Published in final edited form as:
Nat Genet. 2019 August 21; 51(10): 1506–1517. doi:10.1038/s41588-019-0499-3.

Allele-specific NKX2-5 binding underlies multiple genetic associations with human electrocardiographic traits

Paola Benaglio[1], **Agnieszka D'Antonio-Chronowska**[#2], **Wubin Ma**[#3], **Feng Yang**[3], **William W. Young Greenwald**[4], **Margaret K. R. Donovan**[4,5], **Christopher DeBoever**[4], **He Li**[2], **Frauke Drees**[2], **Sanghamitra Singhal**[1], **Hiroko Matsui**[2], **Jessica van Setten**[6], **Nona Sotoodehnia**[7], **Kyle J. Gaulton**[1], **Erin N. Smith**[1], **Matteo D'Antonio**[2], **Michael G. Rosenfeld**[3,*], **Kelly A. Frazer**[1,2,*]

[1]Department of Pediatrics and Rady Children's Hospital, Division of Genome Information Sciences, University of California, San Diego, La Jolla, CA, USA [2]Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA, USA [3]Howard Hughes Medical Institute, Department of Medicine, University of California, San Diego, La Jolla, CA, USA [4]Bioinformatics and Systems Biology, University of California, San Diego, La Jolla, CA, USA [5]Department of Biomedical Informatics, University of California, San Diego, La Jolla, CA, USA [6]Department of Cardiology, University Medical Center Utrecht, University of Utrecht, Utrecht, The Netherlands [7]Departments of Medicine and Epidemiology, Cardiovascular Health Research Unit, Division of Cardiology, University of Washington, Seattle, WA, USA

[#] These authors contributed equally to this work.

Abstract

The cardiac transcription factor (TF) gene *NKX2-5* has been associated with electrocardiographic (EKG) traits through GWAS, but the extent to which differential binding of NKX2-5 at common regulatory variants contributes to these traits has not yet been studied. We analyzed transcriptomic and epigenomic data from iPSC-derived cardiomyocytes (iPSC-CMs) from seven related

[*] mrosenfeld@ucsd.edu, kafrazer@ucsd.edu.

Data availability. All iPSC lines are available through WiCell Research Institute (www.wicell.org; NHLBI Next Gen Collection). All genomic data are available through dbGAP accessions phs000924 (RNA-seq, ChIP-seq, ATAC-seq, Hi-C) and phs001325 (whole-genome sequence SNV and CNV genotypes), NCBI BioProject PRJNA285375. Processed data files are available through GEO accessions GSE125540 and GSE133833.

Code availability. Custom-written code is made available via GitHub (https://github.com/frazer-lab/NKX2-5_ASE_iPSC-CM)

Author contributions

P.B. designed the study, generated ChIP-seq and RNA-seq data, and performed statistical analyses. A.D.-C. generated iPSC-CMs, ChIP-seq, ATAC-seq and RNA-seq data, and performed EMSA. W.M. generated the constructs for luciferase assay and CRISPRi and performed luciferase assays. F.Y. performed CRISPRi experiments. W.W.Y.G. implemented the fgwas analysis pipeline. C.D. implemented the RNA-seq, ATAC-seq, and allele-specific effect analyses pipelines. H.L. processed WGS and ChIP-seq data. F.D. and S.S. generated iPSC-CMs and contributed to data generation. M.K.R.D. and H.M. performed data processing and computational analyses. N.S. and J.v.S. provided summary statistics for the PR interval GWAS. K.J.G. supervised the EMSA experiments. M.D. and E.N.S. performed statistical analyses. M.G.R. supervised experimental validation of the variants. K.A.F. conceived and oversaw the study. P.B., E.N.S. and K.A.F. prepared the manuscript.

Competing Interests

The authors declare no competing interests.

Europe PMC Funders Author Manuscripts

individuals and identified ~2,000 single nucleotide variants (SNVs) associated with allele-specific effects (ASE) on NKX2-5 binding. NKX2-5 ASE-SNVs were enriched for altered TF motifs, for heart-specific eQTLs, and for EKG GWAS signals. Using fine-mapping combined with epigenomic data from iPSC-CMs, we prioritized candidate causal variants for EKG traits, many of which were NKX2-5 ASE-SNVs. Experimentally characterizing two NKX2-5 ASE-SNVs (rs3807989 and rs590041) showed that they modulate the expression of target genes via differential protein binding in cardiac cells, indicating that they are functional variants underlying EKG GWAS signals. Our results show that differential NKX2-5 binding at numerous regulatory variants across the genome contributes to EKG phenotypes.

