8-23-2021

INFIMA leverages multi-omics model organism data to identify effector genes of human GWAS variants.

Chenyang Dong

Shane P Simonett

Sunyoung Shin

Donnie S Stapleton

Kathryn L Schueler

See next page for additional authors

Authors
Chenyang Dong, Shane P Simonett, Sunyoung Shin, Donnie S Stapleton, Kathryn L Schueler, Gary Churchill, Leina Lu, Xiaoxiao Liu, Fulai Jin, Yan Li, Alan D Attie, Mark P Keller, and Sündüz Keleş

Genome Biology

METHOD Open Access

INFIMA leverages multi-omics model organism data to identify effector genes of human GWAS variants

Chenyang Dong[1], Shane P. Simonett[2], Sunyoung Shin[3], Donnie S. Stapleton[2], Kathryn L. Schueler[2], Gary A. Churchill[4], Leina Lu[5], Xiaoxiao Liu[5], Fulai Jin[5], Yan Li[5], Alan D. Attie[2], Mark P. Keller[2*] and Sündüz Keleş[1,6*]

*Correspondence:
mark.keller@wisc.edu;
keles@stat.wisc.edu
[1]Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA
[2]Department of Biochemistry, University of Wisconsin-Madison, Madison, WI, USA
Full list of author information is available at the end of the article

Abstract

Genome-wide association studies reveal many non-coding variants associated with complex traits. However, model organism studies largely remain as an untapped resource for unveiling the effector genes of non-coding variants. We develop INFIMA, *I*ntegrative *Fi*ne-*Ma*pping, to pinpoint causal SNPs for diversity outbred (DO) mice eQTL by integrating founder mice multi-omics data including ATAC-seq, RNA-seq, footprinting, and in silico mutation analysis. We demonstrate INFIMA's superior performance compared to alternatives with human and mouse chromatin conformation capture datasets. We apply INFIMA to identify novel effector genes for GWAS variants associated with diabetes. The results of the application are available at http://www.statlab.wisc.edu/shiny/INFIMA/.

Keywords: Fine-mapping, Molecular quantitative trait loci, Genome-wide association studies, Pancreatic islets, Diversity outbred mouse, ATAC-seq, Generative probabilistic modeling, Transfer learning

Background

Vast majority of disease and complex human trait-associated single nucleotide polymorphisms (SNPs) identified through genome-wide association studies (GWAS) are non-coding [1]. This creates two key challenges for translation of genetic discoveries into disease mechanisms. GWAS have capitalized on large-scale genomic and epigenomic data to address the first challenge of interpreting non-coding risk SNPs and assigning them potential regulatory roles [2, 3]. In many cases, non-coding loci with risk SNPs span broad genomic regions that contain multiple genes [4]. This creates the second challenge of identifying the effector genes through which risk SNPs exert their impact on the phenotype, possibly via long-range chromatin interactions. With the advances in three-dimensional (3D) chromatin structure and interaction profiling, recent studies have successfully shown that a genetic variant is not necessarily causal for the nearest gene

[5, 6]. The consequence of this new perspective is a vast expanse of the set of candidate effector genes for a GWAS risk locus. In addition, the linkage disequilibrium (LD) [7] further complicates the elucidation of effector genes for most GWAS risk SNPs because the causal variant may not be the SNP with the strongest association, but one that is in high LD. Collectively, these challenges hinder the delineation of effector genes for the majority of GWAS risk SNPs.

The recent transcriptome-wide association studies (TWAS) that leverage reference expression panels led to notable progress in identifying candidate disease-associated genes [8, 9]. However, these approaches do not directly link the effector genes to SNPs. In addition, and perhaps more restrictively, they rely on reference transcriptomes which may not be readily available or are difficult to obtain for an array of disease-relevant tissues. Complementary to these, model organism studies continue to provide opportunities to unveil susceptibility genes and investigate findings from human GWAS. Specifically, progress during the last decade confirmed that evolutionary conservation can be used to discover regions of coding and non-coding DNA that are likely to have biological functions [10–12] and thus may harbor functional SNPs. In this paper, we leverage model organism multi-omics data, specifically, data from the diversity outbred (DO) mouse population [13], to develop a framework for identifying candidate effector genes of non-coding human GWAS SNPs.

The DO mouse population [13], a model organism resource derived from eight founder strains (`129`, `AJ`, `B6`, `CAST`, `NOD`, `NZO`, `PWK`, `WSB`), has been widely used to identify QTL for a variety of physiological and molecular phenotypes, including type 2 diabetes and gene expression in pancreatic islets [14–18]. These studies led to novel insights into the genetic architecture of islet gene regulation [14] and insulin secretion [19]. However, a key impediment to maximizing the results of these types of eQTL studies is the lack of genomic resolution required to pinpoint the causal variants and elucidate potential regulatory mechanisms. These in-bred genomes harbor long stretches of genetic variants in high LD [20]. While this is advantageous for achieving gene-level mapping because, compared to a human GWAS, comparatively fewer markers (i.e., tag SNPs) are needed to genotype a larger group of SNPs, it results in groups of SNPs with similarly high LOD scores. Consequently, it hinders identifying enhancer-sized regions (i.e., in the order of hundreds of bases) underlying the detected associations. For example, an eQTL marker with the highest LOD score was identified for the gene *Abcc8* (Fig. 1a), where `PWK` has the lowest allelic effect (Fig. 1b, DO-eQTL allelic effects estimated by R/qtl2 [21]). However, several SNPs within a 0.8-Mb sub-region are in high LD, i.e., at a level that greatly exceeds the applicability of existing GWAS fine-mapping methods [22, 23], and thus have similarly high LOD scores (Fig. 1c).

To facilitate fine-mapping of DO-islet eQTLs, we generated functional multi-omics data by assay for transposase-accessible chromatin using sequencing (ATAC-seq) [24] and transcriptome sequencing (RNA-seq) [25] from the islets of founder DO strains. Analysis of these individual data sets established widespread variation in chromatin architecture and gene expression in the DO founder strains. Next, we developed an integrative statistical model named INFIMA (*In*tegrative *Fi*ne-*Ma*pping with model organism multi-omics data) that leverages multiple multi-omics data modalities to elucidate causal variants underpinning the DO-islet eQTLs. INFIMA exploits differences of the candidate genetic variants in terms of their multi-omics data support such as the chromatin accessibility

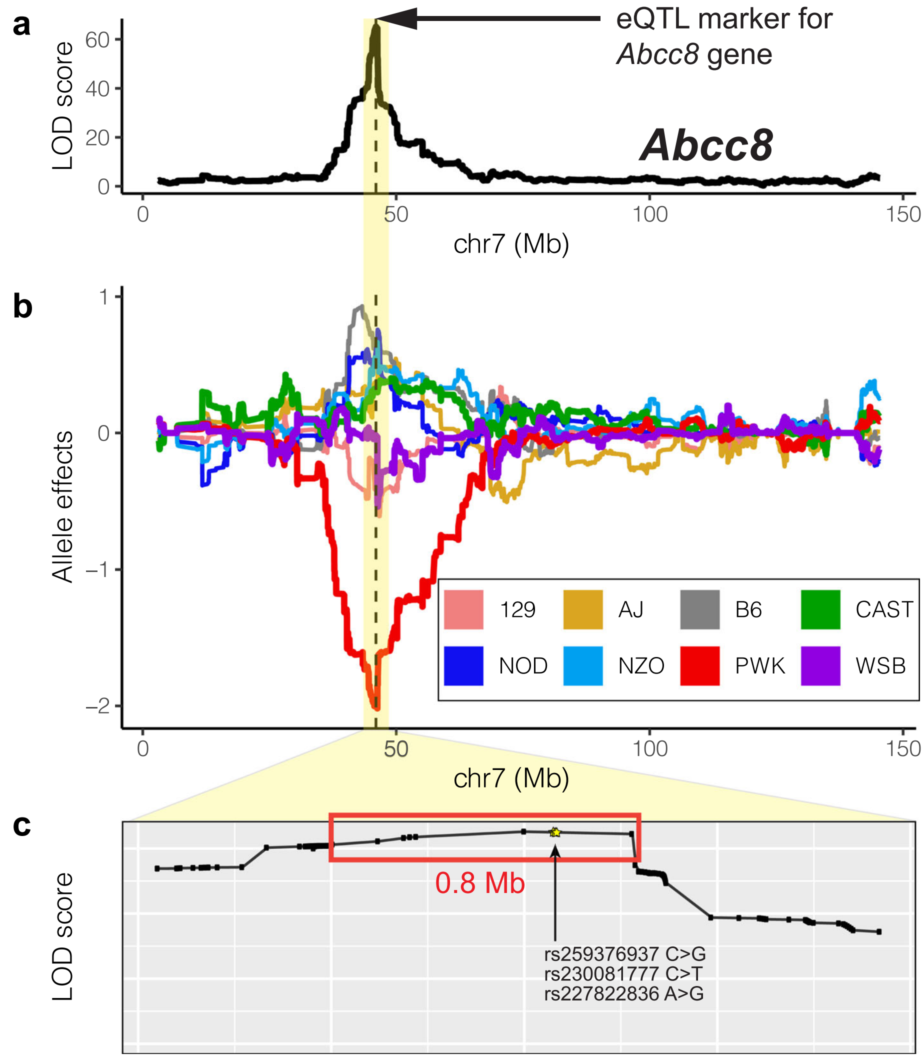

Fig. 1 Diversity outbred (DO) mice expression QTL (eQTL) analysis results at the *Abcc8* locus. **a** LOD score profile for *Abcc8* is maximized at marker location chr7:46,000,542 (dashed line) [14]. **b** Allele dependence for *Abcc8* local eQTL, where `PWK` harbors the low expression allele. **c** Zoomed in version of the LOD score profile from **a** at the SNP level. SNPs tied for the same highest LOD score are marked with the red box. Fine-mapped SNPs by INFIMA are highlighted by yellow stars

of the variant locations, correlations of chromatin accessibility, and transcriptome with variant genotypes and DO mice allelic expression patterns. As a result, it maps genetic variants within the DO founder strains to eQTL genes by quantifying how robustly the multi-omics data explains the allelic patterns observed in the eQTL analysis. Application of INFIMA to islet eQTLs identified in DO mice [14] revealed genetic variants that affect chromatin accessibility, and led to strain-specific expression differences. Leveraging our INFIMA-based fine-mapping of DO-islet eQTLs enabled us to nominate effector genes for ∼ 3.5% of the ∼ 15,000 human GWAS SNPs associated with diabetes. We validated INFIMA fine-mapping predictions with high throughput chromatin capture data from both mouse and human islets. Our results demonstrate that INFIMA provides a foundation for the critical task of capitalizing on model organism multi-omics data to elucidate target susceptibility genes of GWAS risk loci.

Results

ATAC-seq analysis reveals variable chromatin accessibility in islets of founder DO strains

We performed ATAC-seq to survey chromatin accessibility in pancreatic islets of both sexes of the eight founder DO strains (Fig. 2a; Methods). After quality control with transcription start site (TSS) enrichment analysis (Additional file 1: Figure S1) and data processing, we obtained 77.7 ± 4.1 million reads (excluding mitochondrial DNA) per sample which yielded a total of 51,014 accessible chromatin regions (Additional file 1: Figure S2). Specifically, ATAC-seq reads from 16 samples were aligned to the reference

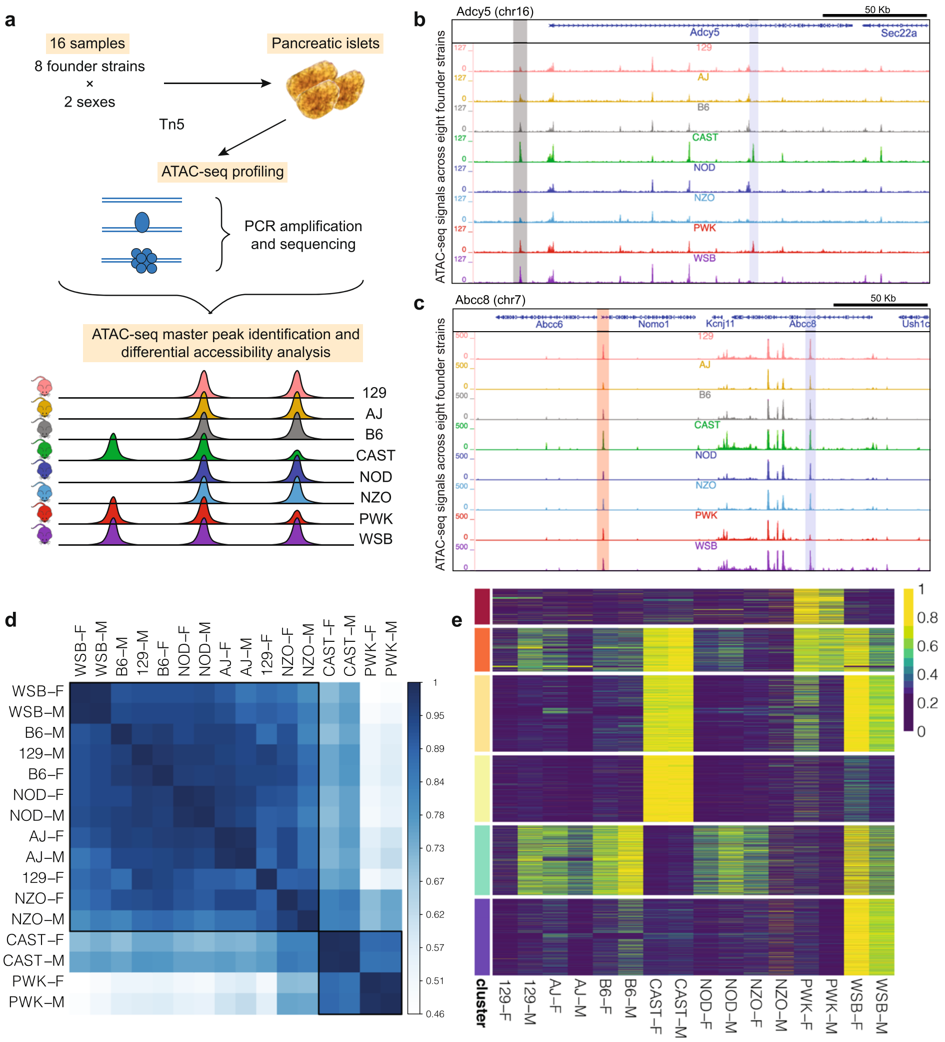

Fig. 2 Variable chromatin accessibility across founder DO strains. **a** Experimental overview and schematic of primary output for chromatin accessibility profiling of founder DO strains by ATAC-seq and differential accessibility analysis. **b**, **c** Genome browser displays of differentially accessible ATAC-seq peaks. **b** A differentially accessible distal intergenic ATAC-seq peak (translucent gray) and a `CAST-PWK` specific ATAC-seq peak at the *Adcy5* intron (translucent blue). **c** A differentially accessible ATAC-seq peak at the *Nomo1* promoter (translucent red) and an ATAC-seq peak less accessible in `PWK` at the *Abcc8* intron (translucent gray). **d** Heatmap of Pearson correlations between each pair of samples based on normalized chromatin accessibility cluster strains consistent with their genetic relatedness. Hierarchical clustering reveals the two clusters of strains outlined in black. **e** Differentially accessible regions (rows) in 16 samples (columns) of eight founder DO strains across two sexes. ATAC-seq peak scores are standardized to the [0, 1] range. Rows are clustered by k-means ($k = 10$). The six wild-derived clusters from top to bottom are: `PWK`, `CAST-PWK-WSB`, `CAST-WSB`, and `CAST`, absent in `CAST-PWK`, `WSB`. Additional file 1: Figure S2 is the full version of this figure

mouse genome (B6) assembly version mm10, yielding an average alignment rate of 92.3 ± 0.7% (Additional file 1: Table S1; Methods). To eliminate potential reference strain bias, we also aligned to individualized genomes, and observed, on average, only 0.86% difference (with a range of 0% and 3.66% across all alignments) between the two alignment strategies (Additional file 1: Table S2). Since these differences were not above the level one would expect from slight variation in alignment parameters [26], we used alignments to the reference mouse genome. We identified regions of accessible chromatin with MOSAiCS [27, 28] and applied irreproducible discovery rate (IDR) analysis [29] to generate ATAC-seq peak sets of each strain (at IDR of 0.05; Additional file 1: Supplementary Notes). The resulting peak sets were then merged to generate a combined peak list. Overall, we observed high concordance of chromatin accessibility (Pearson's $r \sim 0.95$) between the sexes for each strain (Additional file 1: Figure S3).

More than 70% of the accessible chromatin regions shared by all the strains corresponded to promoters and/or enhancers according to H3K27ac and H3K4me3 ChIP-seq based classification of tissue-specific promoters and enhancers from ENCODE (see URLs; Additional file 1: Supplementary Notes). In contrast, only 26.2% of the peaks that were specific to a single strain were annotated as promoters or enhancers (Additional file 1: Figs. S8 and S9). These results suggest that most of the strain-specific ATAC-seq peaks occur in strain-specific enhancers that are not captured in the existing list of mouse enhancers from ENCODE.

