# Single-cell genomics improves the discovery of risk variants and genes of cardiac traits

Alan Selewa[1*], Kaixuan Luo[2*], Michael Wasney[3+], Linsin Smith[4+], Chenwei Tang[5], Heather Eckart[3], Ivan Moskowitz[2,6], Anindita Basu[3], Xin He[2#], Sebastian Pott[3#]

[1]Biophysical Sciences Graduate Program,
[2]Department of Human Genetics,
[3]Section of Genetic Medicine, Department of Medicine,
[4]Committee on Genetics, Genomics and Systems Biology,
[5]The College,
[6]Department of Pediatrics,
The University of Chicago, Chicago, IL 60637, USA;
*, +, # These authors contributed equally.

Correspondence: obasu@uchicago.edu, xinhe@uchicago.edu, spott@uchicago

## Abstract

Genome-wide association studies (GWAS) have linked hundreds of loci to cardiac diseases. However, in most loci the causal variants and their target genes remain unknown. We developed a combined experimental and analytical approach that integrates single cell epigenomics with GWAS to prioritize risk variants and genes. We profiled accessible chromatin in single cells obtained from human hearts and leveraged the data to study genetics of Atrial Fibrillation (AF), the most common cardiac arrhythmia. Enrichment analysis of AF risk variants using cell-type-resolved open chromatin regions (OCRs) implicated cardiomyocytes as the main mediator of AF risk. We then performed statistical fine-mapping, leveraging the information in OCRs, and identified putative causal variants in 122 AF-associated loci. Taking advantage of the fine-mapping results, our novel statistical procedure for gene discovery prioritized 45 high-confidence risk genes, highlighting transcription factors and signal transduction pathways important for heart development. We further leveraged our single-cell data to study genetics of gene expression. An unexpected finding from earlier studies is that expression QTLs (eQTLs) are often shared across tissues even though most regulatory elements are cell-type specific. We found that this sharing is largely driven by the limited power of eQTL studies using bulk tissues to detect cell-type-specific regulatory variants. This finding points to an important limitation of using eQTLs to interpret GWAS of complex traits. In summary, our analysis provides a comprehensive map of AF risk variants and genes, and a general framework to integrate single-cell genomics with genetic studies of complex traits.

## Introduction

Cardiac diseases are a leading cause of mortality across the world[1,2]. GWAS of cardiac traits have uncovered a large number of associations, such as >100 loci linked to atrial fibrillation (AF)[3–7]. However, in most loci the disease-driving causal variants and risk genes remain unknown due to several common challenges. Most trait-associated variants are located in non-coding regions with possible regulatory effects[8] and studies have highlighted enrichment of risk variants in cis-regulatory elements (CREs) in trait-related cell and tissue types[8–10]. Existing disease-related regulatory and epigenomic datasets, however, were often collected from bulk tissue samples that represent complex mixtures of cell types[11,12]. A lack of cell type-resolved epigenomic data thus limits our ability to interpret regulatory effects of variants. Even with comprehensive epigenomic maps, extensive linkage disequilibrium (LD) in the human genome hinders identification of causal variants in trait-associated loci. Additionally, non-coding variants are not easily associated with their target genes because of pervasive long-range gene regulation. Together, these challenges make it difficult to translate GWAS associations into molecular mechanisms.

To address these challenges in the context of heart diseases, we developed an integrated framework that unifies advances in single cell epigenomics, computational fine-mapping and a novel procedure for risk gene discovery. Specifically, we performed single-cell chromatin accessibility profiling to map CREs across major cell types in the heart. Our statistical fine-mapping method utilizes the CRE maps to infer disease-relevant cell types and takes advantage of such information to identify putative causal variants. Our novel gene-mapping approach then aggregates information of all fine-mapped SNPs to predict the risk genes, considering multiple sources of information such as distance and chromatin loops between enhancers and promoters. Application of this framework to AF revealed a number of putative risk variants and genes, highlighting biological processes important to the genetics of AF.

Motivated by our success in studying genetics of heart diseases, we took advantage of our single-cell genomics data to study genetics of gene expression, i.e., expression QTLs (eQTLs). By linking genetic variants with gene expression, eQTLs help annotate the regulatory effects of variants and have been used as key resources for interpreting GWAS findings[13,14]. An unexpected finding from eQTL studies is that eQTLs from diverse tissues show a high degree of sharing[15,16], despite cell type specificity of most CREs[17]. This finding is important, as it suggests the possibility that current eQTL studies may provide limited information for studying diseases, which are generally specific

to certain tissues or organs. Indeed, it was estimated that eQTLs from bulk gene expression may explain only 10-20% of disease heritability[18]. It has been difficult to resolve the puzzle of eQTL tissue sharing, largely due to the lack of cell-type information for identified eQTLs from bulk tissue samples. Taking advantage of our data, we were able to assign cell types to many heart eQTLs. Our analysis suggests two key factors for explaining high tissue-sharing of eQTLs, the sharing of cell types across tissues and the low sensitivity of bulk eQTL studies in detecting cell-type specific regulatory effects.

## Results

**Overview of the experimental and computational approach.** Our approach combines single-cell genomics with novel computational procedures to study genetics of cardiac traits (Fig. 1). Using single nucleus RNA-sequencing[19–21] (snRNA-seq) and single cell ATAC-seq (scATAC-seq)[22,23], we obtained transcriptome and open chromatin regions (OCRs) across all major cell types in the adult human heart (Fig. 1, step 1). These OCR profiles allow us to discover cell types enriched with the genetic risks of traits of interest. To identify specific causal variants in trait-associated loci, we performed Bayesian statistical fine-mapping, a common strategy that uses GWAS statistics as well as LD patterns to infer likely causal variants driving association signals[24]. Compared to standard fine-mapping, our method assigns prior probabilities to favor variants located in OCRs of enriched cell types (Fig. 1, step 2). The use of functionally informed prior has been shown to improve the accuracy of fine-mapping[9,25,26]. We believe this is particularly advantageous with single-cell data. Indeed, compared to OCRs from bulk tissues which include a mixture of all cell types, the OCRs in disease-relevant cell types would be particularly enriched with genetic signals. After fine-mapping, the candidate SNPs and their associated cell-type information allow us to assign the cell type(s) through which the causal variants are likely to act across disease-associated loci.

Finally, we developed a procedure to infer causal genes at each locus (Fig. 1, step 3), addressing some common challenges. In “gene association tests”, researchers test if the set of SNPs near a gene collectively show disease association[27,28]. These types of methods, however, cannot distinguish between multiple genes close to disease-associated variants. Alternatively, researchers may perform fine-mapping first, then link the high-confidence SNPs to target genes using additional information. However, fine-mapping alone rarely leads to a single, or even a few, high

confidence SNPs at associated loci[29], therefore this approach also has limited utility. In contrast, our procedure is directly informed by variant fine-mapping, but instead of considering only high-confidence SNPs, it aggregates information of all fine-mapped ones in a locus. To see its benefit, suppose fine-mapping in a locus implicates 10 putative causal variants without any single one reaching high confidence; however, if all 10 SNPs likely target the same gene, we still achieve high-confidence at the gene level. To implement this idea, our procedure partitions the evidence of a SNP being causal variant into nearby genes, with its likely target genes receiving larger evidence. The information is then aggregated across all SNPs to produce gene level evidence. The details are described below (Fig. 5a) and in Methods.

**Single-cell transcriptome and chromatin accessibility profiling reveals multiple cell types in the human heart.** We performed snRNA-seq and scATAC-seq using the Chromium platform (10x Genomics) (Fig. 1, step 1). The heart samples were obtained from the left and right ventricles (LV and RV), the interventricular septum, and the apex of three adult male donors (Supplementary Table 1). After quality control, we retained data of 49,359 cells in snRNA-seq and 26,714 cells in scATAC-seq, respectively (Extended Data Fig. 1 and 2).

We characterized cell populations with clustering analysis in both snRNA-seq and scATAC-seq datasets. From snRNA-seq[30], we identified eight major cell types based on marker genes and comparison to published single-cell heart atlas data[20] (Fig. 2a, left), with ~70% of cells from cardiomyocytes (CMs), fibroblasts, and endothelial cells. Clustering based on scATAC-seq data[31] revealed similar cell populations (Fig. 2a, right). To match the clusters identified by both technologies, we computationally transferred cluster labels from snRNA-seq onto scATAC-seq clusters[30] (Methods) and unambiguously identified matching cell types (Extended Data Fig. 3a, b). Indeed, expression and chromatin accessibility near marker genes showed high cell-type specificity (Fig. 2b, c). Across the eight clusters, gene scores inferred from scATAC-seq, a metric that summarizes the chromatin accessibility near a gene[31] (Methods), were highly correlated with transcript levels in the matched clusters (Extended Data Fig. 3c). These results supported the cell-type assignments in both modalities.

**Analysis of scATAC-seq data identifies cell-type-specific regulatory elements and their regulators.** We pooled cells of the same cell type and identified OCRs separately in each cell type.

Combining samples of the same cell type (Extended Data Fig. 4a, b), we detected 45,000-150,000 OCRs per cell type (Extended Data Fig. 4c) yielding a union set of 352,904 OCRs. *K*-means clustering of these regions based on their accessibility suggested that most OCRs are active in specific cell types (Fig. 3a). Using differential accessibility (DA) analysis, we identified 173,782 (49%) OCRs with cell-type-specific accessibility (Methods). We divided the remaining 179,122 (51%) OCRs into three categories based on their detection across cell types: shared in 2-3 cell types, shared in >=4 cell types (denoted as Shared 2-3 and Shared 4), and remaining ones, denoted as "non-DA OCRs", which mostly comprise peaks with low read counts (Methods). In agreement with previous observations, shared OCRs were enriched in promoter regions[32] (Fig. 3b, c).

We compared our OCRs to regulatory regions identified in bulk samples from multiple tissues in ENCODE[12]. As expected, a large fraction of OCRs from major heart cell types (e.g., CMs, endothelial, fibroblasts) overlapped with DNase Hypersensitive sites (DHS) from the ventricles (Fig. 3d, top). In contrast, smaller proportions of OCRs from rare cell types (e.g., myeloid) overlapped with bulk DHS, suggesting higher sensitivity of detecting regulatory elements in rare cell types by scATAC-seq (Fig. 3d, top, Extended Data Fig. 4d). Additionally, 60-80% of OCRs from major cell types overlapped with H3K27ac regions from LV and RV, suggesting enhancer activity (Fig. 3d, bottom). Together, these results showed that scATAC-seq identified cell-type specific regulatory elements.

Chromatin accessibility is largely controlled by lineage-specific transcription factors (TFs)[33]. To identify these TFs, we assessed the enrichment of TF motifs in OCRs specific to each cell type and identified 260 significantly enriched motifs (Methods). Because TFs of the same family may share similar motifs, we performed additional analysis to infer the exact TFs driving the enrichment, assuming that for these TFs, their motif enrichment should correlate with gene expression across cells. To test this, we correlated motif accessibility scores of TFs calculated by chromVar[34] with their accessibility-derived gene scores, a proxy of gene expression[31] (Methods). This analysis yielded 76 TFs with enriched motifs and correlation > 0.5 (Fig. 3e, Supplementary Table 2). Many of these TFs are cell type-specific (Fig. 3e) and include known CM regulators, such as TBX5, GATA4, and MEF2A[35] (Fig. 3f). These results provided a compendium of putative transcriptional regulators across major cell types in the human heart.

**Open chromatin regions in CMs are enriched with risk variants of heart diseases and inform statistical fine-mapping.** Using our cell-type-resolved OCRs, we assessed the contribution of different cell types to genetics of heart-related traits[36]. Risk variants from GWAS of two cardiac traits, AF and PR interval, were almost exclusively enriched (>10-fold) in OCRs from CMs (Fig. 4a). In contrast, variants of cardiovascular traits, CAD and blood pressure, were enriched across multiple cell types (Fig. 4a). As control, non-cardiovascular traits showed little or no enrichment in heart cell types (Fig. 4a). These results suggested distinct cell type origins of different heart-related traits, highlighting CMs as the main cell type underlying AF and PR interval.

This observation motivated us to statistically fine-map causal variants in 122 approximately independent AF-associated loci[37]. Our procedure favors putatively functional variants in protein-coding regions, conserved sequences, and OCRs in CMs (Extended Data Fig. 5a, Methods)[38]. Compared to fine-mapping that treats all variants equally (uniform prior), this procedure increased the number of high-confidence risk variants. In total, we identified 54 variants whose probabilities of being causal variants, denoted as Posterior Inclusion Probabilities (PIP), are 0.5 or higher, compared with 39 at PIP >= 0.5 under the uniform prior (Fig. 4bc, Supplementary Table 3). Across 122 loci, our procedure narrowed down putative causal variants to 5 or fewer SNPs in 48 loci (Fig. 4d).

The fine-mapping results inform how the risk variants are partitioned into various functional categories, such as exons and OCRs in different cell types. The sum of PIPs of all SNPs assigned to a category can be interpreted as the expected number of causal variants in that category. We found that >40% of causal signals are from OCRs and 25% of signals from CM-specific OCRs, highlighting the key role of CMs in AF (Fig. 4e). As expected, exons and UTRs explain only 4% of causal signals.