Genome-wide association studies (GWAS) for electrocardiographic (EKG) phenotypes have found >500 risk variants[1], the majority of which are non-coding and enriched in regulatory elements of the genome. Detecting the causal variants and the molecular mechanisms that drive these associations has been challenging[2], and therefore only a handful of genetic associations with EKG traits have been explained by variants with clear molecular mechanisms[3,4].

Altered transcription factor (TF) binding has been proposed as one of the major mechanisms by which non-coding regulatory variants are causally associated with complex traits[5–7]. NKX2-5 is an evolutionary conserved, cardiac-specific TF, which, through cooperative binding with other core cardiac TFs such as TBX5 and GATA4, regulates heart development[8–11] and is implicated in a spectrum of human congenital heart defects[12]. Moreover, common non-coding variants near *NKX2-5*, *TBX5*, and *MEIS1* have been associated through GWAS[13–17] with EKG phenotypes, indicating that variation in developmental pathways plays an important role in these traits. Therefore, it is likely that genetic variation affecting the binding of developmental cardiac TFs also influences the heritability of EKG traits. However, this hypothesis has not yet been examined on a genome-wide scale.

Because the function of regulatory variants that contribute to common traits is often cell-type specific, attention to the appropriate cellular model in which to test the variants is important. Human induced pluripotent stem cell (iPSC)-derived cell types have recently emerged as a novel platform to analyze the functional consequence of genetic variants on molecular phenotypes in target cell types. iPSCs show variation in molecular phenotypes associated with their genetic background[18–20], making them a suitable model to perform expression QTL (eQTL) studies[19–24]. However, there are only a few studies showing similar utility of iPSC-derived cardiomyocytes (iPSC-CMs) to study regulatory variations[22], with potential limitations being cell-type heterogeneity that arises from directed differentiation[24–26] and the functional immaturity of iPSC-CMs[27]. Thus, while human iPSC-CMs are a promising model system, it has yet to be shown that they could enable the identification and characterization of regulatory variants that play important roles in cardiac traits.

Here we conducted a genome-wide analysis to identify regulatory variants affecting the binding of NKX2-5 and investigated their role in cardiac gene expression and EKG phenotypes. We generated iPSC-CM lines from a pedigree of seven whole-genome

sequenced individuals and profiled them with a variety of functional genomic assays, including RNA-seq, ATAC-seq, and ChIP-seq of both NKX2-5 and histone modification H3K27ac. After identifying heterozygous sites that showed ASE, we investigated NKX2-5 ASE-SNVs in detail by examining whether they altered cardiac TF motifs and whether they were enriched for eQTLs and EKG GWAS-SNPs. By applying a fine-mapping statistical approach to three GWAS studies (heart rate, atrial fibrillation, and PR interval), we prioritized putative causal variants at known (as well as novel) loci. As a proof-of-principle, we experimentally interrogated two NKX2-5 ASE-SNVs, providing evidence that they are causal variants underlying genetic associations with EKG traits. Our data show that variation affecting the binding of NKX2-5 and other cardiac TFs likely serves as a molecular mechanism underlying control of numerous EKG loci across the genome, and that fine-mapping approaches, combined with molecular phenotype data from iPSC-CMs, can be used to prioritize causal variants in EKG GWAS loci.

Results

Generation and functional genomic profiling of iPSC-derived cardiomyocytes

We generated iPSC-CMs from seven individuals in a three-generation family that includes three genetically unrelated subjects and two parent-offspring quartets (Fig. 1a and Supplementary Table 1). In total, we differentiated nine iPSC lines[28] into 26 iPSC-CM samples: 12 were harvested at day 25 after lactate selection to obtain purer cardiomyocytes, and 14 were harvested at day 15, of which one was lactate purified (Fig. 1a). After confirming the expression of cardiac markers by flow cytometer and immunofluorescence (TNNT2 and MYL7; Supplementary Fig. 1a,b and Supplementary Note), we further examined the iPSC-CMs, and the iPSCs from which they were derived, by comparing their functional genomics profiles (RNA-seq, ATAC-seq, ChIP-seq of H3K27ac and NKX2-5; Supplementary Tables 2 and 3) with those from the Roadmap Epigenomics project[14]. We confirmed that the iPSC-CMs and iPSCs, respectively, expressed cardiac-specific and stem cell-specific genes and epigenetic signatures (Fig. 1b, Supplementary Note, and Supplementary Figs. 1c and 2).