Among the 51,014 islet ATAC-seq peaks identified, 76.0% showed strain-dependent differences (FDR of 0.05; Methods) in an additive model of strain and sex effect. In contrast, only 50 peaks, 39 of which are located on chromosome X, exhibited sex effects at the same FDR level. The small number of peaks with sex effect is largely driven by the use of strain-specific male and female data to define consistent peaks within a strain and enable irreproducible error rate calculations for robust peak calling. Therefore, our analysis does not reflect the overall chromatin accessibility differences between the sexes of strains. Figs. 2b and c display a variety of peaks with strain differences. Specifically, an intronic region of *Adcy5* is more accessible in CAST and PWK compared to other strains, while a distal intergenic region exhibits more accessibility in CAST, PWK, and WSB (Fig. 2b). An intronic region of *Abcc8* is less accessible in PWK compared to other strains, whereas the *Nomo1* promoter is more accessible in CAST (Fig. 2c). We observed that differentially accessible chromatin regions were, overall, over-represented in promoters and under-represented in distal intergenic regions; however, these differentially accessible regions were more likely to be located in distal intergenic regions compared to peaks that did not exhibit significant strain effect (34.5% versus 28.8%, Additional file 1: Figure S10, quantified by regioneR [30] and ChIPseeker [31]). Clustering of the normalized ATAC-seq signals of the master peaks across the 16 samples (both sexes, eight strains) revealed a grouping structure largely consistent with the phylogenetic relationships among the founder strains (Fig. 2d). CAST, PWK and WSB are wild-derived subspecies of *M. musculus* [32], and represent ≥ 80% of the strain-specific peaks (Fig. 2e). These results suggest that the disproportionate amount of genetic variation contributed by these wild-derived strains mediate much of the differential chromatin accessibility we identified in islets.

Recent computational advances have enabled modeling of the magnitude and the shape of genome-wide chromatin accessibility profiles to infer putative transcription factor (TF) binding sites [33, 34]. We leveraged PIQ [33] to identify putative TF binding sites within the islet ATAC-seq peaks identified in the founder strains. Utilizing 744 known TF motifs in mouse and human, we identified high-confidence binding profiles for 12 TFs, Mzf1, Gata1, Yy1, Sox10, Nfic1, Ets1, Spib, Znf354c, Gata3, Spi1, Nfatc2, and the complex Arnt:Ahr (Fig. 3a). Nfatc2 is a well-established regulator of β-cell proliferation in mouse and human islets [36] and Yy1 [37], Sox10 [38], Ets1 [39], and Sbip1 [40] are TFs abundantly expressed in pancreatic islets. Recent work on a β-cell specific knockout of Arnt supports a key role in glucose-stimulated insulin release and islet gene expression [41, 42]. While the standard footprint analysis considers both the sequence motifs and ATAC-seq signals of binding sites, it cannot discriminate footprints of TFs with similar binding sites. To improve the specificity of the footprint analysis, we integrated the expression levels of TFs in islets from the founder DO strains, with abundant footprints identified from ATAC-seq profiles (Additional file 1: Figure S11), and the sequence similarity between TF motifs (Additional file 1: Figs. S12-S16). These additional criteria revealed that the binding motif of the transcriptional repressor Znf354c, which is not expressed in founder islets, is similar to that of Nkx2-2 (Additional file 1: Figure S17), a well-characterized TF that is abundantly expressed and plays a key role in islet development [43]. Thus, the Znf354c sites may be occupied by Nkx2-2. In addition, Gata1 and Gata3 are not expressed in founder islets, whereas Gata2, a closely related TF to Gata1 and Gata3 [44], is highly expressed (Additional file 1: Figure S11), suggesting that it may bind these sites. As expected, α-cell specific TFs such as Arx, Irx1, Irx2 showed a fewer number of footprints ($\leq$ 100) within the ATAC-seq peaks than β-cell specific TFs [45] (e.g., Pdx1, Mnx1, NFATC2 with an average of $\sim$ 4900 footprints). Additional β-cell specific TFs (e.g., Mafk, Pax4, Nkx2-2, Foxa2, Pax6, Nkx6-1) were collectively enriched in ATAC-seq peaks (p-value = 1.66e−2; Additional file 1: Supplementary Notes), albeit with fewer footprints ($\sim$ 1900).

Genetic variants associate with differential chromatin accessibility in islets of founder DO strains

We next evaluated the contribution of genetic variability present in the eight founder DO strains to differential chromatin accessibility within their islets. We associated the signal of 22,200 ATAC-seq peaks with at least one SNP, with the genotypes of the SNPs that they harbor. Although chromatin accessibility of a genomic region demarcated by an ATAC-seq peak can be modulated by SNPs in proximal and distal ATAC-seq peaks or genomic regions, we considered only the local SNPs to alleviate the multiple testing problem. As a result, we identified 47,062 local-ATAC-seq signal modulating variants (local-ATAC-MVs) within these 22,200 ATAC-seq peaks at FDR of 0.05 (Fig. 3b; "Methods"). The distribution of the number of local-ATAC-MVs within ATAC-seq peaks is right-skewed (Additional file 1: Figure S18) indicating that most peaks have one to three local-ATAC-MVs. Overall, 16,549 (42.7%) of the 38,749 differential peaks do not harbor any local-ATAC-MVs, suggesting that SNPs, or other factors, outside the ATAC-seq peaks contribute to their variable accessibility among the strains. The vast majority (95.6%) of the local-ATAC-MVs are associated with SNPs present in the three wild-derived strains (Additional file 1: Figure S19). Furthermore, a large percentage of the local-ATAC-MVs (77.3%) reside in distal intergenic or intronic regions, while 18.7% occur within promoters (Additional file 1: Figure S20).

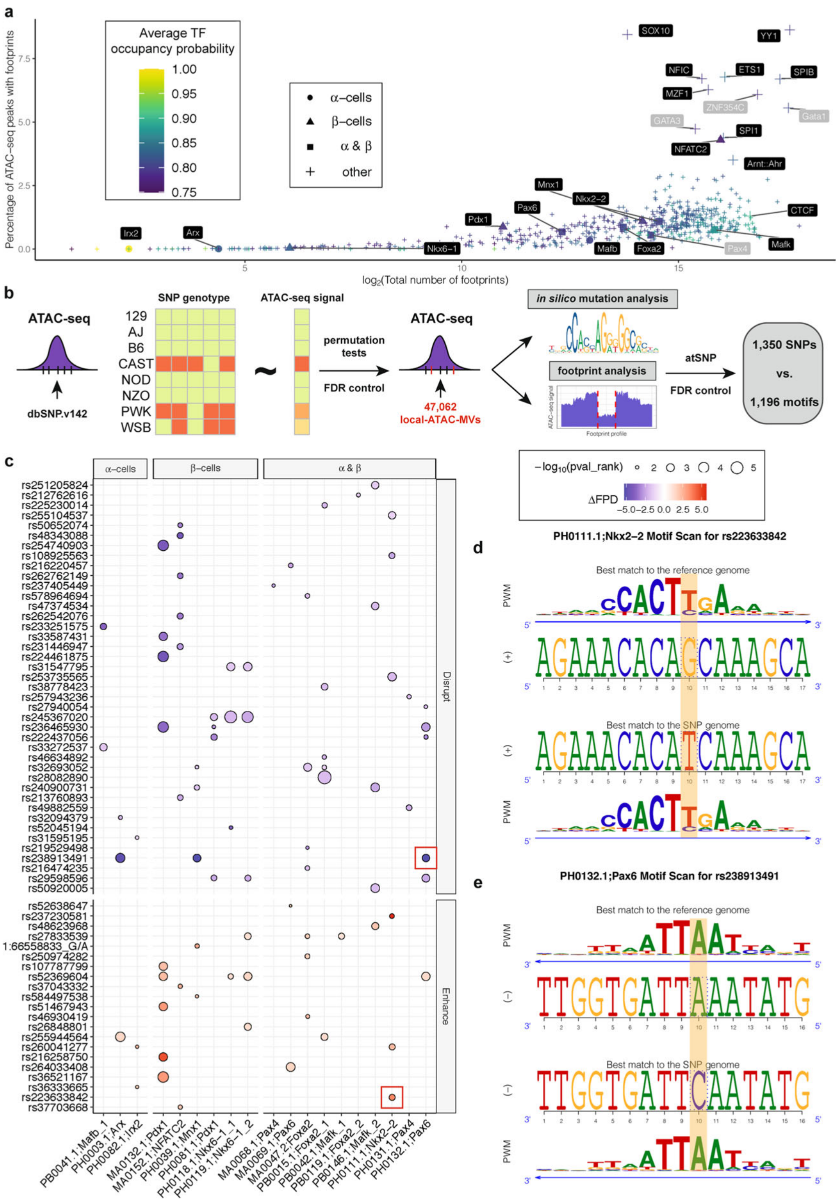

Fig. 3 Genetic variants associate with differential chromatin accessibility in the islets of DO founder strains. **a** Footprint analysis of ATAC-seq peaks. Transcription factors in black labels are expressed in founder islets (Additional file 1: Figure S11). **b** Overview of local-ATAC-MV identification, footprint, and in silico mutation analysis with atSNP. **c** A subset of SNPs (rows) that enhance/disrupt islet specific footprints (columns). The circles depict changes in the footprint depth with the SNP and reference alleles ($\Delta\mathrm{FPD} = \mathrm{FPD}_{SNP} - \mathrm{FPD}_{REF}$). Enhancement and disruption based on comparative FPD are depicted by shades of red and blue, respectively. The circle size indicates the significance of the impacts of the SNP alleles to the motif matches as calculated by atSNP [35]. Larger circles correspond to more significant changes in the motif match. Examples in **d** and **e** are highlighted by red boxes in **c**. **d** atSNP composite logo plot depicting Nkx2-2 binding site enhancement by SNP rs223633842 (G → T). **e** atSNP composite logo plot depicting Pax6 binding site disruption by SNP rs238913491 (A → C)

Genetic variants can affect gene regulation by changing TF binding affinities to genomic sequences [46]. To assess whether local-ATAC-MVs influence TF binding, we first performed an in silico mutation analysis of TF binding using atSNP [47]. In addition, for each SNP-motif pair, we computed the relative change in footprint depth (FPD), a measure of TF activity within ATAC-seq peaks [48], at the motif location across strains with the reference and alternative alleles (Additional file 1: Figure S21). Overall, we identified 8029 loci where local-ATAC-MVs significantly influenced the footprint at TF binding sites after multiplicity adjustment at FDR level of 0.05 (see Fig. 3b for the overall pipeline and Additional file 1: Figure S22 and S23 for evaluation of all the SNP-motif combinations; "Methods"). Despite the stringent multiplicity adjustment, we identified 62 local-ATAC-MVs that impact binding sites of TFs that are highly expressed in α, β, or other islet cell types [45] (Fig. 3c). For example, the SNP rs223633842 enhances a Nkx2-2 motif (Figs. 3d), whereas the SNP rs238913491 disrupts a Pax6 motif (Figs. 3e). Together, these results suggest that strain-specific differences in chromatin accessibility are affected by local-ATAC-MVs residing within ATAC-seq peaks and disrupting or enhancing TF binding.

RNA-seq analysis in islets of founder DO strains reveals variable transcriptome

After establishing widespread association of SNP genotypes with differential chromatin accessibility in the founder DO strains, we sequenced the islet transcriptome of the same eight strains. This enabled us to link local-ATAC-MVs with strain-dependent differences of nearby gene expression. We quantified the expression of 13,568 protein-coding genes with RSEM [49] (Additional file 1: Figure S24; Methods) which appropriately clustered the samples based on strain (Fig. 4a, Additional file 1: Figure S25). To maximize statistical power, we associated only the founder local-ATAC-MVs, instead of all the founder SNPs, with gene expression and identified 34,711 (73.8%) local-ATAC-MVs as associating with *cis* (as defined by 1 Mb neighborhood of genes) gene expression variation ("Methods"). The expression patterns of the genes associated with the local-ATAC-MVs are largely driven by alleles of wild-derived strains `CAST`, `PWK`, and `WSB`. Specifically, alleles of these three strains exert the most significant associations of the genes, i.e., the top 6 genotypes driven by these strains compromise 50.3% of the top associations of the 6418 local-ATAC-MV-associated genes (Fig. 4b). Next, we evaluated the distance between these genes and

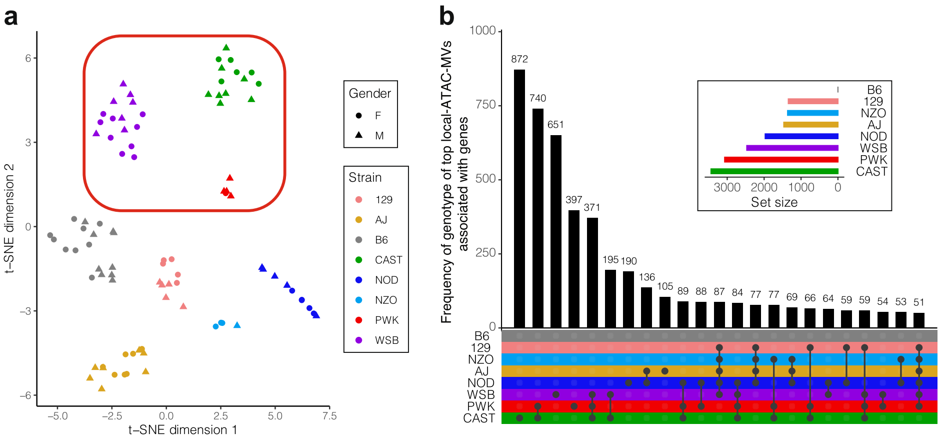

Fig. 4 Variable transcriptome across islets of founder DO strains. **a** Two-dimensional projection of the 91 founder RNA-seq samples with tSNE. Samples from wild-derived strains are boxed in with the red rectangle. **b** UpSet plot [50] for the frequencies of genotypes of local-ATAC-MVs associated with founder islet gene expression. Each gene with at least one significant association contributed its most significant local-ATAC-MV. Genotypes with frequencies less than 50 are not displayed

the proximal associated local-ATAC-MV loci. We found widespread contribution of promoters to expression variation across strains by harboring associated local-ATAC-MVs, i.e., 58% of the genes with at least one local-ATAC-MV association had associated local-ATAC-MV loci in their promoters (Additional file 1: Figure S26). We further investigated how well the differential ATAC-seq peaks within promoters explained the variation in gene expression across the strains. A pairwise differential expression analysis (Methods; FDR of 0.05) for the eight founder strains identified eGenes that were selective for one strain, i.e., B6 eGenes (expressed more in B6) and CAST eGenes (expressed more in CAST). As expected, B6 eGenes have higher promoter accessibility in B6, whereas CAST eGenes have higher promoter accessibility in CAST (Additional file 1: Figure S27). This concordance between strain-selective promoter accessibility and gene expression was observed, on average, for 67% of the eGenes (Additional file 1: Figure S28), suggesting a strong contribution of genetic variance of chromatin architecture within promoters to proximal gene regulation as also observed by others [51–54].

INFIMA model for fine-mapping DO mouse islet eQTLs by leveraging founder strain islet ATAC-seq and RNA-seq

The strain-dependent differences in accessible chromatin and transcriptome landscapes in islets of the DO founder strains allowed us to identify local-ATAC-MVs and their putative effector genes. Next, we leveraged this founder data to fine-map islet eQTLs from DO mice [14] (DO-eQTL, Fig. 1). We developed an integrative framework, named INFIMA, that exploits the high-resolution of the founder ATAC-seq profiles and gene expression data to delineate enhancer-sized loci as the most likely causal locus for individual DO-eQTLs.

INFIMA is an empirical Bayes model that estimates the linkages between founder local-ATAC-MVs and DO-eQTL genes for improving the resolution of DO-eQTL analysis. This is achieved by quantifying how well each non-coding SNP in high LD with the islet DO-eQTL marker explains the observed relationship between the allelic effect of the eQTL, islet ATAC-seq profile and gene expression among the founder strains proximal to the marker locus, and derived TF footprint results (Fig. 5a). This quantification enables inferring the likelihood of each candidate SNP, implied by the marker, to be causal. We summarize the INFIMA framework in Fig. 5 and provide the statistical details in this section.