The same PIP summation approach can also be applied to each locus, with the PIP sum of a functional category, e.g., OCRs or exons, now interpreted as the probability that the causal variant in that locus falls into that category. Using this approach, we estimate that at nearly half of all loci, causal variants have >50% probability to localize to OCRs (Fig. 4f). Further partitioning of OCRs into cell-type-level categories (Fig. 3b), we identified 31 loci where the causal signals almost entirely (>90%) come from CM-OCRs (Fig. 4g). Interestingly, in three loci, the most likely cell types are fibroblast or immune cells, respectively, based on OCR annotations (Fig. 4g, Supplementary Table 4). For example, at one locus (chr17:36809344-38877404), lymphoid OCRs

192 explain 56% causal signal and the most likely target gene is *IKZF3* (see our gene mapping results
193 below), a TF involved in the regulation of lymphocyte development[39]. In another locus
194 (chr17:7317398-8306425), myeloid OCRs explain 53% causal signal, with the most likely gene
195 being *TNFSF13*, another gene with immune functions[40]. Together these results highlighted that
196 our approach can identify cell type contexts of individual loci, including the cell types missed by
197 enrichment analysis.
198
199 **Fine-mapped variants are supported by regulatory annotations and experimental validation.**
200 We characterized the regulatory functions of 54 specific variants at PIP >= 0.5. The majority
201 (31/54) were located in CM-OCRs (Fig. 4h, Supplementary Table 3). 57% (31/54) of all variants
202 and 87% (27/31) of variants in CM-OCRs overlapped H3K27ac marks in the heart, suggesting
203 enhancer activities (Fig. 4h). 37% of variants (20/54) overlapped with fetal DHS[12], suggesting that
204 these variants may act across fetal and adult stages (Fig. 4h). Additionally, 24% of variants were
205 linked to promoters through chromatin loops in Promoter-capture HiC (PC-HiC) from iPSC
206 derived CMs[41] (Fig. 4h). Using mouse ChIP-seq datasets of three key cardiac TFs (*GATA4, TBX5,*
207 *NKX2-5*)[35], we found that five candidate variants are located in human orthologous regions of TF
208 binding sites, representing 7-fold enrichment over expectation by chance (Extended Data Fig. 5b).
209 We also found that 22% (12/54) SNPs alter binding motifs (Fig. 4h) of one of the 76 TFs we
210 identified as likely transcriptional regulators in heart cell types (Fig. 3e). Together, these results
211 supported regulatory functions of many fine-mapped variants.
212 We experimentally tested six non-exonic variants with PIP > 0.95 that were located inside CM-
213 OCRs and overlapped with putative enhancers marked by H3K27ac or H3K4me1/3 (Fig. 4i,
214 Supplementary Table 5). Four out of six variant-containing OCRs induced reporter gene
215 expression in mouse HL-1 cells[42,43] (Extended Data Fig. 6a, Methods), but not in a fibroblast line
216 (3T3), suggesting cell-type-specific activity of the four OCRs (Extended Data Fig. 6b). Three out
217 of these four variants showed allelic changes of reporter activities in HL-1 cells, for at least one
218 alternative allele (Fig. 4i). The most striking effect was observed for rs7172038. Two alternative
219 alleles of this SNP (A and G) strongly reduced activation. The enhancer containing this SNP
220 interacts with the promoter of *HCN4* located about 5 kb away, according to Activity-by-Contact
221 (ABC) score[44] (Supplementary Table 3). *HCN4* is a well-known AF risk gene and is
222 physiologically implicated in cardiac rhythm control[45]. Consistent with these results, deletion of a

223 syntenic 20 kb region in mice containing this enhancer significantly reduced the expression of
224 *HCN4*[46]. Notably, in two out of three SNPs with allelic effects, the use of functional information
225 in fine-mapping significantly boosted their PIPs to >= 0.95 (PIP = 0.36 for rs7172038 and 0.39 for
226 rs1152591 under the uniform prior). These experimental results supported regulatory functions of
227 our high confidence variants.
228 In principle, we expect regulatory variants to affect transcript levels of target genes. Using GTEx
229 eQTL data from the left ventricle (LV), we found that only 29% (16/54) variants are eQTLs
230 (Supplementary Table 6). And only in three cases, the eQTLs showed plausible evidence of
231 colocalization (PP4 > 0.2 using coloc[47]) with the AF risk (Supplementary Table 6). The small
232 overlap of fine-mapped variants with heart eQTLs suggests a limitation of bulk eQTL data to
233 identify regulatory variants, an issue we will address in more detail below.
234
235 **A novel computational procedure utilizes fine-mapping results to identify AF risk genes.**
236 Despite our fine-mapping efforts, there remained considerable uncertainty of causal variants in
237 most loci (Fig. 4d). Even if the causal variants are known, assigning target genes can be difficult
238 due to long-range regulation of enhancers[48]. We developed a novel procedure, called Mapgen, to
239 address these problems (Fig. 5a): **(1)** For every putative causal SNP, we assign a weight to each
240 nearby gene, considering multiple ways a SNP may affect a gene. The weight of a gene can be
241 viewed as the probability with which a particular SNP affects that gene. For a SNP in an exon or
242 in a regulatory region linked to a particular gene, we assign a weight of 1 to that gene. When a
243 SNP cannot be linked to any gene in these ways, its target genes are assigned using a distance
244 weighted function (Fig. 5a, Methods). **(2)** The PIP of each SNP is then distributed among all
245 potential target genes according to the weights of these genes. The "fractional PIP" a gene receives
246 from a SNP can be viewed as the support the SNP provides to that gene. **(3)** For each gene, we
247 then sum over the fractional PIPs it receives from all candidate SNPs in the region. The resulting
248 "gene PIP" approximates the probability of a gene being causal (Methods). Similar to variant-level
249 fine-mapping, we also define a "credible gene set", the set of genes that capture the causal signal
250 at a locus with high probability (Methods).
251 We identified 45 genes with gene PIP >= 0.8, and 88 with gene PIP >= 0.5 (Fig. 5b,
252 Supplementary Table 7, and Table 1 for top prioritized genes). At each locus, we obtained credible
253 gene sets that captured at least 80% of the causal signal. These credible gene sets contained a single

gene in 41 out of 122 blocks, and two genes in 32 blocks (Fig. 5c, Supplementary Table 8). The genes at PIP >= 0.8 included many known AF risk genes such as TFs involved in cardiac development and atrial rhythm control (e.g. *TBX5*[49] and *PITX2*[50]), ion channels (e.g. *KCND2*[51] and *KCNN3*[52]), and genes involved in muscle contraction (e.g. *TTN*).

We note that a key benefit of Mapgen is that even in the absence of high-confidence causal variants, it may still identify putative risk genes. In 20 out of 45 genes at PIP >= 0.8, the SNP level PIPs were diffused, i.e., no single SNP reached PIP >= 0.5 (Supplementary Table 7). As an example, *CAMK2D*, an ion channel gene implicated in AF[53], was supported by eight SNPs (highest PIP = 0.43), all likely targeting *CAMK2D*. This led to a gene level PIP = 0.996 (Extended Data Fig. 7a, Supplementary Table 3). This observation thus highlighted the advantage of aggregating information from all putative causal variants.

We compared the Mapgen results with those of three common approaches for nominating target genes: closest proximity to risk SNPs, chromatin conformation that links variant-containing enhancers to target promoters, and eQTL analysis.

We first assessed the distance between supporting SNPs and their predicted target genes. Among the 45 genes at PIP >= 0.8, six (15%) were not the nearest genes to the top GWAS SNPs: *ETV1*, *TAB2*[54], *FGF9*, *PLN*[55,56], *CALU*[57] and *DBX1*. All except *DBX1* have previously described impact on cardiovascular physiology or rhythm (Supplementary Table 9). For example, *ETV1*[58,59], a TF important in heart development[60] has been recently implicated in atrial remodeling and AF[59]. *FGF9* is supported by rs9506925 (SNP PIP 0.76) which is linked, via PC-HiC, to the *FGF9* promoter 1 Mb away (Fig. 5d). FGF signaling and specifically *FGF9* have been implicated in muscle/heart development and diseases[61,62]. Another important difference between Mapgen and the common practice of choosing the nearest genes is that the latter always chooses a single candidate in a locus but does not quantify the uncertainty. For instance, rs1152591 had a PIP of 0.96, yet it contacts the promoters of four genes in PC-HiC. Two of these genes appear plausible from external evidence (*SYNE2*[63,64] and *AKAP5*[65,66]), but neither is nearest to the SNP (Extended Data Fig. 7b, Supplementary Table 3). Our gene PIPs reflect this uncertainty: all four genes had gene PIP ~0.25. Together, these results show the limitations of assigning nearest genes as targets and suggest that a probabilistic approach incorporating multiple sources of information is preferable.

We also considered the use of chromatin conformation in resolving target genes of high PIP SNPs. We found that while chromatin looping data were useful, as shown in the *FGF9* example above, using such information alone may miss many potential risk genes. Among 54 SNPs at PIP >= 0.5, only three showed chromatin interactions with promoters based on ABC scores[67], and 14 if we included both ABC and PC-HiC data. Additionally, it is common to observe multiple chromatin loops at a single SNP. Among the 14 SNPs with chromatin interactions, 64% (9/14) contact more than one promoter (Supplementary Table 3), highlighting the uncertainty of target genes from chromatin looping data.

Use of expression QTLs is another common strategy for linking SNPs to genes. However, as reported above, few fine-mapped variants colocalized with eQTLs. Even if a GWAS SNP is also an eQTL, it may not identify the correct target gene. For example, in the *TTN* locus, the top SNP (rs3731746) is an eQTL of *FKBP7*, but the true risk gene is very likely *TTN*[68,69].

Altogether, these results demonstrated the improved ability of Mapgen to nominate plausible candidate genes compared to alternative approaches linking SNPs to genes.

**Putative AF risk genes are supported by multiple lines of evidence.** We evaluated our candidate genes using multiple sources of data. Consistent with enrichment of AF variants in CM-OCRs, candidate genes (PIP >= 0.8) tended to have higher expression in CMs, compared with other genes in the AF-associated loci (Fig. 5e). Additionally, high PIP genes were enriched in AF-related Mendelian disorders (Supplementary Table 10) (Fig. 5f). We also compared our genes with those prioritized by earlier work that used additional functional data such as AF-related gene ontology and heart gene expression[5,70]. While such functional data was not used in our analysis, the genes at PIP >= 0.8 scored on average substantially higher in two earlier studies than low PIP genes (Extended Data Fig. 8), and 32 of them (71%) were supported by at least one study (Supplementary Table 7).

We next assessed the functions of candidate genes using Gene Ontology (GO) and gene networks[71]. GO analysis showed enrichment of Biological Processes related to heart development and cardiac function, and of Molecular Functions such as ion channels, hormone binding and protein tyrosine kinase (Fig. 5g, Supplementary Table 11). For network analysis, we used the STRING gene network built with genes at a relaxed PIP threshold of 0.5 (88 genes) to increase the number of interactions. This analysis highlighted some well-known processes in AF, such as ion

channels, and structure components of heart muscle (Fig. 5h). A prominent subnetwork consisted of key TFs, including *GATA4, TBX5, NKX2-5* and *HAND2*, implicated previously in AF genetics and/or heart development[49,72–74] (Fig. 5h). Two other TFs in the network, *PITX2* and *ZFHX3*, are also well-known AF genes[49]. Combined with the fact that putative causal variants were enriched in binding sites of TBX5, NKX2-5 and GATA4 (Fig. 4h, Extended Data Fig. 5b), these results suggested that perturbation of transcriptional regulatory networks consisting of TFs and their targets, plays a critical role in the genetics of AF. Additionally, the interaction network highlighted signal transduction pathways, including MAPK signaling and Ephrin signaling (Fig. 5h). Both processes are important in heart development[75–78]. Indeed, 19 out of 88 genes at PIP >= 0.5 were annotated by the GO term "regulation of intracellular signal transduction" (FDR < 0.02) (Supplementary Table 12).

Finally, we found additional literature support for the candidate genes. 37 out of 45 (82%) genes at PIP >= 0.8 have reported roles in cardiac processes and/or diseases from literature (Supplementary Table 9). The subset of genes at PIP >= 0.95 with literature support, as well as their supporting SNPs, were shown in Table 1. The majority of these genes have not been established as AF risk genes through functional studies, representing novel yet biologically plausible risk genes.

**Cell-type-specific epigenomes reveal insights to extensive tissue-sharing of bulk eQTLs.** While a large fraction of fine-mapped AF SNPs fell inside CM-specific OCRs (Fig. 4e), most of them did not colocalize with heart eQTLs (Supplementary Table 6). This observation is consistent with previous findings that only a small proportion of GWAS variants or heritability are explained by eQTLs[18,79]. We hypothesized that bulk eQTL studies, which are conducted on bulk tissues consisting of multiple cell types, may miss the gene regulatory effects of causative variants because of limited power to detect eQTLs with effects restricted to certain cell types. This hypothesis may also help explain the puzzling observation that despite the cell-type-specific nature of regulatory elements, discovered cis-eQTLs are highly shared across tissues[16]. Cell-type-resolved chromatin accessibility and transcriptome data allowed us to infer cell-type origins of bulk eQTLs and provided an opportunity to investigate eQTL tissue-sharing patterns.