Genetic background underlies variability of molecular phenotypes in iPSC-CMs

Experimental sources of variation across the iPSC-CMs, such as differentiation efficiency, may confound the effects that are driven by different genetic backgrounds[24]. To identify sources of variability in our iPSC-CM datasets, and evaluate the contribution of genetic background to this variation compared with the iPSCs, we performed principal components (PC) analysis on each of the RNA-seq and ChIP-seq datasets, and tested whether known covariates, such as batch, TNNT2 expression (for iPSC-CMs), and subject, were associated with each of the top 10 PCs. While we observed variation in both the iPSC-CMs and iPSCs due to differentiation efficiencies and/or batch effects (Supplementary Fig. 3), the average sample-to-sample Spearman correlation of molecular phenotypes was higher between samples of the same individual than between different individuals (Mann Whitney test $P < 0.05$); additionally, samples of related individuals tended to be more correlated than samples of unrelated individuals (Fig. 1c-g). Of note, the iPSC-CMs showed slightly greater variation (i.e. lower correlation values) than the iPSCs, likely due to cellular heterogeneity[24–26].

These analyses show that genetic background was a major driver of variability in our iPSC-CM molecular datasets.

NKX2-5 peaks commonly show allele-specific effects

We examined the fraction of genetic variants associated with variable NKX2-5 peaks compared with the other molecular phenotypes by identifying heterozygous sites that showed ASE within each individual. We first merged the sequencing reads of different samples from the same subject and calculated ASE, and then, when multiple individuals carried the same heterozygous SNV, we combined the ASE results across individuals in a meta-analysis. For each phenotype, we tested between 19,371 (NKX2-5) to 123,151 (H3K27ac in iPSC-CMs) heterozygous SNVs within 12,492 to 57,631 regions (genes or peaks) (Fig. 2a) and identified the fraction of SNVs with significant imbalance at FDR < 0.05 (ASE-SNVs) (Fig. 2b). The different phenotypes showed over a 30-fold difference in the percent of ASE-SNVs, with NKX2-5 ChIP-seq having the highest fraction (10% of tested SNVs), while H3K27ac (0.7% in iPSC-CMs) and ATAC-seq (0.3% in iPSC-CMs) had considerably lower fractions. The fact that NKX2-5 ChIP-seq was so much more efficient for detecting ASE-SNVs was largely due its higher effect sizes, consistent with the fact that the assay directly measures differential TF binding, whereas ATAC-seq and H3K27ac measure altered chromatin accessibility and histone modification, respectively, which are indirect consequences of differential TF binding (Supplementary Note and Supplementary Fig. 4). Shared ASE-SNVs between iPSC-CMs and iPSCs (519 in RNA-seq and 43 in H3K27ac) showed high concordance of ASE effects (Fig. 2c) – defined as the mean proportion of the alternate allele across heterozygous sites (Spearman correlation $r > 0.85$) – indicating consistency of allelic effects between the two cell types. We further tested whether the ASE observed in heterozygous individuals was consistent with the overall effect size (ß, linear regression) on the phenotype when including homozygous samples and observed a significant ($P < 0.05$), positive relationship for all molecular phenotypes (Fig. 2d-f), with the highest correlation in NKX2-5 peaks ($r = 0.69$, Spearman correlation). These data demonstrate that the majority of allele-specific effects identified in both iPSC-CMs and iPSCs are due to genetic variation, and that, among all molecular phenotypes examined, NKX2-5 peaks had substantially more ASE-SNVs and showed the highest consistency across individuals.