A key step in the INFIMA framework is featurization of the DO-eQTL and founder data. We let $\mathcal{S}_n$ and $\mathcal{S}_w$ denote the index set for the classical in-bred (129, AJ, B6, NOD, NZO) and wild-derived strains (CAST, PWK, WSB), respectively, and let s be the index for the strains. Let G denote the total number of instances of DO-eQTL data, i.e., total number of gene-marker associations, $g = 1 \ldots G$ the index for the DO-eQTL gene of the g^{th} instance, p_g the number of candidate local-ATAC-MVs within a window size W of the eQTL marker for gene g, and $k = 1 \ldots p_g$ the index for the local-ATAC-MVs within this window (Fig. 5d). In our application, we have G = 10,936 contributed by 8046 eQTL markers versus 10,393 genes. A given DO-eQTL marker can have multiple DO genes that it associates with (Additional file 1: Figure S29). Let $\mathbf{Y_g}$ be an 8×1 vector of DO-eQTL allelic expression effects estimated with R/qtl2 [21] at marker location with the highest LOD score. We denote the features extracted from founder ATAC-seq and RNA-seq by

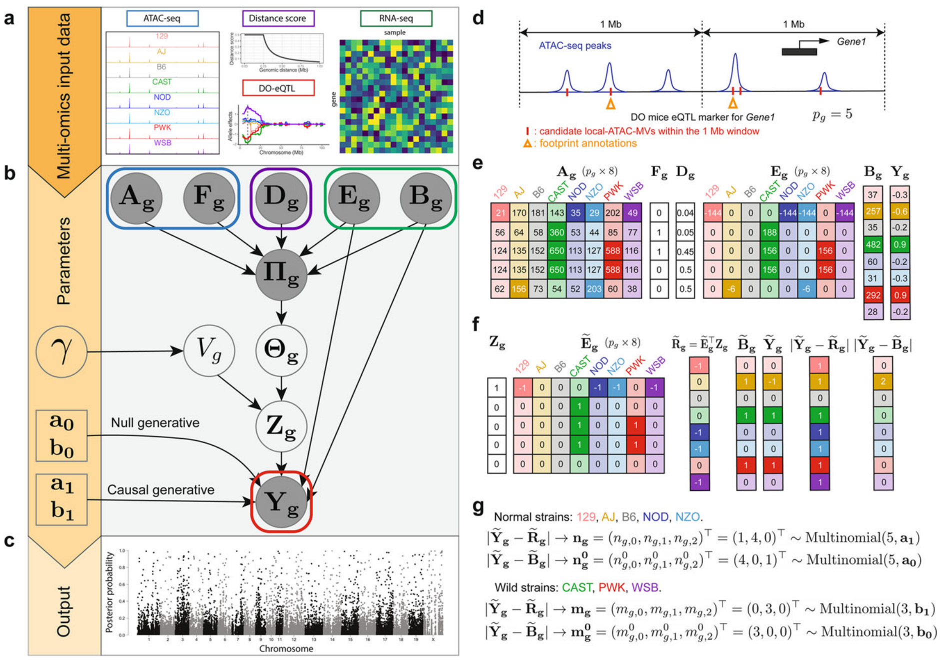
a
Multi-omics input data
ATAC-seq
Distance score
RNA-seq
129
AJ
B6
CAST
NOD
NZO
PWK
WSB
DO-eQTL
sample
gene
b
Parameters
A_g F_g D_g E_g B_g
Π_g
γ
V_g
Θ_g
Z_g
a_0 b_0
a_1 b_1
Null generative
Causal generative
Y_g
c
Output
Posterior probability
Chromosome
d
1 Mb
1 Mb
ATAC-seq peaks
Gene1
DO mice eQTL marker for Gene1
$p_g = 5$
: candidate local-ATAC-MVs within the 1 Mb window
: footprint annotations
e
A_g ($p_g \times 8$)
F_g D_g
E_g ($p_g \times 8$)
B_g Y_g
f
Z_g
$\widetilde{E}_g$ ($p_g \times 8$)
$\widetilde{R}_g = \widetilde{E}_g^\top Z_g$ $\widetilde{B}_g$ $\widetilde{Y}_g$ $|\widetilde{Y}_g - \widetilde{R}_g|$ $|\widetilde{Y}_g - \widetilde{B}_g|$
g
Normal strains: 129, AJ, B6, NOD, NZO.
$|\widetilde{Y}_g - \widetilde{R}_g| \to n_g = (n_{g,0}, n_{g,1}, n_{g,2})^\top = (1,4,0)^\top \sim \text{Multinomial}(5, a_1)$
$|\widetilde{Y}_g - \widetilde{B}_g| \to n_g^0 = (n_{g,0}^0, n_{g,1}^0, n_{g,2}^0)^\top = (4,0,1)^\top \sim \text{Multinomial}(5, a_0)$
Wild strains: CAST, PWK, WSB.
$|\widetilde{Y}_g - \widetilde{R}_g| \to m_g = (m_{g,0}, m_{g,1}, m_{g,2})^\top = (0,3,0)^\top \sim \text{Multinomial}(3, b_1)$
$|\widetilde{Y}_g - \widetilde{B}_g| \to m_g^0 = (m_{g,0}^0, m_{g,1}^0, m_{g,2}^0)^\top = (3,0,0)^\top \sim \text{Multinomial}(3, b_0)$

Fig. 5 INFIMA model overview. **a** Input data for the INFIMA model. INFIMA leverages summaries of model organism multi-omics data to model the relationship between allelic expression patterns ($\mathbf{Y_g}$) of DO-eQTL genes (i) and founder expression patterns ($\mathbf{B_g}$) under a null model of no causal SNPs ($V_g = 0$); (ii) and founder genotype expression patterns $\mathbf{R_g} = \mathbf{E_g}^\top \mathbf{Z_g}$, where $\mathbf{E_g}$ represents genotype effects of candidate SNPs on founder expression and $\mathbf{Z_g}$ encodes the causal SNP for gene g, under an alternative model with causal SNPs ($V_g = 1$), across all the genes indexed by g. **b** Plate representation of the INFIMA model summarizing data and the parameters. Blank circles: latent variables and parameters to be inferred; filled circles: observed variables. **c** INFIMA infers posterior probabilities of association for fine-mapping across all the candidate local-ATAC-MVs. **d**–**g** An example input of the INFIMA model. **d** An overview of a $W = 1$ Mb window around a DO-eQTL marker (centered dashed line) associated with *Gene1*. Two out of five candidate local-ATAC-MVs (red short lines) are decorated with comparative footprint effects (orange triangles). **e** Example input data for the five candidate local-ATAC-MVs. **f** An illustration of data trinarization and edit distance with multinomial distributions. The trinarization details can be found in Methods. **g** The edit distance variables quantify how many strains have 0, 1, or 2 absolute distances between $\widetilde{\mathbf{Y}}_\mathbf{g}$ and $\widetilde{\mathbf{R}}_\mathbf{g}/\widetilde{\mathbf{B}}_\mathbf{g}$ and are modeled by multinomial distributions

$\mathbf{X_g} = (\mathbf{A_g}, \mathbf{F_g}, \mathbf{D_g}, \mathbf{E_g}, \mathbf{B_g})$, where $\mathbf{A_g}$ is a $p_g \times 8$ matrix of the normalized ATAC-seq signal of the peak each candidate local-ATAC-MV resides in; $\mathbf{F_g}$ is the indicator vector ($p_g \times 1$) of whether or not the candidate local-ATAC-MV is affecting a footprint significantly, i.e., it is among the set of 8029 SNP-motif combinations identified in the aforementioned comparative footprint analysis; $\mathbf{D_g}$ is a $p_g \times 1$ vector of distance scores computed from the distances of local-ATAC-MV to the promoter of gene g; $\mathbf{E_g}$ is a $p_g \times 8$ matrix of founder RNA-seq genotype effects of these candidate SNPs for gene g (i.e., marginal regression of gene expression with respect to genotype); $\mathbf{B_g}$ denotes an 8×1 vector of the normalized founder expression of gene g. Figure 5e illustrates an example of the extracted features.

INFIMA model assumes at most one causal local-ATAC-MV per gene for a single marker-gene association. This is encoded by an unobserved random variable $V_g \in \{0, 1\}$ representing the number of causal local-ATAC-MVs for eQTL gene g. While this assumption can be relaxed at the expense of computational cost, it already enables multiple causal loci per gene when the gene is associated with multiple markers. Next, we define an additional unobserved $p_g \times 1$ random variable $\mathbf{Z_g} = (Z_{g,1}, Z_{g,2}, \ldots, Z_{g,p_g})^\top \in \{0, 1\}^{p_g}$ to denote the causal local-ATAC-MV. It immediately follows that $\mathbf{1}^\top \mathbf{Z_g} = V_g$. Finally, in the presence of a local-ATAC-MV, i.e., $V_g = 1$, we define $\mathbf{R_g} = \mathbf{E_g}^\top \mathbf{Z_g}$ as an 8×1 vector of the genotype effects of the causal SNP estimated from founder RNA-seq data for gene g.

For causal SNPs, we expect the allelic effects from DO mice ($\mathbf{Y_g}$ from the eQTL study) to be in agreement with the genotype effect of the causal SNP on the founder expression $\left(\mathbf{R_g}\right)$. We quantify this relationship with a causal generative model of $\mathbf{Y_g}$ conditional on $\mathbf{R_g}$. To avoid parametric assumptions needed for modeling continuous allelic effects $\mathbf{Y_g}$ and $\mathbf{R_g}$, in addition to supporting potential differences in distributions for the classical in-bred and wild-derived strains, we consider an edit distance model. Specifically, we convert $\mathbf{Y_g}$, $\mathbf{R_g}$, and $\mathbf{B_g}$ to trinary indicators encoding three levels of signal strengths: lower, the same, and higher than the reference strain B6 (Fig. 5f; Methods). After trinarizing the effects $\mathbf{Y_g}, \mathbf{R_g} \rightarrow \widetilde{\mathbf{Y}}_\mathbf{g}, \widetilde{\mathbf{R}}_\mathbf{g} \in \{-1, 0, +1\}^8$, we compute absolute values of the differences between their trinarized values $d_{g,s} = |\widetilde{Y}_{g,s} - \widetilde{R}_{g,s}|$ for each strain s. Then, we define the edit distance random variables $n_{g,i} = \sum_{s \in \mathcal{S}_n} \mathbb{I}\left\{d_{g,s} = i\right\}$ and $m_{g,i} = \sum_{s \in \mathcal{S}_w} \mathbb{I}\left\{d_{g,s} = i\right\}$ for $i = 0, 1, 2$. The set of edit distances $\left(n_{g,0}, n_{g,1}, n_{g,2}\right)$ represent numbers of 0's, 1's, and 2's in an experiment that corresponds to rolling a 3-sided dice 5 times. Hence, it follows that $\mathbf{n_g} = \left(n_{g,0}, n_{g,1}, n_{g,2}\right)^\top \sim \text{Multinomial}(5, \mathbf{a_1})$ and, similarly, $\mathbf{m_g} = \left(m_{g,0}, m_{g,1}, m_{g,2}\right)^\top \sim \text{Multinomial}(3, \mathbf{b_1})$. Here, $\mathbf{n_g} = (5, 0, 0)$ and $\mathbf{m_g} = (3, 0, 0)$ indicate that the allelic expression pattern in the DO mice completely matches the genotype effect estimated from the founders for gene g and the causal SNP specified by $\mathbf{Z_g}$. In this model, the lack of a candidate causal SNP is encoded by $V_g = 0$. However, some concordance between DO mice allelic expression $\mathbf{Y_g}$ and founder gene expression $\mathbf{B_g}$ is still warranted. Leveraging this intuition, we develop a null generative model for $\mathbf{Y_g}$ conditional on $\mathbf{B_g}$ with a similar trinarization approach as above. The trinarized data $\mathbf{Y_g}, \mathbf{B_g} \rightarrow \widetilde{\mathbf{Y}}_\mathbf{g}, \widetilde{\mathbf{B}}_\mathbf{g} \in \{-1, 0, +1\}^8$, with absolute differences $d^0_{g,s} = |\widetilde{Y}_{g,s} - \widetilde{B}_{g,s}|$ can be defined similarly as in $V_g = 1$. We define edit distance random variables $n^0_{g,i} = \sum_{s \in \mathcal{S}_n} \mathbb{I}\left\{d^0_{g,s} = i\right\}$ and $m^0_{g,i} = \sum_{s \in \mathcal{S}_w} \mathbb{I}\{d^0_{g,s} = i\}, i = 0, 1, 2$ and assume individual multinomial distributions $\mathbf{n^0_g} = \left(n^0_{g,0}, n^0_{g,1}, n^0_{g,2}\right)^\top \sim \text{Multinomial}(5, \mathbf{a_0})$, $\mathbf{m^0_g} = \left(m^0_{g,0}, m^0_{g,1}, m^0_{g,2}\right)^\top \sim \text{Multinomial}(3, \mathbf{b_0})$, parametrized by $\mathbf{a_0}$ and $\mathbf{b_0}$, respectively. Figure 5g illustrates an example of the trinarized data and the corresponding edit distances.

Next, we combine the two settings, namely $V_g = 1$ and $V_g = 0$, as a mixture over the two generative models. Specifically, we assume that the latent causal indicators are random draws, i.e., $V_g \overset{i.i.d.}{\sim} \text{Bernoulli}(\gamma)$, with the prior probability, $\gamma \in (0, 1)$, for the causal generative model. Let $\Theta_\mathbf{g} = \left(\theta_{g,1}, \theta_{g,2}, \ldots, \theta_{g,p_g}\right)^\top$ denote the probabilities that each candidate SNP is causal for gene g; then, $\mathbf{Z_g}$ is a mixture distribution over a multinomial distribution and a point mass at vector of 0's as

$$\mathbf{Z_g} | V_g, \Theta_\mathbf{g} \sim V_g \text{Multinomial}(1, \Theta_\mathbf{g}) + \left(1 - V_g\right) \delta_\mathbf{0}, \tag{1}$$

where $\delta_\mathbf{0}$ is a size p_g vector of 0's. To leverage the multi-omic data further, we assume a Dirichlet prior for the probability vector $\Theta_\mathbf{g} | \Pi_\mathbf{g} \sim \text{Dirichlet}(\Pi_\mathbf{g})$, where $\Pi_\mathbf{g} = \left(\Pi_{g,1}, \Pi_{g,2}, \ldots, \Pi_{g,p_g}\right)^\top$ is defined as

$$\Pi_{g,k} := F_{g,k} + D_{g,k} + |\text{cor}\left(\mathbf{A_{g,k}}, \mathbf{E_{g,k}}\right)| + |\text{cor}\left(\mathbf{A_{g,k}}, \mathbf{B_g}\right)| + 1. \tag{2}$$

Here, each component of $\Pi_{g,k}$ provides prior multi-omics information that contributes to the likelihood of SNP k to be causal for gene g. Specifically, $F_{g,k} \in \{0, 1\}$ indicates impact on a TF binding site; $D_{g,k} \in (0, 0.5]$ is a function of the distance between the DO-eQTL marker and the candidate SNP to utilize genomic distance; $|\mathrm{cor}\left(\mathbf{A_{g,k}}, \mathbf{E_{g,k}}\right)| \in [0, 1]$ measures the correlation between ATAC-seq signal of the peak harboring SNP k and the genotype effect of SNP k on founder expression; $|\mathrm{cor}\left(\mathbf{A_{g,k}}, \mathbf{B_g}\right)| \in [0, 1]$ similarly quantifies the correlation between ATAC-seq signal and gene expression in the founder strains.

The combined generative model for DO-eQTL effect size $\mathbf{Y_g}$ is then given by

$$\widetilde{\mathbf{Y}}_{\mathbf{g}}|\mathbf{Z_g}, \mathbf{E_g}, \mathbf{B_g}, \mathbf{a_0}, \mathbf{b_0}, \mathbf{a_1}, \mathbf{b_1} \sim \mathbb{I}\left(\mathbf{Z_g} \neq \mathbf{0}\right) f_{\mathbf{a_1},\mathbf{b_1}}\left(\mathbf{n_g}, \mathbf{m_g}\right) + \mathbb{I}\left(\mathbf{Z_g} = \mathbf{0}\right) f_{\mathbf{a_0},\mathbf{b_0}}\left(\mathbf{n_g^0}, \mathbf{m_g^0}\right), \tag{3}$$

where f_{a_x,b_y} denotes the product of Multinomial probability distribution functions parametrized by a_x and b_y. In summary, INFIMA model takes as input DO-eQTL results, summarized functional data from RNA-seq and ATAC-seq analysis of founder strains, as well as ATAC-seq-based comparative footprint and in silico mutation analysis of SNPs and outputs SNP-level quantifications (Fig. 5c).

Simulations reveal improved statistical power and fine-mapping with INFIMA

We first evaluated INFIMA for its ability to improve statistical power of fine-mapping and identification of credible sets of SNPs in marker eQTL applications. We designed Data-driven simulations where the parameters of the generative model are set based on the actual DO-eQTL and summarized founder strain multi-omics data from ATAC-seq, RNA-seq, and comparative footprint and in silico mutation analysis. We varied the prior information extracted from the multi-omics data to be non-informative (NI), moderately informative (MI), and highly informative (HI) by varying the information contributed by the comparative footprint analysis ("Methods"). This allowed modulation of the informativeness of the prior parameters without considering generative models for summaries extracted from ATAC-seq and RNA-seq data. INFIMA model has two key inference variables: $V_g \in \{0, 1\}$ which encodes whether or not a gene has a causal SNP, and $\mathbf{Z_g} \in \{0, 1\}^{p_g}$ which encodes the causal SNP. Although the prior parameter γ for V_g does not depend on the summarized multi-omics data (i.e., is expected to be insensitive to the prior information), varying levels of informativeness in the multi-omics data yield improved area under receiving operating characteristics and precision recall curves, with an average of $0.61 \pm 0.079\%$ improvement in power from moderately to highly informative setting (Additional file 1: Figure S30). Since INFIMA leverages the multi-omics data to specifically infer $\mathbf{Z_g}$ by informing the prior probabilities of causal SNPs, we assessed the impact of levels of informativeness of the priors on fine-mapping. Specifically, we considered the most and least likely causal associations inferred by INFIMA for each gene as "Most Likely", local-ATAC-MV with the highest posterior probability of being causal, and "Least Likely", local-ATAC-MV with the lowest posterior probability of being causal. We compared these INFIMA strategies with three intuitive and model-free baseline strategies of selecting causal SNPs as "Random", a randomly selected local-ATAC-MV; "Closest to Marker", local-ATAC-MV closest to the DO-eQTL marker in genomic distance; and "Closest to Gene", local-ATAC-MV closest to the gene promoter in genomic distance. This comparison revealed that INFIMA predictions provide markedly better fine-mapping compared to baseline strategies regardless of the level of informativeness of the priors. Specifically, the "Most Likely" selection by INFIMA provided the smallest credible proportion (the minimum proportion of ranked candidate local-ATAC-MVs required to encompass the causal variant). The NI, MI, and HI settings yielded 33.90%, 22.22%, and 14.04% credible proportions, respectively (Fig. 6), compared to the minimum of 52.48%, 48.32%, and 50.00% achievable with the baseline strategies. Interestingly, even when the priors are non-informative (NI setting), the INFIMA-produced credible set is, on average, 29.1% smaller than the smallest set that can be achieved by the baseline strategies (33.90% by NI vs. 48.32% by MI). As expected, the least likely predictions with INFIMA performed worse than baseline strategies, confirming INFIMA's ability to rank local-ATAC-MVs with respect to their causal potential. Overall, these simulations highlighted the significance of integrating multi-omics data into fine-mapping.

