 downstream peaks where lead SNP differed between the two models, excluding any peak where the lead variant was an INDEL or CNV. Then we generated the ATAC-seq cleavage (transposase cut site) around the lead SNPs (30 bp on either side).

To investigate motif disruption of the lead variants, we downloaded the 3,059 motifs from CisBP-RIP (version 1.02; see URLs). Within each chromatin accessibility peak, we generated all possible personal genome sequences using phased haplotypes of SNPs and INDELS for our 100 samples. We computed the position weight matrix (PWM) score for each motif and the posterior probability of transcription factor (TF) binding as follows:

\[
p(TF|\text{sequence}) = \frac{1}{Z_{\text{FW}}} \sum_{x \in \{0,1\}} \exp(q(TF,x)|\text{sequence}) - \frac{1}{Z_{\text{FW}}} \sum_{x \in \{0,1\}} \exp(q(TF,x))
\]

where \( Z_{\text{FW}} \) is the partition function for the motif, \( q(TF,x) \) is the score of the motif part of sequence within the peak, and \( Z_{\text{FW}} \) is the PWM score with background probability (0.25 for each nucleotide). We set the prior probability of transcription factor binding as 0.001 for any transcription factor, and defined a transcription factor as bound if \( p(TF|\text{sequence}) \) was greater than 0.5.

Enrichment analysis with PPP. Any enrichment analysis was carried out based on PPP for all \( j \) peak pairs. We compute a \( 2 \times 2 \) table of a binary annotation \( X_j \) (for example, if \( j \)-k peak pair within TAD then \( X_j = 1 \), otherwise 0) and the existence of causality between \( j \)-k peak, such that

\[
\begin{align*}
T &= \sum_{j,k} \left( X_j X_k \right) \left( 1 - X_j \right) \left( 1 - X_k \right) \\
&= \frac{X_j X_k}{\left( 1 - X_j \right) \left( 1 - X_k \right)} \\
&= \frac{X_j X_k}{1 - X_j - X_k + X_j X_k}
\end{align*}
\]

PPC for peak \( j \) and \( k \). We compute the odds ratio from the table to perform hypothesis testing. See Section 3.13 of the Supplementary Note for details.

Simulation strategy. We simulated 17,349,412 peak pairs under each of the 4 hypotheses: causality \((j \rightarrow k)\), causality \((k \rightarrow j)\), linkage, and pleiotropy. To simulate realistic LD, we used real genotype data of 100 samples. To simulate a caQTL at peak \( j \), a causal variant was chosen at random, weighted by the estimated variant-level prior from the real data. The effect size and standard deviation of the error of the simulated causal variant were the same as the estimated effect size and standard deviation for that variant from the simple linear regression and 2SLS of the real data. Chromatin accessibility for each sample at peak \( j \) was then simulated as a draw from a normal distribution, with mean set to the effect size times the genotype dose, and variance equal to the squared standard deviation. For the linkage hypothesis, we repeated this procedure for peak \( k \). For the pleiotropy hypothesis, we generated chromatin accessibility at peak \( k \) with the same causal variant at peak \( j \). For causality from peak \( j \) to \( k \), we used the 2SLS estimator of effect size and standard deviation to generate chromatin accessibility at peak \( k \).

We also assessed two other scenarios where our model assumptions are potentially violated. First, we generated chromatin accessibility with two causal variants for the focal peak under causality or pleiotropy, or four causal variants (two in each peak) under linkage. In addition, we simulated hybrid hypotheses where combinations of linkage, pleiotropy, and causality were considered. See Section 3.14 of the Supplementary Note for more detail.

To compare PHM to MR Steiger, we fit both models (PHM and MR Steiger) to the simulated data under the four different hypotheses (causality \((j \rightarrow k)\), causality \((k \rightarrow j)\), linkage, and pleiotropy), and tested their ability to distinguish the causality \((j \rightarrow k)\) peak pairs from each of the three other scenarios in turn. A positive call was set if the PPC \( > \) 0.5 for PHM. For the pleiotropy \((j \rightarrow k)\) peak pairs from each of the three other scenarios in turn. A positive call was set if the PPC \( > \) 0.5 for PHM. For the pleiotropy \((j \rightarrow k)\) peak pairs from each of the four replicates of the deletion lines (two replicates for D1 and D2 heterozygous genotypes).

Comparison with MPRA. We downloaded the table of combined LCL analysis for all 39,478 variants with MPRA. We extracted 842 variants that showed significant allelic imbalance according to the criteria applied in the paper (referred to as evMar). We then selected the lead eQTL variant for each gene based on the eQTL PPA and asked how many overlapped the validated evMar. When there were ties in PPA for the lead eQTL variants, we randomly selected one variant.

Editing of BLK/FAM167A locus (rs558245864-C>G) using CRISPR-Cas9. The LCL HG00146, which is homozygous for the reference rs558245864 allele, was nucleasefeected with an enhanced Cas9-2a-GFP plasmid and a guide RNA expression plasmid targeting the rs558245864 locus. Deletion clones were selected, expanded, and then subjected to ATAC-seq and RNA-seq. Methods for engineering of the rs558245864 locus are described in full in Sections 4.1–4.6 of the Supplementary Note.