NKX2-5 correlated effects are consistent with dual role as activator and repressor

Genetic loci associated with differential TF binding between individuals often show coordinated effects across different molecular traits[29]. To examine if NKX2-5 loci with ASE were correlated with H3K27ac and gene expression ASEs, we compared the effect sizes (ß) of ASE-SNVs identified within ChIP-seq peaks with the effect size of the same SNV on neighboring regions from different molecular phenotypes (nearest peak or nearest gene) (Fig. 2g,h). The strongest positive correlation was found between NKX2-5 and H3K27ac genetic effects in iPSC-CMs (Spearman correlation coefficient $r = 0.58$, $P = 1.7 \times 10^{-77}$ for NKX2-5 ASE-SNVs (Fig. 2g), and $r = 0.60$, $P = 1.6 \times 10^{-30}$ for H3K27ac ASE-SNVs), supporting the role of NKX2-5 binding in enhancer and promoter activation in these cells. However, genetic effects on NKX2-5 binding were not positively correlated with the expression of neighboring genes (Fig. 2h), possibly due to NKX2-5's dual role as an

activator or repressor[30,31]. We also observed that, when iPSC-CMs or iPSCs had H3K27ac ASE, the effect sizes were positively correlated ($r = 0.39$, $P = 3 \times 10^{-13}$; $r = 0.59$, $P = 1.3 \times 10^{-11}$) with H3K27ac peaks in nearby or overlapping regions in the other cell type, suggesting conserved genetic effects at shared enhancers and promoters. On the other hand, while H3K27ac ASE effect sizes were moderately correlated with gene expression in the corresponding cell type, they were not correlated with gene expression in the other cell type ($r = 0.33$, $P = 4 \times 10^{-12}$ and $r = 0.41$, $P = 1.7 \times 10^{-7}$ within the same cell type, and $r = 0.17$, $P = 7 \times 10^{-4}$ and $r = 0.06$, $P = 0.43$ for mismatched comparisons; Fig. 2h). These results show that, in both the iPSC-CMs and iPSCs, genetic variation underlies coordinated and cell-type specific differences across multiple molecular phenotypes; of note, while NKX2-5 and H3K27ac ASE-SNVs were highly correlated, altered NKX2-5 binding was not positively correlated with gene expression changes, consistent with a more complex function as both an activator and repressor.

Variation in cardiac TF binding motifs underlie NKX2-5 ASE-SNVs

To investigate whether NKX2-5 ASE-SNVs affected sequence motifs of TF binding sites, we selected the most enriched motifs in NKX2-5 peaks, which included the NKX2-5 homeobox motif (cognate motif), as well as motifs of other heart development TFs (GATA4, TBX5, TBX20, MEF2A/C and MEIS1; Supplementary Table 4) (secondary motifs). For both alleles of all heterozygous SNVs tested for ASE within NKX2-5 peaks, we calculated the motif position weight matrix (PWM) score of each motif. We then compared SNVs with ASE to SNVs without ASE and observed that the former were enriched for altered motifs (Fisher's exact test FDR < 0.05) (Fig. 3a). Out of the 1,941 NKX2-5 ASE-SNVs, 735 (37.8%) modified at least one of the twelve tested TF motifs: 94 (4.8%) modified both the cognate and a secondary motif, 247 (12.7%) modified only the cognate motif, and 394 (20.3%) modified one or more secondary motifs. Next, we asked whether the preferred allele (highest read count) of each ASE-SNV was associated with a higher predicted motif score. For most motifs, the preferred allele increased the motif score in 70-88% of SNVs (Fig. 3b), and the allelic proportion of ASE-SNVs positively correlated with the change in motif score, supporting an underlying causal effect for the majority of these SNVs (Fig. 3c,d and Supplementary Fig. 5). We additionally observed that ASE-SNVs tended to affect core, conserved positions within the motif more frequently than they affected less conserved positions (Fig. 3e-h), indicating a stronger effect on TF binding affinity. These data indicate that ~40% of sites containing NKX2-5 ASE-SNVs have altered motifs for NKX2-5 and/or for other known cardiac TFs, suggesting that differential allelic binding of NKX2-5 at these sites likely occurred either directly, due to alterations of its own binding sequence, or indirectly, via alterations of TF binding sites of co-binding partners.