We focused our analysis on 1,216 heart (LV) eQTLs from GTEx where the causal variants have been fine-mapped to single variants with high confidence (PIP >= 0.8) by the GTEx consortium

(Supplementary Table 13). We divided these eQTLs into disjoint classes based on eQTL locations, including exons, UTRs, introns, OCRs in specific cell types, and OCRs shared with varying numbers of cell types. These categories suggested possible cell-type origins of eQTLs and allowed us to compare tissue sharing patterns of different categories, e.g., eQTLs in cell-type-specific OCRs vs. those in shared OCRs across cell types (see Methods).

We first confirmed that the majority of eQTLs were highly shared, i.e., found in >30 tissues in GTEx (Fig. 6a). This high degree of sharing, however, masked heterogeneity across different categories. While eQTLs falling into OCRs shared in multiple cell types were extensively shared across tissues (Fig. 6b), eQTLs in cell-type-specific OCRs showed variable levels of sharing. Fibroblast-eQTLs (eQTLs in fibroblast-specific OCRs) and myeloid-eQTLs were highly shared (median 25 and 38 tissues, respectively), but most CM-eQTLs were found in <10 tissues (Fig. 6b). We hypothesized that this variability reflected different degrees of cell type sharing between the heart and other tissues, with fibroblasts and myeloid cells shared in more tissues and CMs shared in fewer. To test this, we compared heart eQTLs with those from the brain and whole blood. As expected, heart eQTLs from immune cell OCRs had the highest sharing with whole blood, while eQTLs of all heart cell types have low sharing with the brain (Fig. 6c). Together, our results highlighted considerable variability of tissue sharing patterns of heart eQTLs, depending on their likely cell-type origins.

This finding appeared contradictory to the overall high level of tissue sharing of eQTLs. To understand to understand the basis for this observation, we assessed the proportions of heart eQTLs in functional categories, focusing on eQTLs in OCRs, whose cell type origins could be inferred. Unexpectedly, a large proportion of those eQTLs were from OCRs shared in multiple cell types (Fig. 6d), even though more than half of all OCRs were cell-type-specific (Fig. 3a). To better understand these results, we compared the proportions of eQTLs in each category with the proportions of matched random control SNPs (Methods). While eQTLs in OCRs from single cell types showed 2-9 fold enrichment, those shared with 4 or more cell types showed 26-fold enrichment (Fig. 6d). Indeed, the enrichment is highly correlated with the number of cell types in which an OCR is detected (Fig. 6e). We thus concluded that discovered eQTLs are biased towards those with broad effects across multiple cell types, explaining the overall high tissue-sharing across eQTLs.

We reasoned that this bias towards eQTLs with shared effects, or equivalently, the depletion of cell-type-specific eQTLs, can be explained by the nature of bulk eQTL studies. When the effect of an eQTL on a gene is limited to a single cell type, but the gene is expressed in other cell types, the effect of the variant on the bulk gene expression would be diluted, leading to lower power of detecting this eQTL. This argument was supported by the observation that gene expression was less cell-type-specific than accessibility of regulatory elements. In heart eQTLs localized to CM-specific OCRs, the expression of corresponding genes in CMs were only modestly higher than their expression in other cell types (Extended Data Fig. 9a).

We performed simulations to investigate the power loss in detecting cell-type-specific eQTLs. We considered a variant that is an eQTL of a gene in one cell type ("focal" cells). Mathematical analysis showed that the power of detecting association of this variant with bulk expression depends on effect sizes of the variant in all cell types, the cell type proportions, and the variance as well as correlations of gene expression across cell types (Supplementary Notes). Under simplified assumptions about the variance and correlation of expressions across cell types, and a cell type mix similar to our heart data, we estimated that, when the focal cells are 30% of the sample, the power of detecting the eQTL, at sample size 500 and p-value < 1e-3, is about 26-88% (depending on effect size) of the maximum power; and when focal cell proportion is 20%, reduces to only 8-40% (Extended Data Fig. 9b, Supplementary Notes).

In conclusion, our empirical study and power analysis together showed that sharing of cell types across tissues, and the under-detection of cell-type-specific regulatory variants are two factors explaining high level of tissue-sharing of bulk eQTLs. The latter factor may also explain the finding that cis-eQTLs from bulk tissues only mediate 10-20% of disease heritability. Together our finding points out limitations of current eQTL studies and highlights the need of other strategies such as single-cell eQTL mapping[81].

## Discussion

While GWAS have been successful in a range of complex traits, the causal variants, their target genes, and their mechanisms in disease-related cell types have been elucidated in few cases[48]. In this work, we established a cell-type-resolved atlas of chromatin accessibility and transcription of the human heart to study the genetics of heart-related traits, focusing on AF[3–5]. We statistically fine-mapped AF-associated loci, and experimentally validated some of the candidate variants.

Using a novel computational procedure, we identified 45 high confidence genes, implicating key biological processes, in particular TFs and signaling pathways important for heart development. Motivated by our observation that the putative AF variants often were not colocalized with eQTLs, we investigated how heart eQTLs are shared across tissue types. Our analysis suggests that eQTLs with cell-type-specific effects are under-detected and that this is likely a factor explaining both high tissue-sharing of eQTLs and the lack of eQTLs in GWAS variants.

Single-cell epigenomics has been used to aid the genetic studies of several common diseases, including AF[82–85]. These studies, however, often aimed to assess the key cell types of diseases of interest and fell short of comprehensive discovery of disease-causing variants and genes. A unique strength of our work is that it takes full advantage of the single-cell data to identify candidate risk variants and genes. Our computational procedure leverages strong enrichment of genetic signals in CM-specific OCRs to fine-map causal variants, greatly increasing the number of high confidence SNPs[82] (Fig. 4b). Our gene-mapping procedure effectively leverages fine-mapping results and multiple sources of information linking SNPs to putative targets. This avoids the bias of previous work that only considers one metric, e.g., distance, to link SNPs to genes, and increases the sensitivity of detecting risk genes. As a result, we found high confidence genes (PIP >= 0.8) in more than 1/3 of known AF-associated loci.

Our set of 45 candidate genes shed light on the genetics of AF. Earlier linkage studies implicated ion channels and structural proteins, as well as a few TFs[86]. Our results confirmed these earlier findings and showed an even larger role of regulatory genes, including TFs and signaling proteins. In total, we identified 7 TFs with PIP >= 0.8 (Supplementary Table 7), and 18 at PIP >= 0.5. These included known AF genes, *TBX5* (PIP 0.99), *NKX2-5* (0.99), *PITX2* (0.9), *ZFHX3* (0.84) and *GATA4* (0.57), as well as TFs with roles in heart development such as *HAND2* (0.87), *ZEB2* (0.98), and *PRRX1* (0.74). Our results also highlighted signal transduction pathways, including MAPK signaling[75], Ephrin signaling[76–78] (Fig. 5h), G-protein coupled receptor signaling[87], Wnt signaling[88] (Supplementary Table 11) and FGF signaling[61,62] (*FGF9*, PIP = 0.94 and *FGF5* PIP = 0.53), all previously implicated in heart development.

Despite the advances described above, our study has a few limitations. Our experimental data were limited to four anatomical locations of the ventricles, while some AF risk variants might act through atrial-specific CMs. However, it is worth noting that a recent study, using scRNA-seq based cellular atlas of the heart including all anatomic locations, found that AF candidate genes

were strongly enriched in ventricular CMs[20]. Additionally, our fine-mapping leveraged the almost exclusive enrichment in CM-specific OCRs (Fig. 4a), and thus may miss variants acting on the AF risk through other cell types. This possibility is suggested by a small number of candidate variants showing accessibility specific to fibroblasts (Fig. 4g), known contributors to AF etiology[89]. Finally, some disease variants potentially act transiently during development and might be missed using adult heart samples.

Our investigation of tissue-sharing patterns of *cis*-eQTLs found that heart eQTLs located in OCRs were dominated by those with likely broad effects across cell types (Fig. 6de). This result may reflect the limited power of bulk eQTLs in detecting eQTLs acting on a low proportion of cells, a finding supported by our power analysis. As additional support, studies using both sorted cell types[90] and single-cell technology[91,92] have shown that cell-type-specific eQTLs are common.

There are some caveats to our eQTL study. We interpreted the higher enrichment of heart eQTLs in shared OCRs compared to cell-type specific OCRs (Fig. 6de) as the difference of detection power. The assumption was that eQTLs from the two groups have similar effect sizes. However, this assumption could be violated. Another caveat is that in our power analysis we assumed that the variance of gene expression across samples is identical across cell types. This reflects the limitation of our knowledge. Further work using single-cell RNA-seq or sorted cell populations may better inform the power analysis.

In conclusion, by combining novel experimental and computational approaches, our study identified a number of risk variants and genes and revealed key insights of the genetics of AF. These data provide a rich resource for future functional studies. Importantly, our analytic framework, including the software for fine-mapping and risk gene identification, may provide a general model for the study of other complex phenotypes.

## Methods

**Data collection.** *Nuclei isolation from adult heart tissue.* Heart tissue samples were obtained from National Disease Research Interchange (NDRI) without identifying information. The work with these samples was determined to be Non-Human subject research and approved by the IRB committee of the University of Chicago (IRB19-1429). Samples were stored at -80ºC and kept on dry ice whenever outside of the freezer. We included samples from 4 regions (left and right ventricles, interventricular septum, apex) from 3 male individuals (Supplementary Table 1). Aliquots of each heart sample were prepared from frozen heart tissue using a tissue pulverizer, which was cooled prior to pulverization for 20 minutes over dry ice. Aliquots assayed in this study ranged from 86.7 mg to 141.6 mg. Prior to library preparation, we purified nuclei using fluorescence-activated cell sorting (FACS) to remove debris and minimize contamination from ambient RNA.

Single nuclei isolation was performed on the heart tissue aliquots as described in Litvinukova *et al.* 2020[20], with some modifications. Single heart aliquots were kept on dry ice until being transferred into a precooled 2 mL dounce homogenizer (Sigma) with 2 mL homogenization buffer (250 mM sucrose, 25 mM KCl, 5 mM MgCl2, 10 mM Tris-HCl, 1 mM dithiothreitol (DTT), 1x protease inhibitor, 0.4 U/µl, RNaseIn, 0.2 U/µl SUPERaseIn, 0.1% Triton X-100 in nuclease-free water). Samples were dounced 25 times with pestle A (loose) and 15 times with pestle B (tight), filtered through a 40-µm cell strainer, and centrifuged (500g, 5 minutes, 4ºC). Supernatant was discarded and the nuclei pellet was suspended in nuclei resuspension buffer (1x PBS, 1% BSA, 0.2 U/µL RNaseIn) and stained with NucBlue Live ReadyProbes Reagents (ThermoFisher). Hoechst-positive nuclei were enriched using fluorescence-activated cell sorting (FACS) on the FACSAria (BD Biosciences), obtaining between 172,500 and 350,000 nuclei while targeting a maximum of 350,000. Nuclei were sorted into 0.75 ml of resuspension buffer. Flow-sorted nuclei were counted in a C-Chip Disposable Hemocytometer, Neubauer Improved (INCYTO) before commencing with library preparation.

*snRNA-seq library preparation and sequencing*. A portion of the sorted nuclei suspension was removed and brought to a concentration of between 700 and 1,200 nuclei per microliter. An appropriate number of nuclei were loaded on the Chromium controller (10X genomics) in order to target between 6,000-8,000 nuclei, according to V3 of the manufacturer's instructions for the Chromium Next GEM Single Cell 3′ Reagent Kits (10X Genomics)[93]. 3' gene expression libraries

were amplified with 15 cycles during sample index PCR. QC was performed on 3' gene expression cDNA and final libraries using a Qubit Fluorometer (ThermoFisher) and an Agilent 2100 Bioanalyzer (Agilent). Libraries were sequenced on the NovaSeq 6000 (Illumina) or the NextSeq 500 (Illumina) at the University of Chicago's Genomics Facility using paired-end sequencing.

*scATAC-seq library preparation and sequencing*. scATAC-seq libraries were prepared according to v1 of the manufacturer's guidelines for the Chromium Next GEM Single Cell ATAC Reagent Kits (10X Genomics), with the modification that we started from nuclei that were isolated as described above. Between 9,300 and 25,000 nuclei were tagmented using Transposition Mix (10X Genomics) at 37°C for 1 h and loaded on the Chromium controller. We targeted between 6,000 and 10,000 nuclei for library preparation. QC was performed on final ATAC-seq libraries using a Qubit Fluorometer and an Agilent 2100 Bioanalyzer. Libraries were sequenced on the NovaSeq6000 or the NextSeq500 at the University of Chicago's Genomics Facility using paired-end sequencing.