Differential chromatin accessibility and expression analyses. We used DESeq to perform differential chromatin accessibility and differential expression analyses. We compared the two replicates of the parental line against the four replicates of the deletion lines (two replicates for D1 and D2 heterozygous lines, respectively). See Section 4.7 of the Supplementary Note for more detail.

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

Code availability. The PHM is available from GitHub (https://github.com/natsuhiko/PHM).

Data availability. ATAC-seq data for the 100 LCLs used in this study are available from the European Nucleotide Archive (ERP110508). All summary statistics are available from Zenodo (https://doi.org/10.5281/zenodo.1405945).

References.

44. Jiang, H., Lei, R., Ding, S. W. & Zha, S. Skever: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics 15, 182 (2014).
45. Ding, Z. et al. Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. PLoS Genet. 10, e1004798 (2014).
46. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
47. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
48. Kim, D. et al. TopHat2: accurate alignment of transcripts in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36 (2013).
49. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. Am. J. Hum. Genet. 98, 116–126 (2016).
50. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3, 95–98 (2016).
51. Hoffman, M. M. et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat. Methods 9, 473–476 (2012).
52. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods 9, 215–216 (2012).
53. Wood, S. N. Generalized Additive Models: An Introduction with R. (Chapman & Hall/CRC, Boca Raton, Florida, USA, 2006).
54. Weirbach, M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell 158, 1431–1443 (2014).
55. Anders, S. & Huber, W. Differential expression analysis for sequence count data. Genome Biol. 11, R106 (2010).
Experimental design

1. Sample size
   Describe how sample size was determined.
   
   We used 100 lymphoblastoid cell lines. To set the sample size, we first estimated the average variance explained by a lead chromatin accessibility quantitative trait locus (caQTL) in 100Kb cis window (h²=0.24) using 24 samples we previously sequenced (Kumasaka et al. Nature Genetics, 48, pages 206–213 (2016)). Then we computed the power by using the Chi-square distribution (https://genome.sph.umich.edu/wiki/Power_Calculations:_Quantitative_Traits). We estimated that a sample size of 100 would provide 99.8% power to detect caQTLs with h²=0.24.

2. Data exclusions
   Describe any data exclusions.
   0 samples were excluded

3. Replication
   Describe the measures taken to verify the reproducibility of the experimental findings.
   
   1 replicate for each of the 100 cell lines. We also generated 2 CRISPR engineered lines with 2 replicates

4. Randomization
   Describe how samples/organisms/participants were allocated into experimental groups.
   
   caQTL mapping was performed under only one experimental condition with the Mendelian Randomization technique.

5. Blinding
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   
   Our study did not contain any experimental grouping, and blinding was therefore not relevant.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.
6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a Confirmed

☐ The exact sample size \( n \) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)

☐ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ A statement indicating how many times each experiment was replicated

☐ The statistical test(s) used and whether they are one- or two-sided

*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*

☐ A description of any assumptions or corrections, such as an adjustment for multiple comparisons

☐ Test values indicating whether an effect is present

*Provide confidence intervals or give results of significance tests (e.g. \( P \) values) as exact values whenever appropriate and with effect sizes noted.*

☐ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)

☐ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

7. Software

Policy information about availability of computer code

Describe the software used to analyze the data in this study.

We implemented a novel Bayesian pairwise hierarchical model. The software is available from GitHub (https://github.com/natsuhiko/PHM).

We also used the following publicly available softwares:

- CHICAGO 1.1.8 (http://regulatorygenomicsgroup.org/chicago)
- Juicer Tools 0.7.5 (https://github.com/theaidenlab/juicer/wiki/Juicer-Tools-Quick-Start)
- Beagle 4.0 (https://faculty.washington.edu/browning/beagle/b4_0.html)
- DESeq (https://bioconductor.org/packages/release/bioc/html/DESeq.html)
- bwa 0.7.4 (https://sourceforge.net/projects/bio-bwa/files/)
- Bowtie2 2.2.4 (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml)
- TopHat2 2.0.13 (http://ccb.jhu.edu/software/tophat/index.shtml)
- RASQUAL (https://github.com/natsuhiko/rasqual)
- skewer 0.1.127 (https://github.com/relipmoc/skewer)

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). Nature Methods guidance for providing algorithms and software for publication provides further information on this topic.

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

No restrictions

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

We have not used any antibodies in this study.
10. Eukaryotic cell lines
   a. State the source of each eukaryotic cell line used.
   b. Describe the method of cell line authentication used.
   c. Report whether the cell lines were tested for mycoplasma contamination.
   d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

| All lymphoblastoid cell lines were obtained from Coriell (https://www.coriell.org/). |
| ATAC-seq data from each line was checked against known line genotypes. |
| The mycoplasma testing was performed by Coriell (https://www.coriell.org/0/sections/support/global/qcmyco.aspx?pgid=411) and the result was negative for all the cell lines (https://www.coriell.org/0/sections/support/global/qccells.aspx?pgid=409). |
| Provide a rationale for the use of commonly misidentified cell lines OR state that no commonly misidentified cell lines were used. |

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

   Provide all relevant details on animals and/or animal-derived materials used in the study.

   No animals were used in this study.

Policy information about studies involving human research participants

12. Description of human research participants

   Describe the covariate-relevant population characteristics of the human research participants.

   Cell lines derived from 100 healthy volunteers in England and Scotland; 49 males and 51 females. Age unknown.