NKX2-5 ASE-SNVs modulate cardiac-specific gene expression

We examined if NKX2-5 ASE-SNVs were associated with cardiac-specific effects on gene regulation by comparing the enrichment of NKX2-5 and H3K27ac ASE-SNVs with quantitative trait loci (QTL) from diverse cell types, including DNase hypersensitivity QTLs (dsQTLs) in lymphoblastoid cell lines (LCLs)[32], expression QTLs (eQTLs) from iPSCs[21], and eQTLs from 13 combined studies obtained from Haploreg[33] ("combined tissues") (Fig. 4a-c and Supplementary Table 5). In iPSC-CMs, H3K27ac ASE-SNVs were enriched over

Europe PMC Funders Author Manuscripts

SNVs without ASE for all three types of QTLs (Fisher's exact test $P < 0.05$); on the other hand, H3K27ac ASE-SNVs in iPSCs were only enriched for iPSC eQTLs. Of note, NKX2-5 ASE-SNVs were significantly depleted for iPSC and combined tissue eQTLs, suggesting that they exert regulatory functions only in cardiac tissues.

We therefore investigated if NKX2-5 ASE-SNVs were enriched for heart-specific eQTLs. NKX2-5 and H3K27ac ASE-SNVs were compared with SNVs without ASE to assess enrichment for tissue-specific eQTLs (defined in methods) in 26 tissue types from the GTEx project (v6)[34]. ASE-SNVs in both NKX2-5 and H3K27ac peaks in iPSC-CMs were more enriched for heart-specific eQTLs (Fig. 4d and Supplementary Table 5) than other tissue-specific eQTLs, while H3K27ac ASE-SNVs in iPSCs were not enriched for any GTEx tissue-specific eQTL. Notably, there were 55 NKX2-5 ASE-SNVs that overlapped a heart-specific eQTL, of which 9 affected the NKX2-5 binding motif, and 13 affected one or more of the other cardiac TF motifs in Figure 3 (Supplementary Table 5). These results indicate that ASE-SNVs in the iPSC-CM lines are enriched for tissue-specific regulatory variants associated with molecular traits in previous studies. Overall, consistent with its importance as a cardiac identity transcriptional regulator, we found that SNVs affecting the binding of NKX2-5 and other cardiac TFs (with which NKX2-5 cooperatively binds) are likely to underlie cardiac-specific eQTLs.

NKX2-5 ASE-SNVs are enriched for GWAS associations with EKG traits

Based on the fact that GWAS variants near the *NKX2-5* gene have been previously associated with EKG traits[13–15,35,36], we hypothesized that the altered binding of NKX2-5 in other GWAS loci could be causally implicated in these traits. We first examined if NKX2-5, H3K27ac, or ATAC peaks from iPSC-CMs were enriched for GWAS-SNPs for six EKG traits (heart rate, PR interval, QT interval, QRS duration, atrial fibrillation (AF) and P-wave duration), compared with GWAS-SNPs from 119 other traits having a comparable number of associated SNPs. We observed a strong relative enrichment for several EKG traits (Binomial test FDR < 0.05, Fig. 5a-c and Supplementary Fig. 6), with QRS duration GWAS-SNPs and heart rate GWAS-SNPs being the top two enriched traits in NKX2-5 peaks. We also examined H3K27ac and DHS peaks from Roadmap cardiac tissues, which similarly showed high enrichment for all EKG GWAS-SNPs, while H3K27ac and DHS peaks from iPSCs did not (Supplementary Fig. 6). These data show that enhancer regions in iPSC-CM and Roadmap cardiac tissues both show enrichment for EKG trait-specific regulatory variants.

To examine if differential binding of NKX2-5 might have a role in EKG phenotypes, we determined if NKX2-5 ASE-SNVs were enriched for being EKG GWAS-SNPs. In total, there were 121 SNPs that were associated with any of the six EKG traits and were within NKX2-5 peaks, of which 81 were heterozygous in the family and had sufficient read coverage to be tested for ASE. Fourteen of these GWAS-SNPs (17%) were NKX2-5 ASE-SNVs (Table 1), which were significantly enriched compared with the proportion of NKX2-5 ASE-SNVs overlapping heterozygous non-GWAS-SNPs (1,926/19,290 (10%), Fisher's exact test, OR = 1.88, $P = 0.0392$, Fig. 5d). Among these 14 NKX2-5 ASE-SNVs at EKG GWAS loci, seven were evolutionary conserved in mammals (SiPhy conservation[33])

and/or altered a cardiac TF motif (Table 1), and three overlapped heart-specific eQTLs from GTEx. These results suggest a functional link between NKX2-5 binding, cardiac-specific gene expression, and EKG phenotypes at these loci.