**Single-cell genomic data analysis.** *snRNA-seq pre-processing.* FastQ files from 12 sequencing experiments were individually processed using an in-house scRNA-seq pipeline dropRunner[94]. Briefly, dropRunner utilizes FastQC[95,96] to obtain quality control metrics followed by fast and efficient alignment to human reference genome hg38 using STARsolo 2.6.1[97] in GeneFull mode with other parameters set to default. STARsolo performs alignment and quantification of gene expression in one package. We quantified expression at the gene level using Gencode v29 gene annotations[98] utilizing both intronic and exonic reads to improve clustering and downstream analyses of the snRNA-seq data. We extracted the raw gene-by-barcode expression matrices output by STARsolo for downstream analyses. We used Seurat 3.2.1[99] in R 3.6.3 to analyze the snRNA-seq data. We combined all 12 expression matrices into a single Seurat object together with the corresponding metadata such as donor and anatomical region. To filter low-quality nuclei, we removed barcodes that contained less than 1000 UMI. We also used DoubletFinder 2.0.3[100] with pN = 0.015 and pK = 0.005 to account for doublets, which works by generating in-silico doublets and performs clustering to identify nuclei that fall in the neighborhood of the generated doublets. After quality control, we retained a total of 49,359 nuclei.

*scATAC-seq pre-processing.* FastQ files from 12 sequencing experiments were individually processed using 10x Genomics CellRanger-atac 1.2.0[101]. We used the command cellranger-atac count to align the fastq files to human reference genome hg38, followed by marking and removing duplicate reads, and producing a fragment file containing the mapped location of each unique fragment in each nucleus. We used ArchR 0.9.5[31] to further pre-process the data and perform downstream analyses of the scATAC-seq data. Using ArchR, we converted the fragments file into a tile matrix, which is a bin-by-barcode Tn5 insertion count matrix, using a bin-size of 500 bp. We also generated a gene score count matrix using the "model 42" from ArchR, which aggregates Tn5 insertion signals from the entire gene body, scales signals with bi-directional exponential decays from the TSS (extended upstream by 5 kb) and the transcription termination site, and accounts for neighboring gene boundaries. Gene annotations were obtained from Gencode v29. To filter low quality nuclei, we kept nuclei with at least 5,000 unique fragments and a TSS enrichment score of 6. We also used ArchR's doublet removal approach with default parameters, which is based on in-silico doublet generation. We removed nuclei with a doublet enrichment score greater than 1. After quality control, we retained a total of 26,714 nuclei.

*Cell-type identification from snRNA-seq and scATAC-seq.* We performed normalization, dimensionality reduction, and unsupervised clustering on snRNA-seq and scATAC-seq data in order to identify cell-types. For snRNA-seq, we used Seurat's workflow which begins with converting counts to log2 TP10k values using the NormalizeData function. Next, we found the top 2000 variable genes using FindVariableGenes and used these genes as input features for Principal Component Analysis (PCA). We computed the top 30 principal components (PCs) for each cell and used these for downstream analyses. We observed batch effects due to different donors, and corrected this batch effect. This was done using the RunHarmony function from the Harmony 1.0[102] package with default parameters to regress out the donor variable from the PCs. Next, we used the FindClusters in Seurat with a resolution of 0.2 on the harmony-corrected PCs to define clusters. We also computed the corresponding UMAP to visualize the harmony-corrected PCs in two dimensions. We used previously established cell-type markers in order to map clusters to cell types[20,21].

We performed cell-type mapping for scATAC-seq using the ArchR package. We performed dimensionality reduction on the tile matrix using the top 20,000 bins in terms of count across all

cells. We used the function addIterativeLSI with 2 iterations in order to perform latent semantic indexing (LSI) on the scATAC-seq tile matrix and retained the top 50 LSI vectors. Similar to snRNA-seq, we observed batch effects across different donors, and removed this effect using the RunHarmony function. We used addClusters with resolution = 0.2 in order to cluster nuclei based on the harmony-corrected LSI vectors. addUMAP with min.dist = 0.4 was used to compute a 2-dimensional representation of the harmony-corrected LSI vectors. We visualized gene activity scores, as defined in ArchR, using the same marker genes as in snRNA-seq to assign clusters to cell-types.

**Defining and classifying open chromatin regions**. Insertion read counts were aggregated across all cells in each cell-type to form a cell-type pseudo-bulk and peak calling was performed on pseudo-bulk data of each cell-type. Using the function addReproduciblePeakSet in ArchR in conjunction with MACS2[103], a union set of 352,900 peaks were called in total across all cell-types at FDR < 0.1. This set of peaks, called union set, were used for all downstream analyses.

In order to discover cell-type specific regulatory elements, a single-cell insertion count matrix was created using the function addPeakMatrix in ArchR. Cells were grouped into their respective cell-types and differential accessibility (DA) analysis was performed in a one-vs-all fashion, i.e., one cell type vs. all other ones. To perform DA, we used getMarkerFeatures in ArchR with default parameters, which uses the Wilcoxon rank-sum test on the log-normalized insertion count matrix. To control for technical variation, cells from the cell-type group and the group of remaining cell types are matched in terms of TSS enrichment and number of fragments. Using FDR < 10% and $\log_2$ fold-change > 1, we found about 47% of the union set to be cell-type specific.

For OCRs that were not differentially accessible, we reasoned that these are more likely to be shared. To further stratify these OCRs into different classes, based on sharing among different cell types, we used a simple quantile-based method. First, we aggregated the ATAC-seq counts across all cells within each cell-type for each non-DA peak and normalized the counts by the total sum of counts in each cell-type. Next, we binarized the peaks within each cell-type based on whether they are in the top 25% or not in terms of their normalized counts. In this way, we identify the top 25% accessible peaks in each cell-type. Finally, we count how many times a peak is 1, or highly accessible, across cell-types. Through this strategy, we defined three disjoint sets: shared in 2-3 cell types, shared in 4+ cell types and the remaining peaks denoted as “non-DA”. The last category

corresponds to peaks that are only highly accessible (top 25%) in one cell type but are not found to be differentially accessible based on our criteria above.

**Identifying putative TFs regulating chromatin accessibility.** We used a set of 870 human motif sequence instances from CisBP[104]. These motif annotations were added onto the ArchR object using the addMotifAnnotations function. Next, enrichment analysis was performed for each motif in each cell-type-specific set of peaks, using the peakAnnoEnrichment function in ArchR. The function uses the hypergeometric test to assess the enrichment of the number of times a motif overlaps with a given set of peaks, compared to random expectation. After correcting for multiple testing within each cell-type, we used FDR < 1% to ascertain a set of motifs and their enrichment.

Motif enrichment analysis may find multiple TFs with similar motifs. To reduce the redundancy and identify true TFs that drive gene regulation, we correlated the motif accessibility with gene score activity of each TF, expecting that for true TFs, their expression levels should be positively correlated with accessibility of their motifs across cells. We obtained motif accessibility scores from chromVAR (using the addDeviationsMatrix function in ArchR) for each TF across all cells. We obtained the corresponding TF gene activity scores using the "model 42" by ArchR (see "scATAC-seq pre-processing"). These single-cell-level motif accessibility scores and gene scores, however, are noisy given the sparsity of data at individual cells. We thus used a strategy similar to Cicero[105], by aggregating cells into "metacells" based on similarity using a *k*-nearest neighbor approach. Specifically, we found the *k* nearest neighbors to each cell using the LSI vectors of the single-cell ATAC-seq data. We only retained sets of metacells that shared a maximum of 25% of constituting cells. Metacells that shared more than 25% of cells were removed at random. Using *k* = 100, we created about 200 non-redundant meta-cells based on these criteria and averaged the motif accessibility scores and gene scores across cells within each meta-cell. We then computed Pearson's correlation between the gene scores and the motif accessibility scores across meta-cell. We selected all TFs with a Pearson's correlation greater than 0.5.

**Testing enrichment of GWAS risk variants in functional annotations.** We obtained harmonized GWAS summary statistics for cardiovascular and some non-cardiovascular traits from the IEU OpenGWAS project. We removed SNPs with missing values, SNPs on non-autosomal

chromosomes, and indels. Utilizing approximately independent Linkage Disequilibrium (LD) blocks generated by ldetect[37], we assigned each SNP to one of 1700 LD blocks.

We used TORUS[36] to estimate the genome-wide enrichment of risk variants of GWAS traits in various functional annotations, including cell-type specific OCRs obtained from DA testing, and some generic annotations including coding, retrieved from UCSC Genome Browser database, and conserved sequences from Lindblad-Toh, K. *et al.* 2011[106]. We ran TORUS on each annotation, one at a time, to get the marginal enrichment reported in Fig. 4a. P-values for enrichment were estimated from the 95% confidence intervals returned by TORUS and were adjusted for multiple testing across all traits/cell-types using the Benjamini-Hochberg approach.

**Fine-mapping causal variants in AF-associated loci**. We used SuSiE[107] to perform functionally-informed fine-mapping. We used the susie_rss function to fine-map each LD block, which takes GWAS z-scores and an LD matrix for the SNPs in the block. Because only summary statistics were available publicly, we used out-of-sample genotype information from 1000 Genome Project[108] to construct LD matrices. We ran SuSiE with L = 1, which allows a single causal signal for each LD block and is robust to mismatching LD patterns. We allow SNPs to have different prior probabilities in fine-mapping. These prior probabilities were generated by TORUS using a joint-model of the following annotations: CM specific ATAC, CM shared ATAC, non-CM ATAC, UCSC conserved/coding. We fine-mapped a total of 122 LD blocks, each containing at least 1 SNP at genome-wide significance ($P < 5 \times 10^{-8}$).

**Annotating putative AF causal variants with additional functional data**. Fetal DHS and heart H3K27ac data were obtained from ENCODE. PC-HiC interactions were obtained from an earlier study conducted in iPSC derived CMs[41]. Only interactions found in at least 2 out of 3 replicates were included. Motif analysis was performed using R motifbreak package[109] . Only "strong" effects on motif scores, according to the package, were considered.

**Assessing regulatory effects of candidate variants by Luciferase assay.** Candidate regulatory elements were designed from CM-specific accessibility in hg38 and synthesized by IDT, with either the reference allele or SNP allele(s). Sequence was verified and then cloned into the pGL4.23 enhancer luciferase response vector with a minimal promoter. HL-1 cardiomyocytes

were co-transfected with luciferase response vector and a pRL control using Lipofectamine 3000, cultured for 48 hr after transfection, then lysed and assayed using the Dual-Luciferase Reporter Assay system (Promega).

**Gene mapping procedure with Mapgen.** We used the posterior inclusion probabilities (PIPs) generated by SuSiE to calculate a gene-level PIP, reflecting the probability that a gene is a risk gene. We assume there is a single causal gene per disease associated locus. Let $Z_g$ be an indicator variable describing whether gene $g$ is causal ($Z_g = 1$) or not ($Z_g = 0$) for the trait. Assuming a single causal SNP per locus, the probability that the gene is causal, which is denoted as "gene PIP", can be then related to the probabilities of SNPs being causal variants:

$$P(Z_g = 1 \mid D) = \sum_i P(Z_g = 1 \mid \gamma_i = 1) P(\gamma_i = 1 \mid D),$$

where $\gamma_i$ is the indicator variable for whether SNP $i$ is causal or not, and $D$ is the GWAS summary statistics. The term $P(Z_g = 1 \mid \gamma_i = 1)$ is the probability that $g$ is the causal gene if the causal SNP is SNP $i$, and the term $P(\gamma_i = 1 \mid D)$ is simply the PIP of SNP $i$, or $\text{PIP}_i$. So the gene PIP of a gene is a weighted sum of PIPs of all SNPs, weighted by how much that gene is supported by each SNP (see below). Since the PIPs of all SNPs in a block sum to 1, the gene PIP has an upper-bound of 1. In the rare cases where a gene spans two nearby blocks - e.g. when a gene has large introns, the gene PIP may exceed 1, which can be interpreted as the expected number of causal variants targeting the gene $g$.