INFIMA outperforms alternatives for fine-mapping DO mouse eQTLs

We fit INFIMA model with a 1-Mb window size (W) around DO-eQTL markers across all the G=10,936 gene-marker associations (8046 eQTL markers and 10,393 genes). This resulted in a right-skewed distribution for the number of candidate local-ATAC-MVs within a window (Fig. 7a, median = 36.0, sd = 26.9). Figure 7b summarizes the estimated posterior probabilities of having a causal local-ATAC-MV, i.e., $\widehat{V}_g$, across the genes. It indicates that INFIMA infers a causal local-ATAC-MV for 3846 (38.0%) DO-eQTL genes at FDR of 0.05.

We further summarized INFIMA results as we have done for the simulations by identifying the most likely and least likely causal local-ATAC-MVs for genes with an inferred causal SNP, and compared these with the baseline strategies outlined in the simulations. In addition to these baseline methods, we also considered two recent human GWAS fine-mapping methods DAP-G [55, 56] and SuSiE [57], both of which have demonstrated best performances in human GWAS fine-mapping studies. We initially considered applying DAP-G and SuSiE to all the SNPs tagged by the eQTL marker at the individual locus without restricting the set of SNPs to local-ATAC-MVs and by utilizing the multi-omics prior

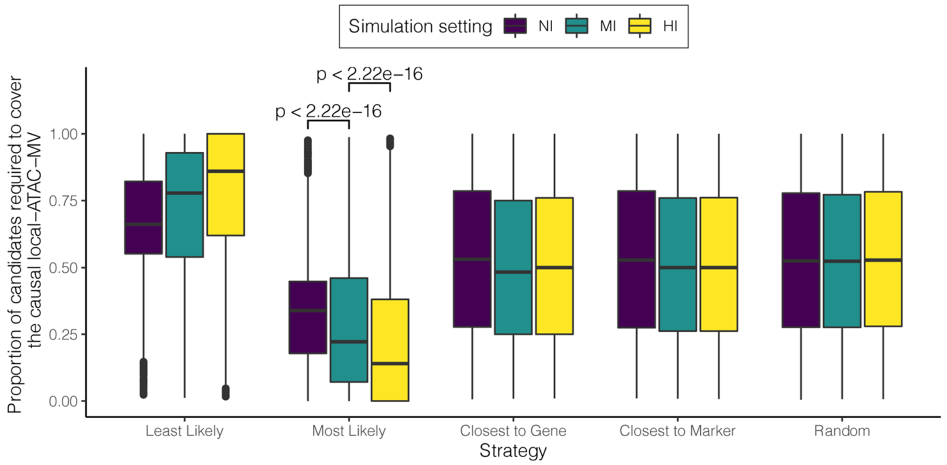

Fig. 6 Simulations reveal improved statistical power and fine-mapping with INFIMA. Comparison of the fine-mapping performances of five strategies across three simulation settings: NI, non-informative; MI, moderately informative; and HI, highly informative. The *y*-axis reports, as the performance metric, the proportion of candidate local-ATAC-MVs required in the credible set to cover the causal SNPs

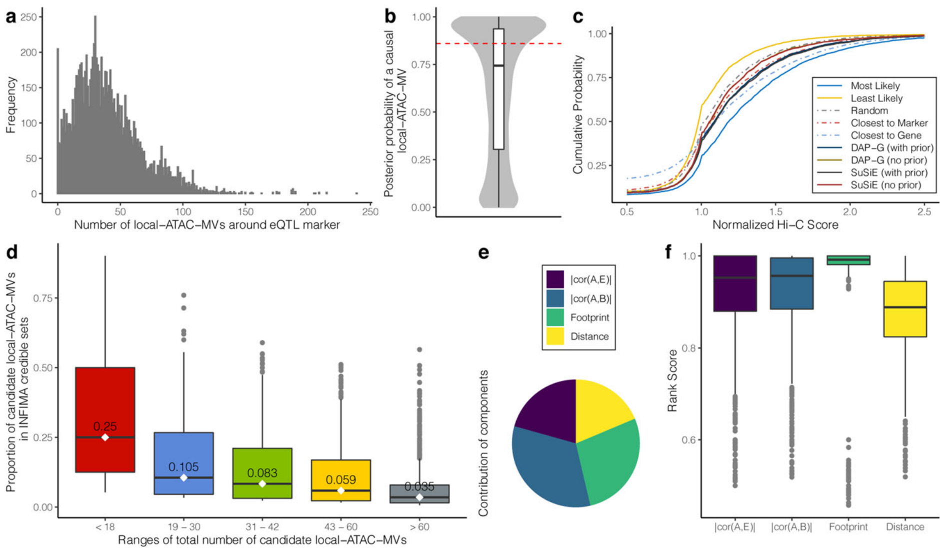

Fig. 7 INFIMA outperforms alternatives for fine-mapping DO mouse eQTLs. **a** Histogram of numbers of local-ATAC-MVs around DO-eQTL markers with window size of $W = 1$ Mb. **b** Boxplot of INFIMA posterior probabilities of association, $\widehat{V}_g$, across all genes. Red dashed line depicts the posterior probability cutoff for FDR of 0.05. **c** Evaluation of fine-mapping strategies with empirical cumulative distribution of normalized easy Hi-C scores. "Most Likely" and "Least Likely" refer to most and least likely predictions from INFIMA, respectively. **d** Boxplots depict the proportion of the candidate causal local-ATAC-MVs that are included in the credible set by INFIMA stratified by the size of candidate sets. The intervals on the *x*-axis are from quantile bins (20%, 40%, 60%, 80%, 100% percentiles) of number of local-ATAC-MVs around the eQTL marker, p_g. Median values are displayed on each boxplot. **e** Proportion of times each of the individual multi-omic components are the leading contributors to the INFIMA prior probability of causality: correlation between ATAC-seq signal and founder eQTL effect sizes $|cor(A, E)|$, 0.207; correlation between ATAC-seq signal and founder gene expression $|cor(A, B)|$, 0.330; footprint, 0.277; and distance, 0.186. **f** The rank scores of the inferred causal local-ATAC-MVs when individual components are the top ranking contributors. The higher the rank scores are, the more INFIMA weights in the component when inferring causal local-ATAC-MVs.

on the full set of SNPs. However, both methods failed to generate credible sets under this setting (Additional file 1: Supplementary Notes) owing to the LD structure of the DO mice (Additional file 1: Figure S31). Therefore, we reduced the candidate SNP set to local-ATAC-MVs for fine-mapping with DAP-G and SuSiE. We leveraged high-resolution easy Hi-C data, processed with a recent computational pipeline [58], from mouse islets and computed the empirical cumulative distribution curve of Hi-C signal between the DO-eQTL genes and their selected local-ATAC-MVs. We expect the local-ATAC-MVs that are likely to be true positives to interact with the gene promoters and, as a result, to exhibit higher Hi-C signal compared to competing approaches. Figure 7c depicts that the "Most Likely" selection by INFIMA outperforms the baseline predictions while the "Least Likely" selection by INFIMA performs worse than the baselines, highlighting an overall goodness-of-fit by INFIMA. The cumulative distribution curve of the "Most Likely" selection is significantly distinct from the baseline strategies (quantified by three different metrics: Kolmogorov-Smirov test, Kullback-Leibler (KL) divergence, and chi-squared test, Additional file 1: Table S4-S6, Addition file 1: Figure S32), confirming that INFIMA prediction of local-ATAC-MVs for DO-eQTL genes tend to be supported by higher Hi-C interaction signals. While the performances of DAP-G and SuSiE improve markedly with the INFIMA multi-omics data prior, they still perform worse than the baseline "Closest to Gene" and are significantly inferior to INFIMA. This is likely attributable to the large numbers of local-ATAC-MVs that are in perfect LD in DO mice compared to typical human GWAS fine-mapping studies (Additional file 1: Figure S33). Hi-C contacts standardized to [0, 1] for each DO-eQTL gene to enable comparison across genes indicate that, concordant with the overall Hi-C score distribution comparison, the "Most Likely" and the "Least Likely" selections by INFIMA harbor the highest and lowest ranked Hi-C scores, respectively (Additional file 1: Figure S34).

After validating that INFIMA inferred causal local-ATAC-MVs are significantly better than those identified by the baseline and alternative strategies, we evaluated the impact on fine-mapping. INFIMA is able to reduce the size of the credible set of local-ATAC-MVs tagged by a marker by 96.5% when $p_g > 60$. When the set size, p_g, is ≤ 18 (the lowest 20%), INFIMA reduces the size of the set of candidate local-ATAC-MVs by 75.0% (Fig. 7d). These are significant reductions at both the high and low ends of the size of the tagged local-ATAC-MV sets of a marker as it markedly reduces the number of loci for follow-up.

Since the multi-omics data INFIMA leverages to inform SNP prior probability of causality is multi-component, we asked whether the individual components contributed differently to the learned priors, i.e., $\Pi_{\mathbf{g}}$. Specifically, for each causal local-ATAC-MV of gene g, we ranked each of the individual components across the same category of components from all the competing p_g local-ATAC-MVs in ascending order, calculated a rank score[1] by normalizing with p_g, and reported the highest ranking contributor for the causal local-ATAC-MV as the component with the highest rank score. We found that, for only 20.1% of the causal local-ATAC-MVs, the Distance is the highest ranking contributor to the prior. The correlation between ATAC-seq signal and gene expression, i.e., $|\text{cor}(A, B)|$, contributes the most at 33.0% (Fig. 7e). Figure 7f shows that when Distance is the leading contributor, the median rank scores of the causal local-ATAC-MV, at 0.889, is lower than other components. This further demonstrates that INFIMA is not biased towards the local-ATAC-MVs closest to the genes. Interestingly, the Footprint component, with the highest median rank score of 0.992 (Fig. 7f), exerts a salient impact on INFIMA's ability to discriminate among the set of candidate causal local-ATAC-MVs.

INFIMA generates candidate susceptibility genes for human GWAS SNPs

The INFIMA model links ATAC-seq peaks and local-ATAC-MVs to candidate effector genes by fine-mapping DO-eQTLs. Next, we asked whether this approach can be leveraged to assign putative target genes in islets for non-coding human GWAS SNPs associated with diabetes.

Specifically, we considered 14,434 SNPs associated with 16 diabetes-related physiological traits from human GWAS [59] (Additional file 1: Figure S35). We employed a two-step peak-based strategy to lift-over human GWAS SNPs to syntenic sequences in the mouse genome. We first lifted-over the GWAS SNPs directly using the UCSC lift-over tool (see URLs) and identified the nearest mouse ATAC-seq peak to the syntenic loci. The remaining GWAS SNPs (81.0%) that did not directly lift-over to the mouse genome were first linked to their nearest human islet ATAC-seq peaks [60] and the peaks were lifted-over to mouse and linked to the nearest mouse islet ATAC-seq peak within 10 Kb (Additional file 1: Figure S36; "Methods"). This resulted in syntenic links between 4268 GWAS SNPs (2749 direct and 1519 ATAC-seq peak-based) and 1532 mouse ATAC-seq peaks. Several

[1] `rank()` function in R was used with `ties.method = "average"`, and then normalized the resulting score by p_g. The rank score is $\in [0, 1]$ and larger magnitudes correspond to higher ranks.

studies [61–63] have proposed that genomic compartment annotations associated with promoters are largely conserved between human and mouse. Similarly, distal regulatory elements across species are more likely to reside in regions with similar genomic compartment annotations [64, 65]. Therefore, we asked if these diabetes-associated syntenic regions had common genomic compartment annotations with their human counterparts. Overall, we observed a large degree of genomic annotation conservation for diabetes-associated GWAS SNPs (Fig. 8a). Specifically, ∼ 70% of the local-ATAC-MVs syntenic to intronic/distal/promoter GWAS SNPs exhibited the same genomic compartment annotation in mouse. Furthermore, we found that mouse syntenic regions of GWAS SNPs associated with diabetes-linked traits, e.g., type 1 diabetes, type 2 diabetes, body mass index, and body weight were enriched for local-ATAC-MVs (Fig. 8b; Bonferroni of 0.05, "Methods"). In contrast, mouse syntenic regions of a separate group of control SNPs associated with non-diabetic traits (e.g., Alzheimer's disease, and white blood cell counts) were not enriched with local-ATAC-MVs. This enrichment analysis further confirmed the relevance of the local-ATAC-MVs discovered in the mouse for the diabetes-associated human GWAS SNPs.

Next, we used INFIMA to predict effector genes of diabetes-associated GWAS SNPs. Among the 1532 mouse ATAC-seq peaks syntenic to GWAS variants, 737 contained local-ATAC-MVs. Of these, 548 were causally linked to at least one DO-eQTL gene, with 18.1% linked to a single gene (Additional file 1: Figure S37). This generated a set of human gene orthologs as candidate effectors of GWAS SNPs. We next used human islet promoter capture Hi-C data (pcHi-C) [66] and assessed whether pcHi-C interactions supported the inferred GWAS SNP-effector gene pairs (Additional file 1: Figure S38). First, we observed that the indirect peak-based lift-over strategy did not exhibit any discernible difference from the direct lift-over in terms of pcHi-C validation (Fisher's exact test p-value = 0.848). Next, we compared INFIMA effector gene predictions for these human GWAS SNPs with two baseline strategies: (1) linking mouse ATAC-seq peaks syntenic to GWAS variants to their nearest genes instead of INFIMA predictions and (2) linking human GWAS SNPs to their nearest genes without going through model organism data and INFIMA predictions. We observed that INFIMA predictions were markedly better supported by the pcHi-C data (Fisher's exact test p-values of 3.48e-96 and 1.20e-21 for comparisons of INFIMA predictions to strategy (1) and (2), respectively; "Methods").