To calculate the term $P(Z_g = 1 \mid \gamma_i = 1)$, we consider the location of the SNP $i$ with relation to the gene $g$, as well as functional genomic data linking SNP $i$ with gene $g$. These data were used to assign the weights, denoted as $w_{ig}$, between SNP $i$ *and* gene $g$, reflecting how likely the SNP $i$ affects gene $g$. For example, if a SNP is inside an exon of a gene, then the SNP-gene will have weight 1. We note that $w_{ig}$ and $P(Z_g = 1 \mid \gamma_i = 1)$ have different semantics: it is possible that a SNP affects multiple genes with weights all equal to 1, but there is only a single causal gene supported by any SNP. In other words, for a causal SNP $i$, the conditional probabilities $P(Z_g = 1 \mid \gamma_i = 1)$ should sum to 1 across all nearby genes $g$. So we need to normalize $w_{ig}$ with:

$$PP\left(Z_g = 1 \,\middle|\, \gamma_i = 1\right) = \frac{w_{ig}}{\sum_g w_{ig}}$$

To assign the weight terms, $w_{ig}$, we follow these four steps capturing several scenarios where a SNP may affect a gene: **1)** If a SNP is in an exon or active promoter (promoter overlapping with OCR) of a gene, we assign the SNP to that gene with weight $w_{ig} = 1$. **2)** If a SNP can be linked to a gene's promoter via "enhancer loops", we assign the linked gene with weight $w_{ig} = 1$. Here, "enhancer loops" are defined based on Activity-By-Contact (ABC) scores (constructed from heart ventricle data with ABC scores >= 0.015)[44] and promoter-capture HiC data (from iPSC-CMs)[41]. Considering the fact that Hi-C and PC-HiC may miss contacts between close regions due to technical reasons, we also consider a SNP in OCR within 20 kb of an active promoter as an "enhancer loop". **3)** If a SNP is in a UTR but not in OCRs, suggesting that the SNP likely regulates the containing gene through RNA processing mechanisms, e.g. RNA stability or alternative polyadenylation, we will assign the SNP to the UTR-containing gene with weight $w_{ig} = 1$. **4)** If a SNP is not linked to any gene via the criteria above, we use a distance-based weighting to assign it to all genes within 1Mb. The weights follow an exponential decay function as below, where $d_{ig}$ is the SNP-gene distance:

$$w_{ig} = e^{-d_{ig}/5\times10^4}.$$

The parameter of this weight function, 50 kb, was chosen based on the fact that most enhancers, estimated to be 84% using CRISPR deletion experiments[110], are located within 100 kb of the target promoters. Using a weight of 50 kb here would lead to 87% of weights within 100 kb, with a simple area-under-curve calculation of the weight function above.

At any locus, having PIPs for all the genes in the locus allows us to define the "credible gene set" of the locus, much like the use of the term for SNPs[107]. Simply speaking, the credible set at the 80% level means the minimum set of genes in the locus whose sum of PIPs is greater than or equal to 80%. One complication is that some of the genes in the locus may span another nearby locus, as described above. In this case, while the final reported gene PIP is computed from both loci, we only use the PIP of the gene from the locus of interest to define the credible gene set of that locus.

**Gene interaction network analysis.** We used the STRING database (STRING) 11.5[111] to construct gene network. The analysis was done using Cytoscape 3.8.2[112]. The input genes are those at PIP >= 0.5 from our gene-mapping analysis. To create the gene network (Fig. 5h), we use all

default settings except that we use the recommended threshold for high-confidence interactions (0.700) for interaction scores. Singletons, i.e., genes not having any interactions with other ones, were not shown from the output network. We also used STRING to run functional enrichment analysis based on sources including Gene Ontology[113,114], Reactome Pathways[115] and KEGG[116].

**eQTL tissue sharing analysis.** We started with the rationale of our eQTL tissue sharing analysis. For simplicity, consider eQTLs found in one tissue (heart in our case), and we study the sharing of these eQTLs in a second tissue. Let $p$ denote the probability of eQTLs in the first tissue being shared in the second tissue. Assuming we have several functional categories of eQTLs, e.g. regulatory elements specific in a cell type, or shared across cell types, we can then break down $p$ into several categories with the simple relation:

$$p = \sum_c p_c w_c ,$$

where $c$ denotes a category, $p_c$ is the probability of tissue sharing in eQTLs from category $c$, and $w_c$ is the proportion of eQTLs in category $c$. We hypothesize that different eQTLs categories have distinct molecular mechanisms of modulating transcript levels, and thus different tissue sharing patterns. This simple analysis thus suggests that both $w_c$ and $p_c$ are important for our understanding of tissue sharing. For instance, some categories may have a highly tissue-specific pattern (low $p_c$), but may constitute a small proportion of all eQTLs (low $w_c$), thus these categories would have limited contribution to the overall level of tissue sharing among eQTLs.

*Summary statistics of GTEx heart eQTLs***.** Summary statistics of eQTLs from the left ventricle were obtained from the GTEX v8 release[14]. We also obtained fine-mapping results using DAP-G[38]. The variants with posterior inclusion probability (PIP) greater than 0.8 were kept for downstream analyses. We refer to these putative causal variants as eQTLs henceforth. The total number of eQTL-gene pairs that passed the threshold is 1,216. Tissue sharing data on the same eQTLs were also obtained from GTEx[14]. These data provide information of whether these heart eQTLs are also associated with gene expression in the other tissues in GTEx.

*Defining functional categories of heart eQTLs.* eQTLs were intersected with genomic features. To obtain a set of disjoint genomic features, we used a combination of the union peak set and generic annotations. For generic annotations, the longest transcript was chosen for each gene body, and its

corresponding exons, UTRs, and introns were obtained for all protein coding genes. We partitioned the union peak set into cell-type-specific categories based on the differential accessibility (DA) analysis, as well as the shared categories defined using the quantile approach, as described earlier. We note that DA analysis does not guarantee disjoint sets of features. Indeed, we find that cell types such as lymphoid and myeloid share about 6% of their DA peaks, while CMs share at most 1% with the other cell-types. To make these cell-type DA sets disjoint, we moved any DA peaks that occurred in multiple cell types from DA analysis, to the "Shared 2-3" and "Shared 4+" categories (see "Defining and classifying OCRs") depending on the number of cell types in which it occurred. A small percentage of peaks (< 1%) were affected by this step. The eQTLs in OCRs that overlap with exons or UTRs, or eQTLs in non-DA OCRs, are ambiguous to assign, so they were filtered from our analysis. The eQTLs in intronic OCRs were assigned based on the OCR categories. Those eQTLs not intersecting with any functional category were designated in an "unassigned" category.

*Estimating extent of tissue sharing in different categories of heart eQTLs.* GTEx has performed eQTL mapping jointly across all tissues. Using these results, we call a SNP an eQTL in a given tissue, if it passes the local false sign rate (LFSR) threshold of 1%. For any eQTL, we can thus determine the number of tissues where it is active.

*Estimating eQTL enrichment in functional categories.* All the fine-mapped heart eQTLs are assigned to our set of categories. The proportion of eQTLs in each category is then compared with the expected proportion by chance to obtain enrichment reported in Fig. 6d and 6e. We used SNPsnap[117] to create a set of random control SNPs that match our eQTLs in LD and minor allele frequency. The LD data is obtained from the European population genotypes from 1000 Genomes. We generated 1000 random SNPs which is roughly how many high-confidence eQTLs were used. The proportion of random SNPs in each category is then used as our estimated proportion by chance.

## Data availability

Our snRNA-seq and scATAC-seq data will be deposited to the Gene Expression Ombinus (GEO).

## Code availability

779 Mapgen R package is available from https://github.com/xinhe-lab/Mapgen. Code for data
780 processing and analyses are available at https://github.com/xinhe-lab/heart_atlas.
781
782
783
784 **Acknowledgements:**
785 This work was funded by National Institutes of Health (NIH) grants, R01MH110531
786 and R01HG010773 (to X.H.), and R21 AI144417-02 (to O.B). This project has been made
787 possible in part by grant number CZF2019-002431 from the Chan Zuckerberg Initiative DAF, an
788 advised fund of Silicon Valley Community Foundation. We thank Xuanyao Liu for helpful
789 comments on the manuscript.
790

# Main figures:

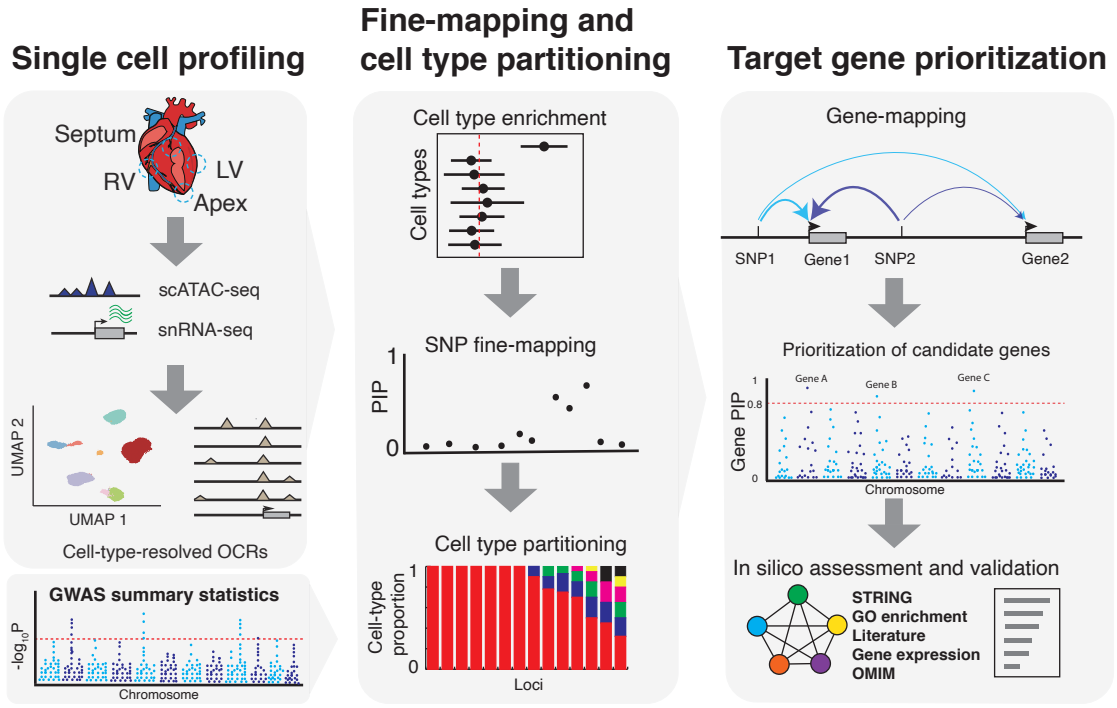


**Fig. 1 | Overview of our experimental and computational framework. Left:** SnRNA-seq and scATAC-seq profiling to cluster cells and obtain open chromatin regions (OCRs) in each cell type. **Middle:** Using OCRs and GWAS summary statistics to assess variant enrichment in cell-type-resolved OCRs. The enrichment results then provide prior for Bayesian statistical fine-mapping. The resulting Posterior Inclusion Probabilities (PIPs) represent the probabilities of variants being causal. The likely cell types through which the causal signals at each locus act can be identified by considering cell type information of likely causal variants. We may not always be able to identify a single cell type per locus, so we assign probabilities to cell types. **Right:** Computational gene-mapping using PIPs from SNP fine-mapping and SNP-to-gene links to obtain gene level PIPs. Note that the PIP of a SNP is partitioned into nearby genes in a weighted fashion, with more likely target genes receiving higher weights (as indicated by thicker arrows). Prioritized genes can be further assessed through external evidence such as gene networks and expression.

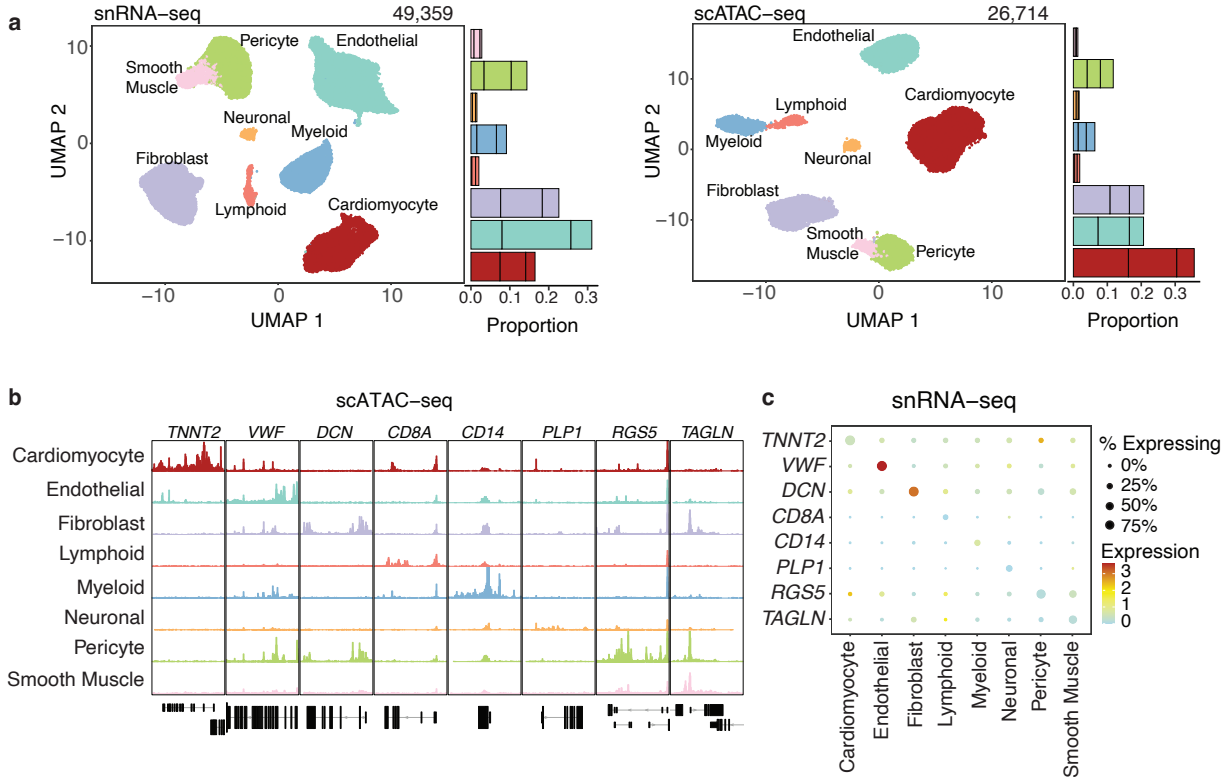


**Fig. 2 | Mapping cell types in the human heart. a,** UMAP projection of individual cells from snRNA-seq and scATAC-seq colored by cell types. Stacked barplots on the right represent the proportions of cell-types from each of the three donors. **b,** Stacked track plots of chromatin accessibility at marker genes across cell types. The bottom part shows the gene track. **c,** Percent of nuclei expressing marker genes in each cell type. Colors represent log-normalized expression values.