Overall, we identified putative effector genes for 587 GWAS SNPs, 499 of which were supported by the candidate effector gene promoter regions exhibiting significant Hi-C signal [66] with either the corresponding GWAS SNPs or human ATAC-seq peaks at enhancer regions (Fig. 8c; "Methods"). Among these effector genes are *ABCC8*, *KCNJ11*, *PDX1*, *ADCY5*, and *KCNQ1*, which are recognized as pancreatic β-cell genes strongly associated with type 2 diabetes [67, 68]. The *ABCC8* promoter is linked to a distal intergenic GWAS SNP rs1557765 (body mass index) as well as three *KCNJ11* intronic GWAS SNPs rs5215, rs5219, and rs757110 (type 2 diabetes) by pcHi-C data. These three human SNPs are syntenic to rs25937937, rs230081777, and rs227822836 in mouse and are identified by INFIMA as causal for a *Abcc8* DO-eQTL, the homolog to human *ABCC8* (Figs. 8d and f). In addition to nominating candidate effector genes, INFIMA analysis also facilitates comparison of potential impacts of human GWAS SNPs and their syntenic mouse local-ATAC-MVs on transcription factor binding. For example, atSNP search [69] results on human SNPs rs5215 and rs1557765 indicate that both rs1557765 and rs5215 lead to

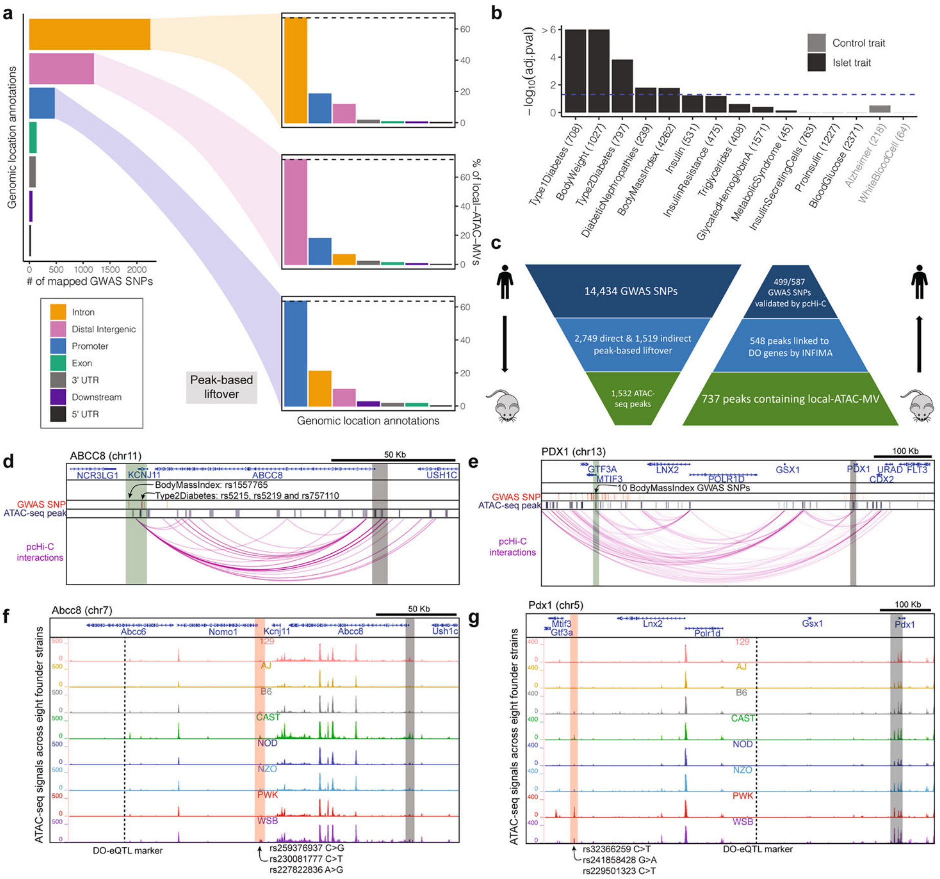

Fig. 8 INFIMA generates candidate susceptibility genes for human GWAS SNPs. **a** Comparison of genomic location annotations between human GWAS SNPs and the orthologous mouse genetic variants. Left: numbers of mapped GWAS SNPs within intronic, distal, promoter, exonic, UTR, and downstream genomic locations. Right: Barplots of the local-ATAC-MVs mapping to the intronic, distal, and promoter groups of GWAS SNPs, highlighting marked conservation of genomic location types. **b** Mapped GWAS SNPs are enriched for local-ATAC-MVs. The numbers in parentheses depict the numbers of GWAS SNPs for each trait and the blue dashed line marks the threshold for the Bonferroni adjusted cutoff at 0.05. **c** Summary of validation of INFIMA suggested susceptibility genes for human GWAS SNPs mapped to mouse with promoter capture Hi-C (pcHi-C). **d**, **f** pcHi-C links *ABCC8* promoter to 4 GWAS SNPs which map to 3 mouse local-ATAC-MVs with INFIMA predicted effector gene *Abcc8*. **e**, **g** Promoter capture Hi-C links *PDX1* promoter to 10 distal GWAS SNPs which map to 3 mouse local-ATAC-MVs with INFIMA predicted effector gene of *Pdx1*. **d**, **e** Human genome depictions of interactions of distal GWAS SNPs (translucent green) with the *ABCC8* and *PDX1* promoters (translucent gray), together with the human ATAC-seq peaks. **f**, **g** Mouse genome depiction of ATAC-seq signal for local-ATAC-MVs (translucent red) where INFIMA fine-maps DO-eQTL marker (dashed line) linked to genes *Abcc8* and *Pdx1* (promoters highlighted in translucent gray)

better sequence motifs for TCF7L2 (atSNP p-values of 5.99e−3 and 6.33e−4 for motif enhancement) and, furthermore, rs5215 also results in a better sequence motif for YY1 (atSNP p-value of 1.68e−3, Additional file 1: Figure S39a). Similarly, their syntenic mouse local-ATAC-MVs rs227822836 and rs230081777 enhance the binding sites for orthologous Tcf712 and Yy1 (atSNP p-values of 2.03e−2 and 8.53e−3 for motif enhancement; Additional file 1: Figure S39b).

pcHi-C data supports a chromatin loop that links *PDX1*, deficiency of which associates with β-cell dysfunction [70], to 10 GWAS SNPs rs1924074, rs9581853, rs9579083, rs9319366, rs9581854, rs4771122, rs12584061, rs12585587, rs9581856, and rs9579084 (also associates with body mass index) at promoter and intronic regions of *MTIF3*. These GWAS SNPs are lifted-over to a mouse locus, with local-ATAC-MVs rs32366259, rs241858428, and rs229501323, and for which INFIMA identifies *Pdx1* as the potential effector (Figs. 8e and g). We further observe that TFAP2A, GABPA, and HIC1 motifs are disrupted while CREB1, NFYA, TP53, NKX3-2, and EGR1 motifs are enhanced by the aforementioned human GWAS SNPs and their syntenic mouse local-ATAC-MVs, suggesting orthologous TF bindings (Additional file 1: Figure S40-S47).

In addition to these examples where the human GWAS SNPs with inferred effector genes are likely to enhance or disrupt TF binding sites, our results include cases where the SNPs exert their effects on expression through H3K27ac modification which is one of the enhancer-defining histone modifications. An example of this is type 2 diabetes GWAS SNP rs11708067 for which INFIMA analysis identified *ADCY5* as the effector gene (Additional file 1: Figure S48). This SNP was shown to contribute to type 2 diabetes by disrupting an islet enhancer and, consequently, resulting in reduction of *ADCY5* expression [71]. In addition, *ADCY5* was also inferred as the effector gene for SNPs rs11708903, rs6438788, and rs4450740 associated with blood glucose and insulin-secreting cells and residing in the intronic region of *ADCY5*. Finally, supporting data for *KCNQ1*, a susceptibility gene for type 2 diabetes [72], is provided in Additional file 1: Figure S49.

Discussion

While advances in genome sequencing improved the power of GWAS studies, elucidating which genes GWAS SNPs might be impacting is still a critical barrier for fully unleashing the power of GWAS. Recent large-scale and innovative efforts that leverage reference transciptome datasets to impute gene expression in GWAS cohorts and leverage co-localization with GWAS results have been successful in suggesting gene-level associations [73–75]. However, these studies are limited by the availability of reference transcriptomes in relevant tissues and accurate predictive models of gene expression. In a complementary approach, we leveraged model organism multi-omics data for this challenging task. Specifically, we developed INFIMA as a statistically grounded framework to capitalize on multi-omics functional data and fine-map model organism molecular quantitative trait loci. Application of INFIMA to DO mouse islet eQTLs fine-mapped previously identified eQTLs. Next, we asked whether INFIMA islet eQTL fine-mapping results could be transferred to human to infer effector genes of non-coding human GWAS SNPs. This reasoning is instigated by the observation that non-coding human GWAS SNPs associated with pancreatic islet functions are overwhelmingly enriched in syntenic accessible chromatin regions in islets of founder DO strains, suggesting potential functional relatedness among the two sets of non-coding regions. We utilized INFIMA resolved DO mouse SNP-effector gene linkages to infer effector genes for about fifteen thousand human GWAS SNPs. This application identified effector genes for 587 GWAS SNPs, linkages of 85% were supported by promoter capture Hi-C data of human islets. Notably, a limitation of pcHi-C data as the gold standard is the lack of specificity compared to, for example, large-scale CRISPR screening experiments. However, it currently serves as a widely used approach for identifying putative links [76–78]. The effector gene set included genes with well-established connections to islet functions (e.g., *ABCC8*, *KCNJ11*, *PDX1*, *ADCY5*, and *KCNQ1*) as well as novel candidates (e.g., *NFATC2IP*). While the ability to infer susceptibility genes for only 3.5% of the GWAS SNPs might appear low, this is due to several potential limiting factors. First, by utilizing multi-omics data from islets, we are aiming to identify effector genes of diabetes-associated GWAS variants in islets. This will inherently exclude SNPs that might be exerting their effects in other tissues. Second, the set of candidate regulatory regions (local-ATAC-MVs) that we have defined in founder strain islets excludes other known potential regulatory mechanisms (e.g., alternative transcriptional regulation and 3D interactions [79, 80]) that the non-coding SNPs might be involved in. Third, only a subset of the trait-associated human GWAS SNPs are likely to be eQTLs [81], and, furthermore, GWAS SNPs can mediate their effects through molecular mechanisms beyond expression modulation. These, in combination with potential organism-specific regulatory mechanisms, impact the extent of effector gene inference from human GWAS SNPs and fine-mapped model organism eQTL data. Despite these shortcomings, we showed with promoter capture Hi-C data validation that INFIMA, with the current lift-over strategies that we employed, can be a powerful transfer learning approach for exploring susceptibility genes of human GWAS loci. The lift-over strategies to identify syntenic non-coding regions between human and mouse are likely to benefit from recent analysis of cross-species enhancers [82].

Conclusions

Model organism studies provide extensive resources for human GWAS; however, effective model organism data integration methods as well as reliable cross organism transfer learning frameworks are lagging behind. INFIMA provides a general framework for fine-mapping model organism molecular quantitative loci by integrating multiple functional data modalities. The availability of such fine-mapping results enables their transfer to the human genome to identify putative effector genes of GWAS variants. The current implementation of INFIMA excludes trans-eQTLs. As the ability to measure inter-chromosomal interactions matures, incorporating trans-eQTLs into INFIMA framework would be a natural extension. The INFIMA software is released at GitHub under the MIT license [83], https://github.com/keleslab/INFIMA. The web application for INFIMA results are available at http://www.statlab.wisc.edu/shiny/INFIMA/.

Methods

ATAC-seq sample preparation

The ATAC-seq samples were prepared using a selection of 50 average sized mouse islets. The islets were washed with 500 μL of PBS at 4C and pelleted by centrifugation at 100×*g* for 1 min. Three hundred microliters of ATAC Lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630) was used to resuspend the islets. The islets were incubated for 20 minutes on ice. After incubating, the islets were lysed by trituration with a 25-gauge needle until intact islets were no longer visible, usually 6 triturations. The lysate was centrifuged at 500×*g* for 10 min at 4C. This generated a crude nuclei pellet and a supernatant. The supernatant was discarded and the nuclei pellet was washed with 100 μL of ATAC Lysis buffer in order to reduce cytoplasmic and mitochondrial contamination. This mixture was centrifuged at 500×*g* for 10 min at 4C and the supernatant was removed. Per ATAC-seq sample, a mixture of 25 μL 2x TDE buffer, 22.5 μL nuclease-free water, and 2.5 μL TDE1 transposase enzyme (Nextera DNA Library Prep kit, Illumina) was applied and incubated for 30 min in a 37 °C water bath. The samples were then purified using a MinElute Reaction cleanup kit (Qiagen) and eluted using two sequential aliquots of 10 μL EB buffer. After purification all ATAC samples were kept at −80 °C. All ATAC-seq samples were transposed and frozen prior to preparing all libraries. Libraries were amplified using 20 μL of ATAC sample, 2.5 μL Primer-1 (Ad1_noMX, 25 μM working stock), 2.5 μL Primer-2 (Ad2.X, 25 μM working stock), and 25 μL of NEBNext High Fidelity 2x PCR Master Mix. Each ATAC sample was amplified by 12 cycles which was determined by qPCR to be saturating for the libraries. The PCR thermocycler was set to 72 °C for 5 min, 98 °C for 30 s, and then 12 total cycles of 98 °C for 10 s, 63 °C for 30 s, 72 °C for 1 min. After amplification the libraries were purified using MinElute PCR purification cleanup kit (Qiagen). The libraries were sequenced to a depth of 134.8 ± 8.2 million reads using paired-end 125 bp reads on a HiSeq2500 (Illumina) at the University of Wisconsin Biotechnology Center DNA Sequencing Facility.

ATAC-seq data analysis

Alignment of ATAC-seq reads

Illumina Nextera adapters were trimmed with cutadapt (version 2.0) [84] using the option "-q 30 –minimum-length 36". Paired-end ATAC-seq reads were aligned to the mouse genome assembly (mm10) with bowtie2 (version 2.3.4.1) [85] with option "-X 800 –no-mixed –no-discordant". For each sample, unmapped reads were filtered out by SAMtools view (version 1.8) [86] with option "-F 4" and mitochondrial reads were removed. Duplicated reads were removed with Picard tools (version 2.9.2) [87]. This resulted in an average of 77.7 ± 4.1 million reads per sample. TSS enrichment analysis was performed with ataqv [88].

Generation of a master peak list from the ATAC-seq samples

Peaks from individual and pooled samples across sexes of each strain were identified using MOSAiCS [27, 28] at FDR of 0.05. Blacklisted regions (see URLs) and Chr Y regions were filtered. We employed IDR analysis [29] to obtain reproducible sets of peaks between male and female samples at IDR of 0.05 and leveraged "SignalValue" and "p-value" outputs from IDR analysis as measures of peak-level signal to noise. The "SignalValue" output was normalized across strains by multiplying $10^8/(\#\text{ of reads})$ to adjust for differences in the sample sequencing depths. IDR identified peaks from the pooled peak sets were trimmed to exclude peaks with the lowest 10% "SignalValue" for each strain and then merged to form the master peak list across all strains. "SignalValue" and "$-\log_{10}(p\text{-value})$" columns were aggregated as "MeanSignal" and "MeanP" in the master peak list.

Strain-specific ATAC-seq peaks tended to have lower ATAC-seq signals compared to peaks present in multiple strains (Additional file 1: Figure S4). We mitigated the potential for this bias by trimming the combined peak list to maximize the overlap of the trimmed set with the ENCODE chromHMM annotations depicting non-quiescent regions of the genome (See URLs; Additional file 1: Figs. S4, S5, and S6; Additional file 1: Supplementary Notes). We reasoned that ATAC-seq peaks across the strains should largely be within non-quiescent chromatin states. We utilized 15-state chromHMM data for mm10 across 12 tissues from the ENCODE portal [89] and annotated the master peak list according to the pooled set of the non-quiescent chromHMM regions across the 12 tissues. For each level of "Total", i.e., the number of strains a master peak is identified in, we varied two tuning parameters: percentile of "MeanSignal" and percentile of "MeanP", both of which varied in {0, 1, ..., 50}. Additional file 1: Figure S5 depicts the heatmaps for the percentage of non-quiescent peaks and the percentage of remaining peaks as a function of these two trimming parameters. In order to maximize these two quantities, we chose tuning parameters for each level of the "Total" and generated the trimmed master peak list. Finally, the reference strain B6 did not have more strain-specific peaks compared to other strains regardless of the trimming procedure, further demonstrating that alignments to the reference mouse genome did not amplify B6 ATAC-seq peak signals (Additional file 1: Table S3, Additional file 1: Figure S7).

Differential accessibility analysis

The ATAC-seq count matrix for the set of master peaks was computed by the R package ChromVAR [90]. We used DESeq2 [91] to identify strain effects (the model "∼ strain" vs. the null model) and sex effect (the model "∼ sex + strain" vs. "∼ strain") by corresponding likelihood ratio tests at FDR of 0.05.

Footprint analysis of ATAC-seq peaks

We utilized PIQ [33] to identify footprints of the 1316 curated JASPAR motifs [92] in B6 ATAC-seq samples with purity score cutoff 0.75, i.e., TF occupancy probability. To investigate whether ATAC peaks were enriched for footprints of TFs highly expressed in islets, we first quantified the ATAC-seq signal genome-wide at base pair resolution by counting the 5′ end Tn5 cut sites for each strain and normalized the cut sites by the sequencing depths. Then, for each potential transcription factor binding site along the genome, we computed the average Tn5 cuts at (1) the binding site, (2) 25 bp flanking regions of the binding site, and (3) 26–50 bp flanking regions of the binding site. We adapted the footprint depth (FPD) metric [48] as the proportional decrease in cut sites at the binding site compared to flanking regions (Additional file 1: Figure S21). The footprint profiles for the individual binding sites were computed from the base pair level ATAC-seq signal in B6 ATAC-seq samples and aggregated for each individual motif. We evaluated the significance of average FPD of each islet TF by comparing it to average FPDs of motifs that are similar in width (width within ± 1 of the islet TF motif width) and information content (information content within ± 0.2 of the islet TF motif information content). A randomization test was performed to evaluate the collective enrichment of islet TFs (Additional file 1: Supplementary Notes).

Identification of local-ATAC-MVs

In order to evaluate the impact of SNPs on ATAC-seq signal, we first extracted genetic variants within differential ATAC-seq peaks for the eight founder strains from the dbSNP (v142) database (see URLs, [93]) with the R package VariantAnnotation (version 1.34.0) [94]. Retaining only the SNPs with "FILTER = PASS" and "QUAL = 999" resulted in 630,349 SNPs. In order to identify genetic variants genotypes of which are associated with the ATAC-seq signal, we conducted a permutation test and retained for each differential ATAC-seq peak only the SNP which associated the best with the local-ATAC-seq signal while including all the SNPs with the same exact best association statistics. This resulted in 22,200 ATAC-seq peaks harboring a total of 47,062 local-ATAC-MVs at FDR of 0.05, with an average (median) of 2.1 (1.0) local-ATAC-MVs per peak.

In silico mutation and footprint analysis

Variant-level comparative footprint analysis

We applied atSNP [35] to 47,062 local-ATAC-MVs with the 1316 curated JASPAR motifs [92] and quantified the in silico effect of SNPs on TF binding by labeling SNP-motif combinations with atSNP `pval_rank` < 0.05 as significant.

Next, to quantify the impact of SNPs on the realized ATAC-seq footprints, for each SNP $\times$ motif interaction, FPD with/without SNP were computed by aggregating the results for strains with/without the alternative allele. This ensured the disruption/enhancement of motif by a SNP to be consistent with a decrease/increase in FPD. In order to evaluate whether the change in FPD (ΔFPD) due to the SNP is significant, we generated motif-specific empirical null distributions of ΔFPD by treating insignificant results from atSNP as the null set since this approximated the distribution of ΔFPD when the SNP is not affecting the motif. Only the SNP-motif combinations with `pval_fpd` < 0.05 were retained for the downstream analysis.

Accounting for both the in silico effect of SNP on TF binding and change in ATAC-seq FPD, resulted in 1,211,807 candidate SNP-motif interactions with consistent changes across the two metrics (640,038 Gain of function combinations: `pval_ref` > 0.05, `pval_snp` ≤ 0.05, ΔFPD > 0; 571,769 Loss of function combinations: `pval_ref` ≤ 0.05, `pval_snp` > 0.05, ΔFPD < 0). Finally, for each SNP, we recorded the minimum `pval_fpd` as the p-value for the null hypothesis that the SNP is not affecting any TF binding. Collectively, we identified 8029 significant SNP $\times$ motif interactions comprising 1350 SNPs and 1196 motifs (FDR of 0.05).

RNA-seq sample preparation

Islet RNA profiling methods are described in detail in [14].

RNA-seq data analysis

Quantification of transcript abundance

We used RSEM [49] with GENCODE vm18 [95] gene annotation and obtained the gene expression count matrix across protein-coding genes on Chromosomes 1-19, and X. Genes with the lowest 10% variance across the samples were removed from the downstream analysis. Upper quartile normalization [96] and retaining the genes with non-zero counts in at least 85% of the samples resulted in 13,568 protein-coding genes.

Association analysis of founder local-ATAC-MVs and gene expression

We applied MatrixEQTL [97] with default settings to all local-ATAC-MVs and obtained 96,309 associated local-ATAC-MV and gene pairs (34,711 distinct local-ATAC-MVs, only *cis* regulatory local-ATAC-MVs were considered, 1 Mb window) at FDR of 1e−5.

INFIMA implementation details

INFIMA model fitting with an Expectation-Maximization algorithm

We estimated the INFIMA parameters with maximum likelihood using an expectation-maximization (EM) algorithm. We provide below the detailed derivations. Let $\Gamma_{\mathbf{g}} = (\Theta_{\mathbf{g}}, \mathbf{a_0}, \mathbf{b_0}, \mathbf{a_1}, \mathbf{b_1}, \gamma)$ denote the full set of model parameters and $\mathbf{1_{g,k}}$ be a $p_g \times 1$ vector with the kth entry equal to 1 and 0 elsewhere. The joint likelihood of the data $(\widetilde{\mathbf{Y}}_{\mathbf{g}})$

and the latent variables, conditional on features $\mathbf{X_g}$ extracted from founder RNA-seq and ATAC-seq, for $\mathbf{Z_g} = \mathbf{1_{g,k}}$ is given by

$$\mathbb{P}\left(\widetilde{\mathbf{Y}}_\mathbf{g}, \mathbf{Z_g} = \mathbf{1_{g,k}} | \mathbf{X_g}, \Gamma_\mathbf{g}\right) \propto \mathbb{P}\left(\widetilde{\mathbf{Y}}_\mathbf{g} | \mathbf{X_g}, \mathbf{Z_g} = \mathbf{1_{g,k}}, \Gamma_\mathbf{g}\right) \mathbb{P}\left(\mathbf{Z_g} = \mathbf{1_{g,k}} | \mathbf{X_g}, \Gamma_\mathbf{g}\right), \tag{4}$$

where the first term is given by

$$\begin{aligned} \mathbb{P}\left(\widetilde{\mathbf{Y}}_\mathbf{g} | \mathbf{X_g}, \mathbf{Z_g} = \mathbf{1_{g,k}}, \Gamma_\mathbf{g}\right) &= \mathbb{P}\left(\widetilde{\mathbf{Y}}_\mathbf{g} | \mathbf{X_g}, \mathbf{Z_g} = \mathbf{1_{g,k}}, \mathbf{a_1}, \mathbf{b_1}\right) & (5) \\ &= \prod_{i=0,1,2} a_{1,i}^{n_{g,i,k}} b_{1,i}^{m_{g,i,k}}, & (6) \end{aligned}$$

and the second term is given by

$$\begin{aligned} \mathbb{P}\left(\mathbf{Z_g} = \mathbf{1_{g,k}} | \mathbf{X_g}, \Gamma_\mathbf{g}\right) &= \mathbb{P}\left(\mathbf{Z_g} = \mathbf{1_{g,k}} | \mathbf{X_g}, \Gamma_\mathbf{g}, V_g = 1\right) \mathbb{P}\left(V_g = 1 | \mathbf{X_g}, \Gamma_\mathbf{g}\right) & (7) \\ &= \theta_{g,k}\gamma. & (8) \end{aligned}$$

Similarly, the joint likelihood when $\mathbf{Z_g} = \mathbf{0}$ is then

$$\begin{aligned} \mathbb{P}\left(\widetilde{\mathbf{Y}}_\mathbf{g}, \mathbf{Z_g} = \mathbf{0} | \mathbf{X_g}, \Gamma_\mathbf{g}\right) &\propto \mathbb{P}\left(\widetilde{\mathbf{Y}}_\mathbf{g} | \mathbf{X_g}, \mathbf{Z_g} = \mathbf{0}, \Gamma_\mathbf{g}\right) \mathbb{P}\left(\mathbf{Z_g} = \mathbf{0} | \mathbf{X_g}, \Gamma_\mathbf{g}\right) & (9) \\ &= \prod_{i=0,1,2} a_{0,i}^{n^0_{g,i}} b_{0,i}^{m^0_{g,i}} (1 - \gamma). & (10) \end{aligned}$$

We next derive the full parameter joint posterior distribution given the latent variables $\mathbf{Z_g}, V_g$ as

$$\mathbb{P}\left(\Gamma_\mathbf{g}; \widetilde{\mathbf{Y}}_\mathbf{g} | \mathbf{X_g}, \mathbf{Z_g}, V_g\right) \propto \mathbb{P}\left(\widetilde{\mathbf{Y}}_\mathbf{g} | \mathbf{X_g}, \mathbf{Z_g}, V_g, \Gamma_\mathbf{g}\right) \mathbb{P}\left(\Gamma_\mathbf{g} | \mathbf{X_g}, \mathbf{Z_g}, V_g\right), \tag{11}$$

where

$$\begin{aligned} \mathbb{P}\left(\widetilde{\mathbf{Y}}_\mathbf{g} | \mathbf{X_g}, \mathbf{Z_g}, V_g, \Gamma_\mathbf{g}\right) &= \underbrace{\mathbb{P}\left(\mathbf{a_0}, \mathbf{b_0}, \mathbf{a_1}, \mathbf{b_1}; \widetilde{\mathbf{Y}}_\mathbf{g} | \mathbf{X_g}, \mathbf{Z_g}, V_g\right)}_{L_{g,1}} & (12) \\ \mathbb{P}\left(\Gamma_\mathbf{g} | \mathbf{X_g}, \mathbf{Z_g}, V_g\right) &= \underbrace{\mathbb{P}\left(\Theta_\mathbf{g} | \mathbf{X_g}, \mathbf{Z_g}, V_g\right)}_{L_{g,2}} \underbrace{\mathbb{P}\left(\gamma | V_g\right)}_{L_{g,3}}. & (13) \end{aligned}$$

With the combined generative model, we have

$$\begin{aligned} L_{g,1} &= \left[f_{\mathbf{a_1},\mathbf{b_1}}\left(\mathbf{n_g}, \mathbf{m_g}\right)\right]^{\mathbb{I}(\mathbf{Z_g} \neq \mathbf{0})} \left[f_{\mathbf{a_0},\mathbf{b_0}}\left(\mathbf{n^0_g}, \mathbf{m^0_g}\right)\right]^{\mathbb{I}(\mathbf{Z_g} = \mathbf{0})} & (14) \\ &= \left[f_{\mathbf{a_1},\mathbf{b_1}}\left(\mathbf{n_g}, \mathbf{m_g}\right)\right]^{V_g} \left[f_{\mathbf{a_0},\mathbf{b_0}}\left(\mathbf{n^0_g}, \mathbf{m^0_g}\right)\right]^{1-V_g}, & (15) \end{aligned}$$

where$f_{\cdot,\cdot}$ denotes the product of Multinomial probability mass functions with appropriate parameters. The log likelihood aggregated over $g \in \{1, 2, \ldots, G\}$ is given by

$$\begin{aligned} \log(L_1) &= \sum_{g=1}^{G} \log(L_{g,1}) & (16) \\ &= \sum_{g=1}^{G} \left\{ V_g \sum_{k=1}^{p_g} Z_{g,k} \sum_{i=0,1,2} \left(n_{g,i,k} \log a_{1,i} + m_{g,i,k} \log b_{1,i}\right) + \right. & (17) \\ &\quad \left. (1 - V_g) \sum_{i=0,1,2} \left(n^0_{g,i} \log a_{0,i} + m^0_{g,i} \log b_{0,i}\right) \right\}. & (18) \end{aligned}$$

We define the weighted sums of the edit distance random variables $\mathbf{n_g}, \mathbf{n_g^0}, \mathbf{m_g}, \mathbf{m_g^0}$ as

$$N_i = \sum_{g=1}^{G} V_g \sum_{k=1}^{p_g} Z_{g,k} n_{g,i,k}, \quad N = N_0 + N_1 + N_2, \tag{19}$$

$$N_i^0 = \sum_{g=1}^{G} (1 - V_g) n_{g,i}^0, \quad N^0 = N_0^0 + N_1^0 + N_2^0, \tag{20}$$

$$M_i = \sum_{g=1}^{G} V_g \sum_{k=1}^{p_g} Z_{g,k} m_{g,i,k}, \quad M = M_0 + M_1 + M_2, \tag{21}$$

$$M_i^0 = \sum_{g=1}^{G} (1 - V_g) m_{g,i}^0, \quad M^0 = M_0^0 + M_1^0 + M_2^0. \tag{22}$$

Then, the Maximum Likelihood Estimators (MLEs) of the parameters are given by:

$$\widehat{\mathbf{a_1}} = \frac{(N_0, N_1, N_2)^\top}{N}, \quad \widehat{\mathbf{a_0}} = \frac{\left(N_0^0, N_1^0, N_2^0\right)^\top}{N^0}, \tag{23}$$

$$\widehat{\mathbf{b_1}} = \frac{(M_0, M_1, M_2)^\top}{M}, \quad \widehat{\mathbf{b_0}} = \frac{\left(M_0^0, M_1^0, M_2^0\right)^\top}{M^0}. \tag{24}$$

We note that $L_{g,2}$ is the posterior distribution of $\Theta_\mathbf{g}$. By the Dirichlet-Multinomial conjugacy, we have

$$\Theta_\mathbf{g} | \mathbf{X_g}, \mathbf{Z_g}, V_g \sim \text{Dirichlet}\left(\Pi_\mathbf{g} + \mathbf{Z_g} V_g\right), \tag{25}$$

and the maximum a posteriori (MAP) estimator can be computed as

$$\widehat{\Theta}_{g,k} = \frac{\Pi_{g,k} + Z_{g,k} V_g - 1}{\sum_{k=1}^{p_g} \left(\Pi_{g,k} + Z_{g,k} V_g\right) - p_g}. \tag{26}$$

Maximizing $L_{g,3} = \mathbb{P}(\gamma | V_g)$ with respect to the prior probability γ that an association is driven by causal SNP, we get $\widehat{\gamma} = \frac{1}{G} \sum_{g=1}^{G} V_g$.

In the DO-eQTL application, the INFIMA model was fit with the EM algorithm described in Algorithm 1, where V_g and $Z_{g,k}$ values in the above equations were imputed in the E-step. Multiple initial values of parameters were employed to avoid local optima.

Trinarization of allelic expression effect sizes into allelic patterns

For the DO-eQTL data $\mathbf{Y_g}$, we first standardized the 8 ×1 vector to [0,1] and subtracted the allelic expression effect of the reference strain B6. We then trinarized the entries with values > 0.2, < -0.2 to $1, -1$ respectively, and set other entries to 0 to obtain $\widetilde{\mathbf{Y}}_\mathbf{g}$. The cutoffs were selected by balancing the number of entries with the 3 values. The same trinarization scheme was applied to the normalized founder gene expression vector $\mathbf{B_g} \rightarrow \widetilde{\mathbf{B}}_\mathbf{g}$ as well. For each row of the founder RNA-seq genotype effect matrix $\mathbf{E_g}$, if the effect size from the marginal regression of gene expression on the genotype was significant at level 0.05, we replaced the effect size with 1 or – 1 depending on the sign of the effect size; otherwise, the effect size was replaced by 0. Therefore, we obtained $\widetilde{\mathbf{E}}_\mathbf{g}$ and $\widetilde{\mathbf{R}}_\mathbf{g} = \widetilde{\mathbf{E}}_\mathbf{g}^\top \mathbf{Z_g}$. Figure 5f illustrates a specific example in detail.

Algorithm 1 INFIMA Model Fitting with Expectation-Maximization

```
procedure INFIMA(DO-eQTL, ATAC-seq, RNA-seq)
    Initialize $\mathbf{a_1}$, $\mathbf{a_0}$, $\mathbf{b_1}$, $\mathbf{b_0}$, $\Theta$, and $\gamma$.
    repeat
        □ E-step:
        for $g \in \{1, 2, \ldots, G\}$ do
            for $k \in \{1, 2, \ldots, p_g\}$ do
                $Z_{g,k}^{(t)} = \prod_{i=0,1,2} ({a_{1,i}^{(t)}}^{n_{g,i,k}} {b_{1,i}^{(t)}}^{m_{g,i,k}}) \gamma^{(t)} \theta_{g,k}^{(t)}$
            end for
            ${\mathbf{Z_g}}^{(t)} = (Z_{g,1}^{(t)}, \ldots, Z_{g,p_g}^{(t)})^\top / (\sum_{k=1}^{p_g} Z_{g,k}^{(t)} + \prod_{i=0,1,2} ({a_{0,i}^{(t)}}^{n_{g,i,k}^0} {b_{0,i}^{(t)}}^{m_{g,i,k}^0})(1-\gamma^{(t)}))$
            $V_g^{(t)} = \mathbf{1}^\top {\mathbf{Z_g}}^{(t)}$
        end for
        □ M-step:
        Update ${\mathbf{a_1}}^{(t+1)}$, ${\mathbf{a_0}}^{(t+1)}$, ${\mathbf{b_1}}^{(t+1)}$, ${\mathbf{b_0}}^{(t+1)}$ according to Eqs. 23 and 24.
        for $g \in \{1, 2, \ldots, G\}$ do
            for $k \in \{1, 2, \ldots, p_g\}$ do
                $\theta_{g,k}^{(t+1)} = Z_{g,k}^{(t)} V_g^{(t)} + \Pi_{g,k} - 1$
            end for
            ${\Theta_\mathbf{g}}^{(t+1)} = {\Theta_\mathbf{g}}^{(t+1)} / \mathbf{1}^\top {\Theta_\mathbf{g}}^{(t+1)}$ (Eq. 26).
        end for
        $\gamma^{(t+1)} = \frac{1}{G} \sum_{g=1}^{G} \mathbf{1}^\top {\mathbf{Z_g}}^{(t)}$
        $t = t + 1$
    until $t \geq$ max_iteration or $\Delta change \leq$ threshold.
end procedure
```

Distance prior

A well-known bias of Hi-C data is that Hi-C signal decreases exponentially as the distance between promoters and enhancers increases [98]. In order to avoid the bias towards the local-ATAC-MVs closest to the gene promoter, we chose not to penalize the distance until 250 Kb. When distance is above 250 Kb, the score function has a decreasing trend in order to slightly favor closer local-ATAC-MVs. We set the window size W equal to 1 Mb and defined the distance score function as $D(x) = 0.5$ if $x \leq 0.25$ Mb; $D(x) = \frac{5}{12} / \left(10x - \frac{5}{3}\right)$ if $x > 0.25$ Mb, where x is the distance between local-ATAC-MV and DO gene promoter. As a component of the prior $\Pi_\mathbf{g}$, the maximum value of distance score $\mathbf{D_g}$ is 0.5, which serves as a "tie-breaker" rather than overwhelming the other three components (Fig. 5a).