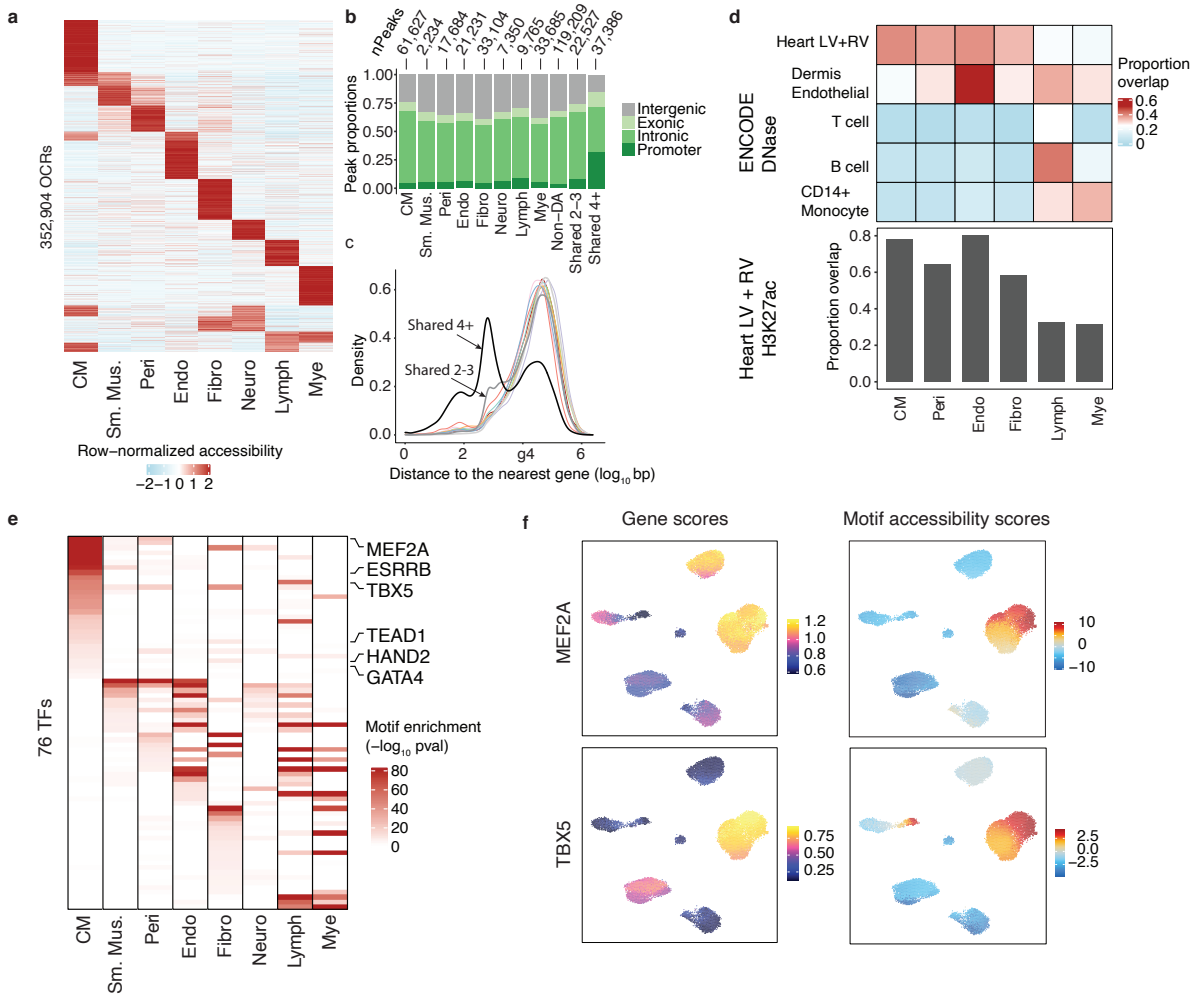


**Fig. 3 | Discovery of OCRs and transcriptional regulators in the human heart. a,** Row-normalized accessibility of OCRs across all cell types. **b,** Number of cell-type-specific and shared OCRs and their genomic distributions. **c,** Density plot of the $\log_{10}$ distance to nearest gene for all cell-type-specific and shared OCRs. Colors of the lines for cell-type-specific OCRs follow the same convention as in Figure 2a. Gray and black lines represent shared 2-3 and shared 4 OCRs. **d,** Proportions of cell-type specific OCRs that overlap with DHS (upper panel). Bar graph (lower panel) shows the proportions of cell-type specific OCRs that overlap with H3K27ac regions (LV = left ventricle, RV = right ventricle). Smooth muscle cells and neuronal cells are not shown due to the small numbers of peaks in these cell types. **e,** Enrichment of TF motifs in the OCRs specific to each cell type. Shown are 76 TFs with FDR < 1% from motif enrichment analysis in at least one cell-type, and correlation between motif enrichment and gene activity > 0.5. **f,** Gene scores (from ArchR) and motif accessibility scores calculated with chromVar in OCRs for MEF2A (top) and TBX5 (bottom) across all cells.

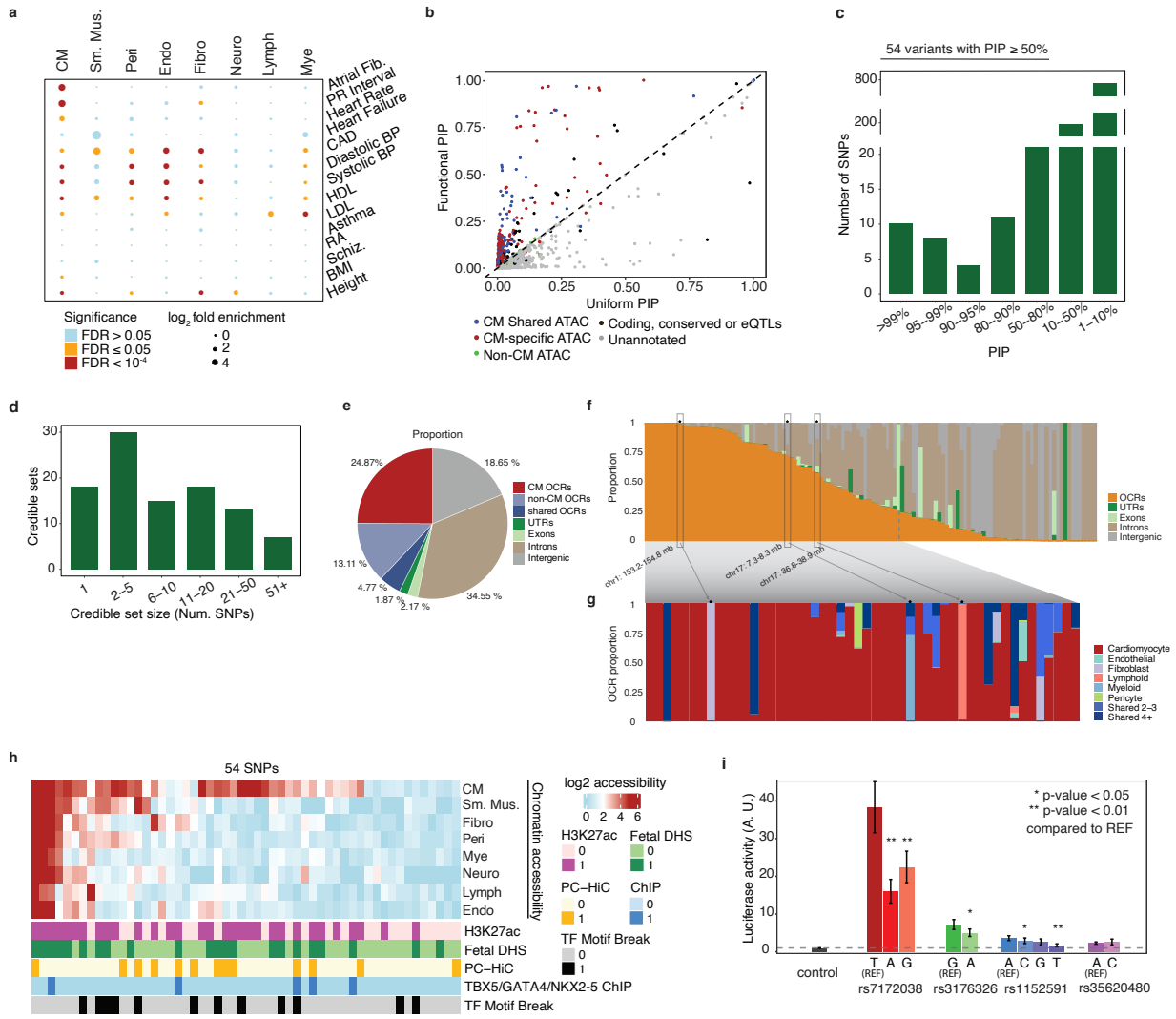


**Fig. 4 | Statistical fine-mapping of loci associated with the AF risk. a,** $\log_2$ fold enrichment (from the tool TORUS) of risk variants of various traits in cell-type-specific OCRs. **b,** Comparison of AF fine-mapping results under the informative prior using OCRs (Y-axis) vs. the results under the uniform prior (X-axis). Each dot is a SNP, and color represents the annotation of SNPs. Dashed line has a slope of 1. **c,** Summary of PIPs of variants. **d,** Summary of credible set sizes from fine-mapping of AF. **e**, Proportions of summed PIPs in disjoint functional annotation categories among all the loci. **f**, Proportion of summed PIPs in disjoint functional annotation categories at each individual locus. **g**, Proportion of summed PIPs into cell type-specific OCRs at each individual locus, for loci with summed PIPs in OCR >= 0.25. Highlighted are three loci with high proportions in fibroblast, myeloid, lymphoid specific OCRs. **h,** Chromatin accessibility and additional functional genomic annotations of 54 SNPs (PIP >= 50%). **i**, Reporter activities in HL-1 cells of regions containing selected SNPs, with both reference and alternative alleles. P-values were calculated using a paired two-sided t-test.