Pseudocounts for the edit distance random variables

To promote the consistency between the trinarized DO-eQTL data and founder data, i.e., to tilt the edit distance random variables to favor lower values, we utilized pseudocount parameters $\lambda_0 = 0.1, \lambda_1 = 0.01$, and $\lambda_2 = 0$ for the multinomial edit distance random variables. Specifically, pseudocounts $\lambda_i p_g$ were added to the weighted sums of edit distance variables (Eqs. 19 to 22) in estimation of $\mathbf{a_0}$, $\mathbf{b_0}$, $\mathbf{a_1}$, and $\mathbf{b_1}$ as:

$$
\begin{aligned}
N_i &= \sum_{g=1}^{G} V_g \sum_{k=1}^{p_g} Z_{g,k} n_{g,i,k} + \lambda_i p_g, && (27)\\
N_i^0 &= \sum_{g=1}^{G} (1 - V_g) n_{g,i}^0 + \lambda_i p_g, && (28)\\
M_i &= \sum_{g=1}^{G} V_g \sum_{k=1}^{p_g} Z_{g,k} m_{g,i,k} + \lambda_i p_g, && (29)\\
M_i^0 &= \sum_{g=1}^{G} (1 - V_g) m_{g,i}^0 + \lambda_i p_g. && (30)
\end{aligned}
$$

Under $\lambda_0 >> \lambda_1 >> \lambda_2$, INFIMA formulation promotes the resulting causal SNPs to have consistent relationships between the founder data and the DO-eQTL data; therefore, the ordering of the SNPs is relatively insensitive to the actual values of these pseudocount parameters.

Data-driven simulations

In order to simulate realistic data for our evaluations, we leveraged the parameters estimated by the INFIMA on the DO-eQTL data fit with all the summarized data from ATAC-seq, local-ATAC-MVs, and RNA-seq data. We used these parameter values as well as the actual summarized ATAC-seq, local-ATAC-MVs, and RNA-seq to simulate V_g, $\mathbf{Z_g}$, and $\mathbf{Y_g}$ from the plate model in Fig. 5b. We varied the informativeness level of the summarized data by varying the prior parameter $\Pi_{g,k} := F_{g,k} + D_{g,k} + |\text{cor}(\mathbf{A_{g,k}}, \mathbf{E_{g,k}})| + |\text{cor}(\mathbf{A_{g,k}}, \mathbf{B_g})| + 1$ according to the following three settings:

NI: The prior parameter $\Pi_{g,k}$ is set to be 1 for all candidate SNPs, corresponding to an uninformative prior.

MI: The prior parameter $\Pi_{g,k}$ set to its observed value in the actual data and accommodates multiple SNPs with $F_{g,k} = 1$. Multiple SNPs are affecting footprints under this setting. Causal SNPs are distinguished by other components of the prior parameter.

HI: $F_{g,k}$ is set to 10 for a randomly selected SNP k and 0 for other SNPs. Under this setting, the SNPs that affect footprints are more likely to be chosen as causal due to the dominant contribution of the footprint component.

Statistical power for V_g was calculated at FDR of 0.05 by using a direct posterior probability approach [99].

Linking human GWAS SNPs to mouse islet ATAC-seq peaks

The peak-based lift-over consisted of two steps: (1) direct and (2) indirect (Additional file 1: Figure S36b). After removing blacklisted and chr Y human ATAC-seq peaks [60], we obtained 156,861 human islet ATAC-seq peaks. For indirect mapping, we used “nearest()” function in “GenomicAlignments” R package [100] to link GWAS SNPs to their nearest human ATAC-seq peaks within 10 Kb distance. We used ‘liftOver()’ function in “rtracklayer” R package [101] and the hg19 to mm10 reciprocal chain file (see URLs). For each human genomic region, we merged gaps less than 10 bp among its mapped regions in the mouse genome and selected the one with maximum width as the syntenic region.

We then linked these syntenic regions to their nearest mouse ATAC-seq peaks within 10 Kb distance. The distance constraints aided to remove potential false positives to preserve conservation of genomic compartments between the syntenic regions of the two organisms. We observed a decline in level of conservation without imposing the distance constraints (Additional file 1: Figure S50).

Enrichment analysis of human GWAS SNPs associated with islet function traits

We carried out an enrichment analysis for the associated SNPs of islet function-related GWAS traits with more than 40 SNPs. Enrichment *p*-values were calculated based on a resampling based null distribution that matched the phylogenic conservation score, width, and chromosomal distribution of the syntenic regions of each GWAS trait. Specifically, for each trait, we sampled the same number of random syntenic regions as the size of the set lifted-over to mouse genome by matching the phylogenic conservation score, width, and chromosomal distribution of the sampled regions to those of the actual syntenic regions. The random syntenic regions were mapped to mouse ATAC peaks within 10 Kb distance, and the overlap with the local-ATAC-MVs were recorded. Repeating this procedure one million times generated a null distribution for the actual observed number of local-ATAC-MVs that mapped to GWAS. The resulting enrichment *p*-values were corrected for multiple testing with the Bonferroni procedure at the significance level of 0.05.

Validation of INFIMA predicted SNP-effector gene linkages with promoter capture Hi-C

For validation purposes, we filtered out 8 out of 1540 mouse ATAC-seq peaks because the human ortholog of the genes that they were linked to resided in different chromosomes than the corresponding GWAS SNPs that they mapped to. Then, we processed the INFIMA results that fine-mapped 737 local-ATAC-MV containing peaks that corresponded to syntenic regions of human GWAS SNPs. INFIMA resulted in mappings for 587 GWAS SNPs (548 local-ATAC-MV containing peaks) by considering the local-ATAC-MVs with aggregated posterior probability of being causal larger than 0.80 and with a credible set less than 50% of the all the candidate SNPs. We leveraged 175,784 significant promoter capture Hi-C contacts from [66] for validation of the inferred links. With a median bin size $\sim$ 4 Kb, the median interaction distance of the pcHi-C data is $\sim$ 300 Kb. We required one end of pcHi-C interaction to be within 10 Kb upstream and 2 Kb downstream around TSS of human orthologous genes while the other end of pcHi-C to reside within 10 Kb distance of GWAS SNPs and human ATAC-seq peaks. We identified 346 GWAS SNPs that were supported by pcHi-C through at least one effector gene. Furthermore, at least one LD partner ($R^2 > 0.8$, 1000 Genomes Phase 3 v5 European population, SNiPA v3.3 [102]) of the 153 GWAS SNPs were in contact with the inferred effector genes. Comparison of INFIMA predictions to the baseline strategies was carried out with a Fisher's exact test.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02450-8.

Additional file 1: Supplementary Information. Additional file 1 contains Supplementary Notes, Supplementary Figs.S1-S50 and Supplementary Tables S1-S6.

Additional file 2: Review history

Acknowledgements

We are grateful to Keleş Group members for helpful discussions and we thank two anonymous reviewers for their constructive comments that led to an improved version of the manuscript.

Peer review information

Anahita Bishop was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as additional file 2.

Authors' contributions

S.K., M.K., and A.A. conceived the project. Sh.S., D.S., and K.S. performed ATAC-seq assays. G.C. contributed RNA-seq data. C.D. and S.K. analyzed ATAC-seq and RNA-seq datasets and performed footprint analysis. Su.S. and S.K. optimized the ATAC-seq data analysis pipeline. C.D. and S.K. developed and evaluated the INFIMA model. L.L., X.L., F.J., and Y.L. contributed easy Hi-C data. C.D. and S.K. developed the initial version of the manuscript. C.D., S.K., and M.K. wrote the manuscript with input from other authors. The authors read and approved the final manuscript.

Funding

This work was supported by National Institutes of Health grants DK101573, HG007019, HG003747, HG011371, and U54AI117924 for the Center for Predictive Computational Phenotyping (CPCP) (SK).

Availability of data and materials

The founder mice ATAC-seq, founder mice RNA-seq, and B6 mice Hi-C datasets generated in this publication have been deposited in NCBI's Gene Expression Omnibus [103] and are accessible through GEO Series accession number GSE180810 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE180810). All the used third party data that support the findings of this study are available at the following resources: The DO mice eQTL data [14], https://churchilllab.jax.org/qtlviewer/attie/islets. ENCODE 15-state chromHMM data [89], https://www.encodeproject.org/search/?type=Annotation&annotation_type=chromatin+state&assembly=mm10&files.file_type=bed+bed9. ENCODE H3K27ac and H3K4me3 ChIP-seq based classification of tissue-specific promoters/enhancers [89], http://zlab-annotations.umassmed.edu/enhancers/ and http://zlab-annotations.umassmed.edu/promoters/. Blacklisted regions [104, 105], http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm10-mouse/. dbSNP142 [93], http://ftp://ftp-mouse.sanger.ac.uk/current_snps/mgp.v5.merged.snps_all.dbSNP142.vcf.gz. JASPAR motifs [92], http://jaspar.genereg.net/. GENCODE vm18 [95], https://www.gencodegenes.org/mouse/release_M18.html. 16 diabetes-related physiological traits from GWAS central [59], https://www.gwascentral.org/. Human islet ATAC-seq peaks [60]. The reciprocal chain file [106–109], https://hgdownload-test.gi.ucsc.edu/goldenPath/hg19/vsMm10/reciprocalBest/. Human islet promoter capture Hi-C data [66]. Human GWAS atSNP results [69], http://atsnp.biostat.wisc.edu/search. The processed data and results for this study are deposited at Zenodo [110], https://doi.org/10.5281/zenodo.4679897. The source code for reproducing the key results is released at GitHub under the GPL 3.0 license [111], https://doi.org/10.5281/zenodo.5099585. The INFIMA software is released at GitHub under the MIT license [83], https://github.com/keleslab/INFIMA.

Declarations

Ethics approval and consent to participate

The animal protocol number is A005821-R01-A02. The IRB approval number is 2017-0793. The euthanization method of the mice is decapitation.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA. [2]Department of Biochemistry, University of Wisconsin-Madison, Madison, WI, USA. [3]Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX, USA. [4]The Jackson Laboratory, Bar Harbor, ME, USA. [5]Case Western University, Cleveland, OH, USA. [6]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA.

Received: 14 October 2020 Accepted: 2 August 2021
Published online: 23 August 2021

References

1. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42(D1):1001–6.
2. Nicolae D, Gamazon E, Zhang W, Duan S, Dolan M, Cox N. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6(4):1000888. https://doi.org/10.1371/journal.pgen.1000888.
3. Dimas A, Deutsch S, Stranger B, Montgomery S, Borel C, Attar-Cohen H, Ingle C, Beazley C, Arcelus M, Sekowska M, Gagnebin M, Nisbett J, Deloukas P, Dermitzakis E, Antonarakis S. Common regulatory variation impacts gene expression in a cell type–dependent manner. Science. 2009;325(5945):1246–50.

4. Mahajan A, Taliun D, Thurner M, Robertson N, Torres J, Rayner N, Payne A, Steinthorsdottir V, Scott R, Grarup N, Cook J, Schmidt E, Wuttke M, Sarnowski C, Mägi R, Nano J, Gieger C, Trompet S, Lecoeur C, Preuss M, Prins B, Guo X, Bielak L, Below J, Bowden D, Chambers J, Kim Y, Ng M, Petty L, Sim X, Zhang W, Bennett A, Bork-Jensen J, Brummett C, Canouil M, Ec kardt K, Fischer K, Kardia S, Kronenberg F, Läll K, Liu C, Locke A, Luan J, Ntalla I, Nylander V, Schönherr S, Schurmann C, Yengo L, Bottinger E, Brandslund I, Christensen C, Dedoussis G, Florez J, Ford I, Franco O, Frayling T, Giedraitis V, Hackinger S, Hattersley A, Herder C, Ikram M, Ingelsson M, Jørgensen M, Jørgensen T, Kriebel J, Kuusisto J, Ligthart S, Lindgren C, Linneberg A, Lyssenko V, Mamakou V, Meitinger T, Mohlke K, Morris A, Nadkarni G, Pankow J, Peters A, Sattar N, Stančáková A, Strauch K, Taylor K, Thorand B, Thorleifsson G, Thorsteinsdottir U, Tuomilehto J, Witte D, Dupuis J, Peyser P, Zeggini E, Loos R, Froguel P, Ingelsson E, Lind L, Groop L, Laakso M, Collins F, Jukema J, Palmer C, H.Grallert, Metspalu A, Dehghan A, Köttgen A, Abecasis G, Meigs J, Rotter J, Marchini J, Pedersen O, Hansen T, Langenberg C, Wareham N, Stefansson K, Gloyn A, Morris A, Boehnke M, McCarthy M. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet. 2018;50(11):1505–13. https://doi.org/10.1038/s41588-018-0241-6.
5. Smemo S, Tena J, Kim K-H, Gamazon E, Sakabe N, Gómez-Marín C, Aneas I, Credidio F, Sobreira D, Wasserman N, et al. Obesity-associated variants within fto form long-range functional connections with irx3. Nature. 2014;507(7492):371–5.
6. Claussnitzer M, Dankel S, Kim K-H, Quon G, Meuleman W, Haugen C, Glunk V, Sousa I, Beaudry J, Puviindran V, et al. Fto obesity variant circuitry and adipocyte browning in humans. N Engl J Med. 2015;373(10):895–907.
7. Gallagher M, Chen-Plotkin A. The post-GWAS era: from association to function. Am J Hum Genet. 2018;102(5): 717–30.
8. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira A, Knowles D, Golan D, Ermel R, Ruusalepp A, Quertermous T, Hao K, et al. Opportunities and challenges for transcriptome-wide association studies. Nat Genet. 2019;51(4):592–9.
9. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson M, Powell J, Montgomery G, Goddard M, Wray N, Visscher P, et al. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. Nat Genet. 2016;48(5): 481.
10. Cheng Y, Ma Z, Kim B, Wu W, Cayting P, Boyle A, Sundaram V, Xing X, Dogan N, Li J, Euskirchen G, Lin S, Lin Y, Visel A, Kawli T, Yang X, Patacsil D, Keller C, Giardine B, Kundaje A, Wang T, Pennacchio L, Weng Z, Hardison R, Snyder M, Consortium M. Principles of regulatory information conservation between mouse and human. Nature. 2014;515(7527):371–5.
11. Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, Hansen R, Stehling-Sun S, Sabo P, Byron R, Humbert R, Thurman R, Johnson A, Vong S, Lee K, Bates D, Neri F, Diegel M, Giste E, Haugen E, Dunn D, Wilken M, Josefowicz S, Samstein R, Chang K-H, Eichler E, De Bruijn M, Reh T, Skoultchi A, Rudensky A, Orkin S, Papayannopoulou T, Treuting P, Selleri L, Kaul R, Groudine M, Bender M, Stamatoyannopoulos J. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science. 2014;346(6212):1007–12.
12. Hook P, Mccallion A. Leveraging mouse chromatin data for heritability enrichment informs common disease architecture and reveals cortical layer contributions to schizophrenia. Genome Res. 2020;30. https://doi.org/10.1101/gr.256578.119.
13. Churchill G, Gatti D, Munger S, Svenson K. The diversity outbred mouse population. Mamm Genome. 2012;23(9-10):713–8.
14. Keller M, Gatti D, Schueler K, Rabaglia M, Stapleton D, Simecek P, Vincent M, Allen S, Broman RbsuffixAandB, Kendziorski C, Broman K, Yandell B, Churchill G, Attie A. Genetic drivers of pancreatic islet function. Genetics. 2018;209(1):335–56. https://doi.org/10.1534/genetics.118.300864.
15. Shorter J, Huang W, Beak J, Hua K, Gatti D, Villena F, Pomp D, Jensen B. Quantitative trait mapping in diversity outbred mice identifies two genomic regions associated with heart size. Mamm Genome. 2017;29. https://doi.org/10.1007/s00335-017-9730-7.
16. Deasy S, Uehara R, Vodnala S, Yang H, Dass R, Hu Y, Lee M, Crouch R, Hunter K. Aicardi-goutières syndrome gene rnaseh2c is a metastasis susceptibility gene in breast cancer. PLoS Genet. 2019;15:1008020. https://doi.org/10.1371/journal.pgen.1008020.
17. Keenan B, Galante R, Lian J, Simecek P, Gatti D, Zhang L, Lim D, Svenson K, Churchill G, Pack A. High-throughput sleep phenotyping produces robust and heritable traits in diversity outbred mice and their founder strains. Sleep. 2020;43(5):278. https://doi.org/10.1093/sleep/zsz278.
18. Recla J, Bubier J, Gatti D, Ryan J, Long K, Robledo R, Glidden N, Hou G, Churchill G, Maser R, Zhang Z-W, Young E, Chesler E, Bult C. Genetic mapping in diversity outbred mice identifies a Trpa1 variant influencing late-phase formalin response. PAIN. 2019;160(8):1740–53. https://doi.org/10.1097/j.pain.0000000000001571.
19. Keller M, Rabaglia M, Schueler K, Stapleton D, Gatti D, Vincent M, Mitok K, Wang Z, Ishimura T, Simonett S, et al. Gene loci associated with insulin secretion in islets from nondiabetic mice. J Clin Investig. 2019;129(10):4419–32.
20. Nicod J, Davies R, Cai N, Hassett C, Goodstadt L, Cosgrove C, Yee B, Lionikaite V, Mcintyre R, Remme C, Lodder E, Gregory J, Hough T, Joynson R, Phelps H, Nell B, Rowe C, Wood J, Walling A, Flint J. Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. Nat Genet. 2016;48. https://doi.org/10.1038/ng.3595.
21. Broman K, Gatti D, Simecek P, Furlotte N, Prins P, Sen S, Yandell B, Churchill G. R/qtl2: software for mapping quantitative trait loci with high-dimensional data and multiparent populations. Genetics. 2019;211(2):495–502.
22. Kichaev G, Yang W-Y, Lindstrom S, Hormozdiari F, Eskin E, Price A, Kraft P, Pasaniuc B. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. 2014;10(10):1004722. https://doi.org/10.1371/journal.pgen.1004722.
23. Chen W, McDonnell S, Thibodeau S, Tillmans L, Schaid D. Incorporating functional annotations for fine-mapping causal variants in a bayesian framework using summary statistics. Genetics. 2016;204(3):933–58.