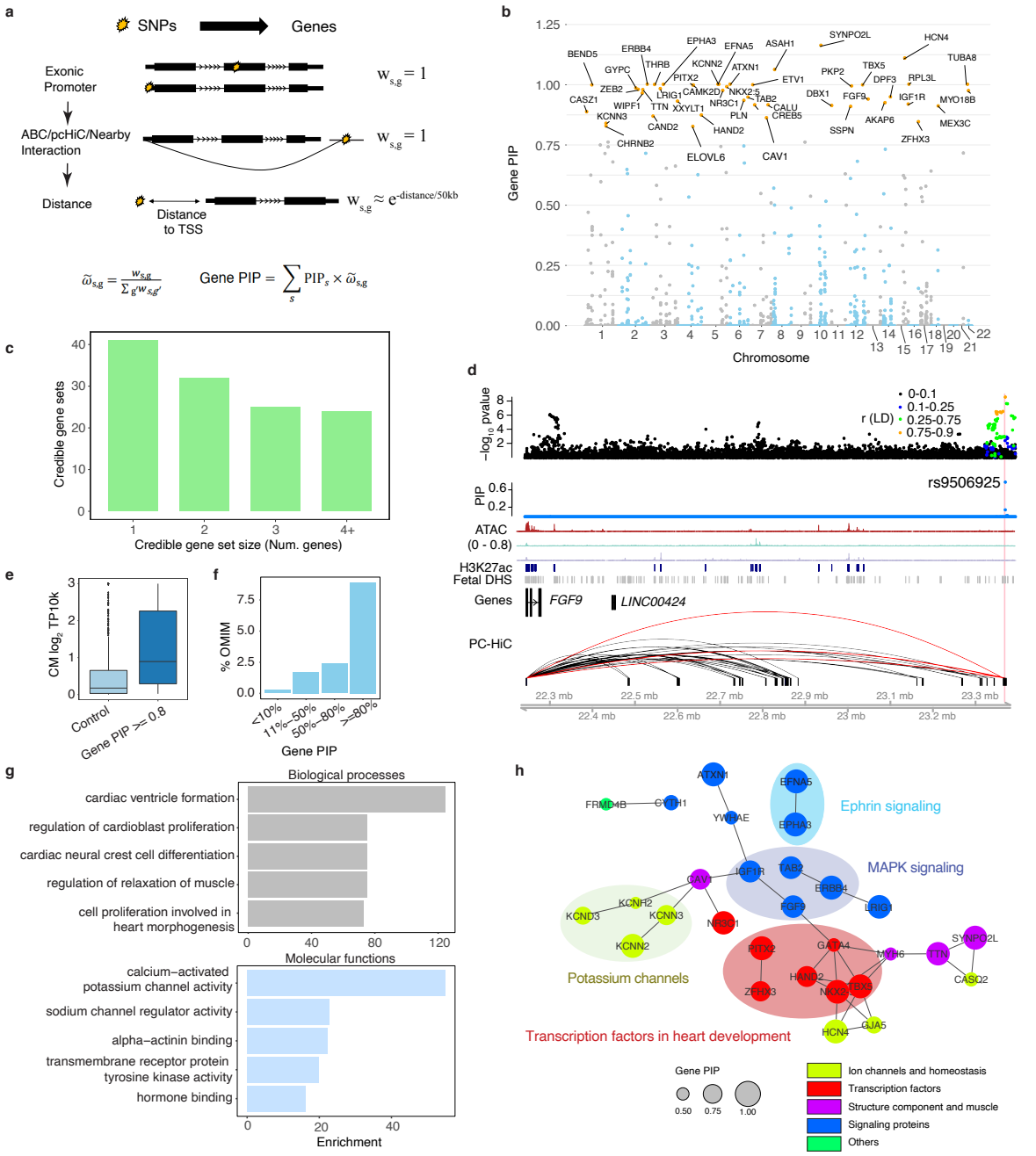


**Fig. 5 | Mapping putative risk genes of AF. a,** Schematic demonstrating the calculation of gene-level PIPs. $w_{s,g}$ represents the weight of a gene (g) with respect to a SNP (S). If a SNP is in exon or promoter of a gene, then the weight is 1 for that gene. If this is not the case, but the SNP can be linked to a target gene via regulatory interactions, we also set weight as 1. If none of these conditions apply, all nearby genes of a SNP receive distance-dependent weights. The weights are then normalized so that the total weight of all genes for a given SNP is 1. See Methods for details. **b,** Manhattan plot of gene PIPs. Genes at PIP >= 0.8 at labeled. **c,** Summary of the sizes of 80% credible gene sets from gene mapping. **d,** *FGF9* locus: the top two tracks represent the -$\log_{10}$ p-value of SNPs from AF GWAS (with color representing LD) and their PIPs from SNP-level fine-mapping, respectively. Middle three tracks represent cell-type aggregated ATAC-seq signals (CM:

red, endothelial: green; fibroblast: purple), followed by heart H3K27ac and fetal DHS peak calls. The links represent interactions identified from promoter-capture HiC data in iPSC-derived CMs. The red links show interactions centered on the likely causal SNP. **e,** Log-normalized CM expression of genes at PIP >= 80% vs. other genes from the AF loci. **f,** Percentage of Mendelian disease genes from OMIM in each gene PIP bin. **g,** Top 5 Biological Processes (BP) and Molecular Functions (MF) GO terms from gene-set enrichment analysis of the 45 genes with PIP >= 80%. **h,** Gene interaction network of candidate AF genes (PIP >= 0.5) using STRING. Only genes with interactions are shown. Interactions are defined using a confidence threshold of 0.7 by STRING. Node sizes represent gene PIPs. Colors of genes indicate their shared molecular functions.

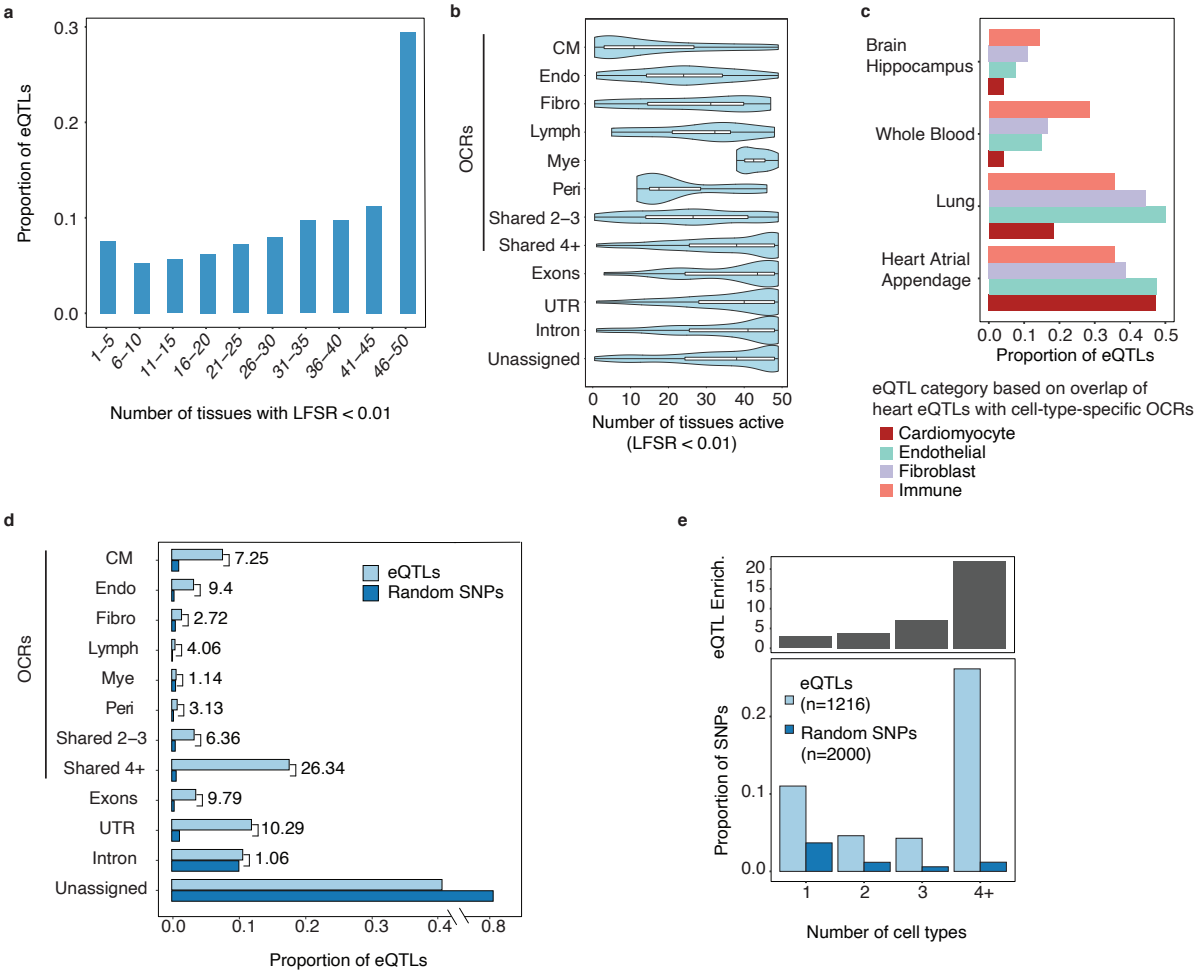


**Fig. 6 | Tissue-sharing patterns of heart (LV) eQTLs from GTEx. a,** Number of tissues where LV-eQTLs are detected at local false sign rate (LFSR) < 1%. **b,** Violin plot showing the number of tissues in which a specific eQTL is detected. Each row represents a different class of eQTLs, assigned based on their overlap with OCRs categories and other genomic locations. Unassigned: eQTLs that cannot be assigned to any functional class. **c,** Proportion of LV-eQTLs located in OCRs of selected cell types (Cardiomyocytes, Endothelial cells, Fibroblast, and Immune cells) that were also detected as eQTLs in a second tissue. **d,** Proportion of LV-eQTLs (n = 1216) in each functional class. For comparison, the proportions of random SNPs in all the classes are also shown. The numbers near the bars represent the fold enrichment in heart eQTLs compared to random SNPs. **e,** Enrichment of GTEx heart eQTLs in OCRs vary with the number of cell types where the OCRs are active. Lower panel shows the proportion of eQTLs (light blue) and control SNPs (dark blue, chosen to match eQTLs in LD and MAF) overlapping OCRs. The OCRs are divided into 4 categories, based on the degree of sharing across cell types in heart: 1= not shared, 4+= shared in >=4 cell types. The upper panel shows the enrichment of eQTLs in each OCR class compared to expectation based on control SNPs.

893 **Table 1 | Top prioritized genes (gene PIP >= 0.95).**
894

| Gene | Gene PIP | Supporting SNPs | SNP PIP | Link Method* | OMIM | CM-specific expression | Known AF risk gene | Reference |
|---|---|---|---|---|---|---|---|---|
| *SYNPO2L* | 1.161 | rs60632610 | 0.959 | exon | | | ✓ | [20215401, 33768119] |
| *HCN4* | 1.108 | rs7172038 | 0.959 | ABC | ✓ | | ✓ | [29987112] |
| *ASAH1* | 1.061 | rs7508 | 1 | exon | | ✓ | | [32015399] |
| *ATXN1* | 1.000 | rs73366713<br>rs113755256<br>rs73724866<br>rs7770062<br>rs59430691 | 0.267<br>0.205<br>0.187<br>0.158<br>0.157 | PC-HiC<br>PC-HiC<br>PC-HiC<br>PC-HiC<br>PC-HiC | | | | [21475249, 22306179] |
| *EFNA5* | 1.000 | rs6871532 | 0.288 | distance | | ✓ | | [23562676, 30909943, 25359705] |
| *ERBB4* | 1.000 | rs6738011 | 0.112 | distance | | ✓ | | [19632177] |
| *KCNN2* | 1.000 | rs337705<br>rs337708 | 0.477<br>0.119 | distance<br>distance | | | ✓ | [19139040] |
| *RPL3L* | 1.000 | rs140185678 | 1 | exon | | | | [32870709, 32514796] |
| *TUBA8* | 1.000 | rs464901<br>rs361834 | 0.853<br>0.147 | nearby OCR<br>nearby OCR | | | | [31398994] |
| *EPHA3* | 0.999 | rs35124509<br>rs6771054<br>rs2117137 | 0.375<br>0.154<br>0.108 | exons<br>distance<br>distance | | | | [17046737] |
| *THRB* | 0.999 | rs73041705<br>rs73032363<br>rs9841040<br>rs1865712 | 0.173<br>0.136<br>0.127<br>0.118 | distance<br>distance<br>distance<br>distance | | ✓ | | [28740583] |
| *ETV1* | 0.998 | rs55734480<br>rs12154315<br>rs12112152 | 0.394<br>0.335<br>0.223 | distance<br>distance<br>distance | | | ✓ | [27775552, 29930145] |
| *PITX2* | 0.997 | rs2220427 | 0.193 | distance | | | ✓ | [28217939, 29367545, 32309338] |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *TBX5* | 0.997 | rs7312625<br>rs883079<br>rs7955405<br>rs10507248 | 0.404<br>0.199<br>0.141<br>0.107 | distance<br>exon<br>PC-HiC<br>PC-HiC | | ✓ | ✓ | [28057264] |
| *CAMK2D* | 0.996 | rs17446418<br>rs2285703 | 0.43<br>0.166 | PC-HiC<br>PC-HiC | | ✓ | ✓ | [24030498] |
| *PKP2* | 0.993 | rs2045172 | 0.841 | PC-HiC | | ✓ | | [28740174] |
| *NKX2-5* | 0.990 | rs6882776<br>rs6891790<br>rs2277923<br>rs10071514 | 0.343<br>0.294<br>0.202<br>0.14 | active promoter<br>ABC/nearby OCR<br>exon<br>ABC/nearby OCR | ✓ | | | [26805889] |
| *LRIG1* | 0.982 | rs34080181<br>rs900171 | 0.322<br>0.156 | distance<br>exon | | | | [23558895, 19632177] |
| *ZEB2* | 0.979 | rs10496971 | 0.805 | distance | | | | [33398012] |
| *TTN* | 0.977 | rs3731746<br>rs2857265<br>rs3829748<br>rs3731748 | 0.317<br>0.265<br>0.182<br>0.15 | exon<br>exon<br>exon<br>exon | ✓ | ✓ | ✓ | [30535219] |
| *MYO18B* | 0.973 | rs133902<br>rs133885 | 0.609<br>0.196 | distance<br>exon | | ✓ | | [27858739] |

* In the Supporting SNPs column, only SNPs that contribute a fractional PIP (SNP PIP multiplied by the weight of the SNP to that gene) of 0.1 or more are shown. *NR3C1* is not included because it does not have any SNPs with fractional PIP >= 0.1**.**

* Nearby OCR is defined as OCR within 20 kb of active promoter of the gene.

## References

1. Benjamin, E. J. *et al.* Independent Risk Factors for Atrial Fibrillation in a Population-Based Cohort: The Framingham Heart Study. *JAMA* **271**, 840–844 (1994).

2. Kornej, J., Börschel, C. S., Benjamin, E. J. & Schnabel, R. B. Epidemiology of Atrial Fibrillation in the 21st Century. *Circ. Res.* **127**, 4–20 (2020).