24. Buenrostro J, Giresi P, Zaba L, Chang H, Greenleaf W. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. Nat Methods. 2013;10(12):1213.
25. Wang Z, Gerstein M, Snyder M. Rna-seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.
26. Zhang Q, Zeng X, Younkin S, Kawli T, Snyder M, Keleş S. Systematic evaluation of the impact of ChIP-seq read designs on genome coverage, peak identification, and allele-specific binding detection. BMC Bioinforma. 2016;17(1):96.
27. Kuan P, Chung D, Pan G, Thomson J, Stewart R, Keleş S. A statistical framework for the analysis of chip-seq data. J Am Stat Assoc. 2011;106(495):891–903.
28. Sun G, Chung D, Liang K, Keleş S. Statistical analysis of ChIP-seq data with MOSAiCS. In: Deep sequencing data analysis. Totowa: Humana Press; 2013. p. 193–212.
29. Li Q, Brown J, Huang H, Bickel P, et al. Measuring reproducibility of high-throughput experiments. Ann Appl Stat. 2011;5(3):1752–79.
30. Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado M, Malinverni R. regioner: an r/bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics. 2016;32(2):289–91.
31. Yu G, Wang L-G, He Q-Y. Chipseeker: an r/bioconductor package for chip peak annotation, comparison and visualization. Bioinformatics. 2015;31(14):2382–3.
32. Morgan A, Welsh C. Informatics resources for the collaborative cross and related mouse populations. Mamm Genome. 2015;26(9):521–39.
33. Sherwood R, Hashimoto T, O'Donnell C, Lewis S, Barkal A, Hoff J, Karun V, Jaakkola T, Gifford D. Discovery of directional and nondirectional pioneer transcription factors by modeling dnase profile magnitude and shape. Nat Biotechnol. 2014;32. https://doi.org/10.1038/nbt.2798.
34. Zhijian L, Schulz M, Look T, Begemann M, Zenke M, Costa I. Identification of transcription factor binding sites using atac-seq. Genome Biol. 2019;20:. https://doi.org/10.1186/s13059-019-1642-2.
35. Zuo C, Shin S, Keleş S. atSNP: transcription factor binding affinity testing for regulatory SNP detection. Bioinformatics. 2015;31(20):3353–5.
36. Keller M, Paul P, Rabaglia M, Stapleton D, Schueler K, Broman A, Ye S, Leng N, Brandon C, Neto E, Plaisier C, Simonett S, Kebede M, Sheynkman G, Klein M, Baliga N, Smith L, Broman K, Yandell B, Kendziorski C, Attie A. The transcription factor nfatc2 regulates β-cell proliferation and genes associated with type 2 diabetes in mouse and human islets. PLOS Genet. 2016;12(12):1–26. https://doi.org/10.1371/journal.pgen.1006466.
37. Cao Y, Gao Z, Li L, Jiang X, Shan A, Cai J, Peng Y, Li Y, Jiang X, Huang X, Wang J, Wei Q, Qin Gn, Zhao J-J, Jin X-L, Liu L, Li Y, Wang W, Wang J, Ning G. Whole exome sequencing of insulinoma reveals recurrent t372r mutations in yy1. Nat Commun. 2013;4:2810.
38. Lioubinski O, Müller M, Wegner M, Sander M. Expression of sox transcription factors in the developing mouse pancreas. Dev Dyn Off Publ Am Assoc Anatomists. 2003;227:402–8. https://doi.org/10.1002/dvdy.10311.
39. Zhang X-F, Zhu Y, Liang W-B, Zhang J-J. Transcription factor ets-1 inhibits glucose-stimulated insulin secretion of pancreatic β-cells partly through up-regulation of cox-2 gene expression. Endocr. 2013;46. https://doi.org/10.1007/s12020-013-0114-9.
40. Ebrahimi JbsuffixAandH-L, Sullivan B, Tsuchida R, Bonner-Weir S, Weir G. Beta cell identity changes with mild hyperglycemia: Implications for function, growth, and vulnerability. Mol Metab. 2020;35. https://doi.org/10.1016/j.molmet.2020.02.002.
41. Pillai R, Huypens P, Huang M, Schaefer S, Sheinin T, Wettig S, Joseph J. Aryl hydrocarbon receptor nuclear translocator/hypoxia-inducible factor-1β plays a critical role in maintaining glucose-stimulated anaplerosis and insulin release from pancreatic β-cells. J Biol Chem. 2011;286(2):1014–24.
42. Pillai R, Paglialunga S, Hoang M, Cousteils K, Prentice K, Bombardier E, Huang M, Gonzalez F, Tupling A, Wheeler M, et al. Deletion of arnt/hif1β in pancreatic beta cells does not impair glucose homeostasis in mice, but is associated with defective glucose sensing ex vivo. Diabetologia. 2015;58(12):2832–42.
43. Doyle M, Sussel L. Nkx2.2 regulates beta-cell function in the mature islet. Diabetes. 2007;56(8):1999–2007.
44. Fujiwara* T, O'Green* H, Keleş* S, Blahnik K, Linneman A, Kang Y-A, Choi K, Farnham P, Bresnick E. Discovering hematopoietic mechanisms through genomewide analysis of GATA factor chromatin occupancy. Mol Cell. 2009;36(4):667–81. *: co-first authors.
45. van der Meulen T, Huising M. The role of transcription factors in the transdifferentiation of pancreatic islet cells. J Mol Endocrinol. 2015;54(2):103.
46. Cano-Gamez E, Trynka G. From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases. Front Genet. 2020;11:424. https://doi.org/10.3389/fgene.2020.00424.
47. Zuo C, Shin S, Keleş S. atsnp: transcription factor binding affinity testing for regulatory snp detection. Bioinformatics. 2015;31(20):3353–5.
48. Baek S, Goldstein I, Hager G. Bivariate genomic footprinting detects changes in transcription factor activity. Cell Rep. 2017;19(8):1710–22.
49. Li B, Dewey C. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinforma. 2011;12(1):323.
50. Conway J, Lex A, Gehlenborg N. Upsetr: an r package for the visualization of intersecting sets and their properties. Bioinformatics. 2017;33(18):2938–40.
51. Klemm S, Shipony Z, Greenleaf W. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet. 2019;20(4):207–20.
52. Love M, Huska M, Jurk M, Schöpflin R, Starick S, Schwahn K, Cooper S, Yamamoto K, Thomas-Chollier M, Vingron M, et al. Role of the chromatin landscape and sequence in determining cell type-specific genomic glucocorticoid receptor binding and gene regulation. Nucleic Acids Res. 2017;45(4):1805–19.
53. Ong C-T, Corces V. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet. 2011;12(4):283–93.

54. Liu L, Leng L, Liu C, Lu C, Yuan Y, Wu L, Gong F, Zhang S, Wei X, Wang M, et al. An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos. Nat Commun. 2019;10(1):1–11.
55. Wen X, Lee Y, Luca F, Pique-Regi R. Efficient integrative multi-snp association analysis via deterministic approximation of posteriors. Am J Hum Genet. 2016;98(6):1114–29.
56. Lee Y, Luca F, Pique-Regi R, Wen X. Bayesian multi-snp genetic association analysis: control of fdr and use of summary statistics. bioRxiv. 2018:316471.
57. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. J R Stat Soc Ser B Stat Methodol. 2020;82(5):1273–300.
58. Lu L, Liu X, Huang W-K, Giusti-Rodríguez P, Cui J, Zhang S, Xu W, Wen Z, Ma S, Rosen J, et al. Robust hi-c maps of enhancer-promoter interactions reveal the function of non-coding genome in neural development and diseases. Mol Cell. 2020;79(3):521–34.
59. Beck T, Hastings R, Gollapudi S, Free R, Brookes A. Gwas central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. Eur J Hum Genet. 2014;22(7):949–52.
60. Greenwald W, Chiou J, Yan J, Qiu Y, Dai N, Wang A, Nariai N, Aylward A, Han J, Kadakia N, et al. Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. Nat Commun. 2019;10(1):1–12.
61. Cheng Y, Ma Z, Kim B-H, Wu W, Cayting P, Boyle A, Sundaram V, Xing X, Dogan N, Li J, et al. Principles of regulatory information conservation between mouse and human. Nature. 2014;515(7527):371–5.
62. Gjoneska E, Pfenning A, Mathys H, Quon G, Kundaje A, Tsai L-H, Kellis M. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. Nature. 2015;518(7539):365–9.
63. Villar D, Berthelot C, Aldridge S, Rayner T, Lukk M, Pignatelli M, Park T, Deaville R, Erichsen J, Jasinska A, et al. Enhancer evolution across 20 mammalian species. Cell. 2015;160(3):554–66.
64. Lynch M. Intron evolution as a population-genetic process. Proc Natl Acad Sci. 2002;99(9):6118–23.
65. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet. 2010;11(5):345–55.
66. Miguel-Escalada I, Bonàs-Guarch S, Cebola I, Ponsa-Cobas J, Mendieta-Esteban J, Atla G, Javierre B, Rolando D, Farabella I, Morgan C, et al. Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. Nat Genet. 2019;51(7):1137–48.
67. Mattis K, Gloyn A. From genetic association to molecular mechanisms for islet-cell dysfunction in type 2 diabetes. J Mol Biol. 2020;432(5):1551–78.
68. van de Bunt M, Manning Fox J, Dai X, Barrett A, Grey C, Li L, Bennett A, Johnson P, Rajotte R, Gaulton K, et al. Transcript expression data from human islets links regulatory signals from genome-wide association studies for type 2 diabetes and glycemic traits to their downstream effectors. PLoS Genet. 2015;11(12):1005694.
69. Shin S, Hudson R, Harrison C, Craven M, Keleş S. atsnp search: a web resource for statistically evaluating influence of human genetic variation on transcription factor binding. Bioinformatics. 2019;35(15):2657–9.
70. Fujimoto K, Polonsky K. Pdx1 and other factors that regulate pancreatic β-cell survival. Diabetes Obes Metab. 2009;11 Suppl 4:30–7. https://doi.org/10.1111/j.1463-1326.2009.01121.x.
71. Roman T, Cannon M, Vadlamudi R, Buchkovich M, Wolford B, Welch R, Morken M, Kwon G, Varshney A, Kursawe R, Wu Y, Jackson A, Erdos M, Kuusisto J, Laakso M, Scott L, Boehnke M, Collins F, Parker S, Mohlke K. A type 2 diabetes-associated functional regulatory variant in a pancreatic islet enhancer at the adcy5 locus. Diabetes. 2017;66:2521–30. https://doi.org/10.2337/db17-0464.
72. Kasuga M. Kcnq1, a susceptibility gene for type 2 diabetes. J Diabetes Investig. 2011;2:413–4. https://doi.org/10.1111/j.2040-1124.2011.00178.x.
73. Gamazon E, Wheeler H, Shah K, Mozaffari S, Aquino-Michaels K, Carroll R, Eyler A, Denny J, GTEx Consortium, Nicolae D, Cox N, Im H-K. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47(9):1091.
74. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx B, Jansen R, Geus E, Boomsma D, Wright F, Sullivan P, Nikkola E, Alvarez M, Civelek M, Lusis A, Lehtimäki T, Raitoharju E, Kähönen M, Seppälä I, Pasaniuc B. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48. https://doi.org/10.1038/ng.3506.
75. Barbeira A, Dickinson S, Bonazzola R, Zheng J, Wheeler H, Torres J, Torstenson E, Shah K, Garcia T, Edwards T, Stahl E, Huckins L, Nicolae D, Cox N, Im H-K. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018;9. https://doi.org/10.1038/s41467-018-03621-1.
76. Jung I, Schmitt A, Diao Y, Lee A, Liu T, Yang D, Tan C, Eom J, Chan M, Chee S, et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat Genet. 2019;51(10):1442–9.
77. Song M, Yang X, Ren X, Maliskova L, Li B, Jones I, Wang C, Jacob F, Wu K, Traglia M, et al. Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. Nat Genet. 2019;51(8):1252–62.
78. Montefiori L, Sobreira D, Sakabe N, Aneas I, Joslin A, Hansen G, Bozek G, Moskowitz I, McNally E, Nóbrega M. A promoter interaction map for cardiovascular disease genetics. Elife. 2018;7:35788.
79. Chen F, Keleş S. Surf: integrative analysis of a compendium of rna-seq and clip-seq datasets highlights complex governing of alternative transcriptional regulation by rna-binding proteins. Genome Biol. 2020;21. https://doi.org/10.1186/s13059-020-02039-7.
80. Cavalli M, Baltzer N, Umer H, Grau J, Lemnian I, Pan G, Wallerman O, Spalinskas R, Sahlén P, Grosse I, Komorowski J, Wadelius C. Allele specific chromatin signals, 3d interactions, and motif predictions for immune and b cell related diseases. Sci Rep. 2019;9(1):2695.
81. Nicolae D, Gamazon E, Zhang W, Duan S, Eileen Dolan M, Cox N. Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. PLoS Genet. 2010;6(4). https://doi.org/10.1371/journal.pgen.1000888.
82. Minnoye L, Taskiran I, Mauduit D, Fazio M, Aerschot L, Hulselmans G, Christiaens V, Makhzami S, Seltenhammer M, Karras P, Primot A, Cadieu E, van Rooijen E, Marine J-C, Egidy G, Ghanem G, Zon L, Wouters J, Aerts S.

Cross-species analysis of enhancer logic using deep learning. Genome Res. 2020:260844–120. https://doi.org/10.1101/gr.260844.120.

83. Dong C. keleslab/INFIMA: INFIMA. 2021. https://doi.org/10.5281/zenodo.5099583.
84. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–2.
85. Langmead B, Salzberg S. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9(4):357.
86. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and samtools. Bioinformatics. 2009;25(16):2078–9.
87. Wysoker A, Tibbetts K, Fennell T. Picard tools version 1.90. 2013;107(17):308. https://doi.org/http://picard.sourceforge.net. Accessed 14 Dec 2016.
88. Orchard P, Kyono Y, Hensley J, Kitzman J, Parker S. Quantification, dynamic visualization, and validation of bias in atac-seq data with ataqv. Cell Syst. 2020;10(3):298–306.
89. Consortium E, et al. The ENCODE (ENCyclopedia of DNA elements) project. Science. 2004;306(5696):636–40.
90. Schep A, Wu B, Buenrostro J, Greenleaf W. chromvar: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat Methods. 2017;14(10):975–8.
91. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):106.
92. Fornes O, Castro-Mondragon J, Khan A, van der Lee R, Zhang X, Richmond P, Modi B, Correard S, Gheorghe M, Baranašić D, et al. Jaspar 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2020;48(D1):87–92.
93. Keane T, Goodstadt L, Danecek P, White M, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 2011;477(7364):289–94.
94. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. Variantannotation: a bioconductor package for exploration and annotation of genetic variants. Bioinformatics. 2014;30(14):2076–8.
95. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge J, Sisu C, Wright J, Armstrong J, et al. Gencode reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47(D1):766–73.
96. Tarazona S, Furió-Tarí P, Turrà D, Pietro A, Nueda M, Ferrer A, Conesa A. Data quality aware analysis of differential expression in rna-seq with noiseq r/bioc package. Nucleic Acids Res. 2015;43(21):140.
97. Shabalin A. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. Bioinformatics. 2012;28(10):1353–8.
98. Mora A, Sandve G, Gabrielsen O, Eskeland R. In the loop: promoter–enhancer interactions and bioinformatics. Brief Bioinform. 2016;17(6):980–95.
99. Newton M, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics. 2004;5(2):155–76.
100. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan M, Carey V. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118.
101. Lawrence M, Gentleman R, Carey V. rtracklayer: an r package for interfacing with genome browsers. Bioinformatics. 2009;25(14):1841–2.
102. Arnold M, Raffler J, Pfeufer A, Suhre K, Kastenmüller G. Snipa: an interactive, genetic variant-centered annotation browser. Bioinformatics. 2015;31(8):1334–6.
103. Edgar R, Domrachev M, Lash A. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10.
104. Amemiya H, Kundaje A, Boyle A. The encode blacklist: identification of problematic regions of the genome. Sci Rep. 2019;9(1):1–5.
105. ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. Nature. 2012;489(7414):57.
106. Harris R. Improved pairwise alignment of genomic DNA: The Pennsylvania State University; 2007.
107. Chiaromonte F, Yap V, Miller W. Scoring pairwise genomic sequence alignments. In: Biocomputing 2002. Kauai: World Scientific; 2001. p. 115–26.
108. Kent W, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad. 2003;100(20):11484–9.
109. Schwartz S, Kent W, Smit A, Zhang Z, Baertsch R, Hardison R, Haussler D, Miller W. Human–mouse alignments with blastz. Genome Res. 2003;13(1):103–7.
110. Dong C, Keleş S. Processed data and results of the INFIMA paper. 2021. https://doi.org/10.5281/zenodo.4679897.
111. Dong C. keleslab/INFIMA-paper: Code for the INFIMA-paper. 2021. https://doi.org/10.5281/zenodo.5099585.

Publisher's Note