3. Roselli, C. *et al.* Multi-ethnic genome-wide association study for atrial fibrillation. *Nat. Genet.* **50**, 1225–1233 (2018).

4. Roselli, C., Rienstra, M. & Ellinor, P. T. Genetics of Atrial Fibrillation in 2020. *Circ. Res.* **127**, 21–33 (2020).

5. Nielsen, J. B. *et al.* Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat. Genet.* **50**, 1234–1239 (2018).

6. Arking, D. E. et al. Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. *Nat. Genet.* **46**, 826–836 (2014).

7. Pfeufer, A. et al. Genome-wide association study of PR interval. *Nat. Genet.* **42**, 153–159 (2010).
8. Maurano, M. T. et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science **337**, 1190–1195 (2012).
9. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
10. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
11. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
12. The ENCODE Project Consortium., Moore, J.E., Purcaro, M.J. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
13. Nicolae, D. L. et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
14. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
15. The GTEx Consortium.. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
16. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).
17. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
18. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).
19. Habib, N. *et al.* Massively-parallel single nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).
20. Litviňuková, M. *et al. Cell*s of the adult human heart. *Nature* **588**, 466–472 (2020).
21. Tucker, N. R. *et al.* Transcriptional and Cellular Diversity of the Human Heart. *Circulation* **142**, 466–482 (2020).
22. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
23. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
24. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
25. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. *Am. J. Hum. Genet.* **98**, 1114–1129 (2016).

955 26. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of
956 complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
957 27. Segrè, A. V. *et al.* Common Inherited Variation in Mitochondrial Genes Is Not Enriched for
958 Associations with Type 2 Diabetes or Related Glycemic Traits. *PLoS Genetics* **6**, e1001058
959 (2010).
960 28. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set
961 analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
962 29. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-
963 density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
964 30. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell
965 transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**,
966 411–420 (2018).
967 31. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin
968 accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
969 32. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**,
970 75–82 (2012).
971 33. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory
972 Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).
973 34. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring
974 transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **528**,
975 142 (2017).
976 35. Olson, E. N. Gene regulatory networks in the evolution and development of the heart.
977 *Science* **313**, 1922–7 (2006).
978 36. Wen, X. Molecular QTL discovery incorporating genomic annotations using Bayesian false
979 discovery rate control. *Ann. Appl. Statistics* **10**, 1619–1638 (2016).
980 **37**. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in
981 human populations. *Bioinformatics* **32**, 283–285 (2016).
982 38. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient Integrative Multi-SNP Association
983 Analysis via Deterministic Approximation of Posteriors. *Am. J. Hum. Genet.* **98**, 1114–1129
984 (2016).
985 39. Morgan, B. *et al.* Aiolos, a lymphoid restricted transcription factor that interacts with Ikaros
986 to regulate lymphocyte differentiation. *EMBO J.* **16**, 2004–2013 (1997).
987 40. Baert, L., Ahmed, M. C., Manfroi, B. & Huard, B. The number 13 of the family: a
988 proliferation inducing ligand. *Curr. Opin. Immunol.* **71**, 132–137 (2021).
989 41. Montefiori, L. E. *et al.* A promoter interaction map for cardiovascular disease genetics. *Elife*
990 **7**, e35788 (2018).
991 42. Kapoor, A. et al. Multiple SCN5A variant enhancers modulate its cardiac gene expression
992 and the QT interval. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 201808734 (2019).
993 43. Claycomb, W. C. et al. HL-1 cells: A cardiac muscle cell line that contracts and retains
994 phenotypic characteristics of the adult cardiomyocyte. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 2979–

995 2984 (1998).
996 44. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**,
997 238–243 (2021).
998 45. DiFrancesco, D. HCN4, Sinus Bradycardia and Atrial Fibrillation. *Arrhythmia Electrophysiol*
999 *Rev.* **4**, 9 (2015).
1000 46. Ouwerkerk, A. F. van *et al.* Identification of Functional Variant Enhancers Associated With
1001 Atrial Fibrillation. *Circ. Res.* **127**, 229–243 (2020).
1002 47. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic
1003 Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).
1004 48. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189
1005 (2020).
1006 49. Ouwerkerk, A. F. van *et al.* Epigenetic and Transcriptional Networks Underlying Atrial
1007 Fibrillation. *Circ. Res.* **127**, 34–50 (2020).
1008 50. Nadadur, R. D. *et al.* Pitx2 modulates a Tbx5 -dependent gene regulatory network to
1009 maintain atrial rhythm. *Sci. Transl. Med.* **8**, 354ra115 (2016).
1010 51. Drabkin, M. *et al.* Nocturnal Atrial Fibrillation Caused by Mutation in KCND2 , Encoding
1011 Pore-Forming (α) Subunit of the Cardiac Kv4.2 Potassium Channel. *Circulation Genom. Precis.*
1012 *Medicine* **11**, e002293 (2018).
1013 52. Mahida, S. et al. Overexpression of KCNN3 results in sudden cardiac death. *Cardiovasc.*
1014 *Res.* **101**, 326–334 (2013).
1015 53. Purohit, A. et al. Oxidized Ca2+/Calmodulin-Dependent Protein Kinase II Triggers Atrial
1016 Fibrillation. *Circulation* **128**, 1748–**175**7 (2013).
1017 54. Yin, H. *et al.* TAB2 deficiency induces dilated cardiomyopathy by promoting RIPK1-
1018 dependent apoptosis and necroptosis. *J. Clin. Invest.* (2022) doi:10.1172/jci152297.
1019 55. Laforest, B. *et al.* Atrial fibrillation risk loci interact to modulate Ca2+-dependent atrial
1020 rhythm homeostasis. *J Clin Invest* **129**, 4937–4950 (2019).
1021 56. Dai, W. *et al.* A calcium transport mechanism for atrial fibrillation in Tbx5-mutant mice.
1022 *Elife* **8**, e41814 (2019).
1023 57. Sahoo, S. K. & Kim, D. H. Characterization of calumenin in mouse heart. *BMB Rep.* **43**,
1024 **158**–63 (2010).
1025 58. Yamaguchi, N. *et al.* Cardiac Pressure Overload Decreases ETV1 Expression in the Left
1026 Atrium, Contributing to Atrial Electrical and Structural Remodeling. *Circulation* **143**, 805–820
1027 (2021).
1028 59. Rommel, C. *et al.* The Transcription Factor ETV1 Induces Atrial Remodeling and
1029 Arrhythmia. *Circ. Res.* **123**, 550–563 (2018).
1030 60. Shekhar, A. *et al*. Transcription factor ETV1 is essential for rapid conduction in the heart. *J.*
1031 *Clin. Invest.* **126**, 4444–4459 (2016).
1032 61. Huang, J. *et al*. Fibroblast growth factor 9 (FGF9) inhibits myogenic differentiation of
1033 C2C12 and human muscle cells. *Cell Cycle* **18**, 1–19 (2019).
1034 62. Itoh, N. & Ohta, H. Pathophysiological roles of FGF signaling in the heart. *Front. Physiol.* **4**,

247 (2013).
63. Hsu, J. *et al.* Genetic Control of Left Atrial Gene Expression Yields Insights into the Genetic Susceptibility for Atrial Fibrillation. *Circulation Genom. Precis. Medicine* **11**, e002107 (2018).
64. Tsai, C.-T. *et al*. Next-generation sequencing of nine atrial fibrillation candidate genes identified novel de novo mutations in patients with extreme trait of atrial fibrillation. *J. Med. Genet.* **52**, 28 (2015).
65. Scott, J. D. & Santana, L. F. A-Kinase Anchoring Proteins. *Circulation* **121**, 1264–1271 (2010).
66. Huang, T. *et al.* AKAP5 anchors PKA to enhance regulation of the HERG channel. *Int. J. Biochem. Cell Biology* **122**, 105741 (2020).
67. Nasser, J. , Bergman, D.T., Fulco, C.P. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593,** 238–243 (2021).
68. Ahlberg, G. *et al.* Rare truncating variants in the sarcomeric protein titin associate with familial and early-onset atrial fibrillation. *Nat. Commun.* **9**, 4316 (2018).
69. Choi, S. H. *et al.* Monogenic and Polygenic Contributions to Atrial Fibrillation Risk. *Circ. Res.* **126**, 200–209 (2020).
70. Ouwerkerk, A. F. van *et al*. Identification of atrial fibrillation associated genes and functional non-coding variants. *Nat. Commun.* **10**, 4755 (2019).
71. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, gky1131 (2018).
72. Benaglio, P. *et al.* Allele-specific NKX2-5 binding underlies multiple genetic associations with human electrocardiographic traits. *Nat. Genet.* **51**, 1506–1517 (2019).
73. Schindler, Y. L. *et al.* Hand2 elevates cardiomyocyte production during zebrafish heart development and regeneration. *Development* **141**, 3112–3122 (2014).
74. Cohen, A. S. A. *et al.* Haploinsufficiency of the basic helix–loop–helix transcription factor HAND2 causes congenital heart defects. *Am. J. Med. Genet.* **182**, 1263–1267 (2020).
75. Romero-Becerra, R., Santamans, A. M., Folgueira, C. & Sabio, G. p38 MAPK Pathway in the Heart: New Insights in Health and Disease. *Int. J. Mol. Sci.* **21**, 7412 (2020).
76. O'Neal, W. T. *et al*. Ephrin–Eph signaling as a potential therapeutic target for the treatment of myocardial infarction. *Med. Hypotheses* **80**, 738–744 (2013).
77. Su, S. *et al.* Essential roles of EphrinB2 in mammalian heart: from development to diseases. *Cell Commun. Signal.* **17**, 29 (2019).
78. Chen, K. *et al.* EphB4 Forward-Signaling Regulates Cardiac Progenitor Development in Mouse ES Cells. *J. Cell Biochem.* **116**, 467–475 (2015).
79. Chun, S. *et al.* Limited statistical evidence for shared genetic effects of eQTLs and autoimmune disease-associated loci in three major immune cell types. *Nat. Genet.* **49**, 600–605 (2017).
80. Degner, J. F. *et al.* DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).

81. Wijst, M. G. P. van der *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).
82. Hocker, J. D. *et al*. Cardiac cell type–specific gene regulatory programs and disease risk association. *Sci. Adv.* **7**, eabf1444 (2021).
83. Chiou, J. *et al*. Single-cell chromatin accessibility identifies pancreatic islet cell type– and state-specific regulatory programs of diabetes risk. *Nat. Genet.* **53**, 455–466 (2021).
84. Chiou, J. *et al. I*nterpreting type 1 diabetes risk with genetics and single-cell epigenomics. *Nature* **594**, 398–402 (2021).
85. Corces, M. R. *et al.* Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.* **52**, 1158–1168 (2020).
86. Lozano-Velasco, E., Franco, D., Aranega, A. & Daimi, H. Genetics and Epigenetics of Atrial Fibrillation. *Int. J. Mol. Sci.* **21**, 5717 (2020).
87. Wang, J., Gareri, C. & Rockman, H. A. G-Protein–Coupled Receptors in Heart Disease. *Circ. Res.* **123**, 716–735 (2018).
88. Foulquier, S. *et al.* WNT Signaling in Cardiac and Vascular Disease. *Pharmacol. Rev.* **70**, 68–141 (2018).
89. Nattel, S. Molecular and Cellular Mechanisms of Atrial Fibrosis in Atrial Fibrillation. *JACC Clin. Electrophysiol.* **3**, 425–435 (2017).
90. Schmiedel, B. J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175**, 1701-1715.e16 (2018).
91. Study, L. C. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).
92. Wijst, M. G. van der *et al.* The single-cell eQTLGen consortium. *Elife* **9**, e52155 (2020).
93. Chromium Single Cell 3' Reagent Kits User Guide (v3.1 Chemistry). https://support.10xgenomics.com/single-cell-gene-expression/library-prep/doc/user-guide-chromium-single-cell-3-reagent-kits-user-guide-v31-chemistry (2019).
94. Selewa, A. *et al.* Systematic Comparison of High-throughput Single-Cell and Single-Nucleus Transcriptomes during Cardiomyocyte Differentiation. *Sci. Rep.* **10**, 1535 (2020).
95. FastQC. https://qubeshub.org/resources/fastqc (2015).
96. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
97. Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *Biorxiv* 2021.05.05.442755 (2021) doi:10.1101/2021.05.05.442755.
98. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, gky955- (2018).
99. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
100. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst.* **8**, 329-337.e4

(2019).
101. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
102. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
103. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137–R137 (2008).
104. Weirauch, M. T. *et al.* Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158**, 1431–1443 (2014).
105. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell.* **71**, 858-871.e8 (2018).
106. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–82 (2011).
107. 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. Statistical Methodol.* **82**, 1273–1300 (2020).
108. Clarke, L. *et al.* The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Res.* **45**, D854–D859 (2017).
109. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847–3849 (2015).
110. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
111. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, gkaa1074- (2020).
112. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
113. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
114. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2020).
115. Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2019).
116. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
117. Pers, T. H., Timshel, P. & Hirschhorn, J. N. SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics* **31**, 418–420 (2015).