Validation of NKX2-5 ASE-SNV in the *SSBP3* locus as a functional regulatory variant

To provide evidence that NKX2-5 ASE-SNVs within EKG GWAS loci could be functional, we experimentally investigated the SNV that showed the strongest evidence for allelic imbalance: rs590041 (NC_000001.10:g.54742471T>C) (Table 1). Two SNPs in the transcription factor *SSBP3* locus are in perfect LD and showed ASE in the same peak; while rs562408 (NC_000001.10:g.54742618A>G) was the lead variant in a P-wave duration GWAS[37], our data suggested that rs590041 is the likely functional variant, as it is more centrally located in the peak and alters both TBX5 and NKX2-5 motifs (Fig. 5e). We confirmed that rs590041 had a direct causal effect on NKX2-5 binding by electrophoretic mobility shift assay (EMSA), showing that the alternate (C) allele, which creates an NKX2-5 motif, had stronger binding to nuclear extract from iPSC-CMs (Fig. 5f), consistent with the allelic imbalance that we identified in NKX2-5 ChIP-Seq (Fig. 5e). Interestingly, the stronger NKX2-5 binding C allele was associated with lower *SSBP3* expression in human atrial appendages (GTEx) (Fig. 5g), suggesting a repressive function of the regulatory element harboring rs590041. In luciferase assays in iPSC-CMs (Fig. 5h), sequences encoding both alleles showed lower expression than the control, but the stronger NKX2-5 binding C allele was significantly lower than the T allele, additionally supporting a repressive function of NKX2-5 binding in this region. This hypothesis was further substantiated by the fact that specific dCas9-KRAB blocking (CRISPRi) of the region resulted in increased expression of *SSBP3* in iPSC-CMs (Fig. 5i). Of note, there is no previously described role for *SSBP3* in EKG phenotypes. Altogether, these data show that rs590041 is a regulatory variant that represses the expression of *SSBP3* in cardiac cells, and suggest that it likely underlies the association of P-wave duration in this locus.

NKX2-5 ASE-SNVs prioritize causal variants in heart-rate GWAS loci

To examine more broadly whether NKX2-5 ASE-SNVs could help prioritize causal variants for EKG traits, we utilized fgwas[38], a statistical framework that integrates functional genomics annotations and GWAS summary statistics to identify putative causal variants at know loci, as well as at potentially novel loci. We initially applied a single annotation model to examine a heart rate[15] meta-analysis to determine if genetic associations were enriched within each individual iPSC-CM genomic annotation (NKX2-5, H3K27ac, and ATAC-seq peaks, and NKX2-5 ASE-SNVs and H3K27ac ASE-SNVs). We found NKX2-5 ASE-SNVs were the most enriched annotation, followed by NKX2-5 peaks (Supplementary Fig. 7). We next applied a joint model, where the association enrichment was quantified simultaneously for all five annotations and refined using 10-fold cross-validation, and found again NKX2-5 ASE-SNVs to be the most significantly enriched, followed by H3K27ac peaks (Fig. 6a). Then, to prioritize causal variants, we used the enrichment estimates from the joint model as priors to update the probability for a variant to be causal (posterior probability of association, PPA) within consecutive 1-Mb windows across the genome. We found 21 variants with greater than 30% probability of being causal, of which seven (30%) were NKX2-5 ASE-SNVs (Supplementary Table 6), suggesting that altered binding of NKX2-5

accounts for a considerable fraction of the genome-wide genetic contribution underlying variable heart rate. Out of these seven NKX2-5 ASE-SNVs (Fig. 6b), four were from "sub-threshold" loci that did not reach genome-wide significance in the heart rate[15] meta-analysis. One of these variants, rs6801957 (NC_000003.11:g.38767315T>C), identified with a 35% PPA, did not reach genome-wide significance in the heart rate[15] meta-analysis, but was significantly associated in a larger heart-rate GWAS[39], as well as in several GWAS for multiple EKG traits[13,14,40–44]. While we predicted that rs6801957 altered a T-box binding sequence and resulted in differential co-binding of NKX2-5 (Fig. 6c), previous functional experiments showed that this variant affects binding of TBX3 and TBX5 and expression of *SCN5A*, the main cardiac sodium channel[3,45]. Thus, rs6801957 serves as a proof of principle for using NKX2-5 ASE-SNVs to identify causal variants at known EKG trait GWAS loci as well as identify novel associated loci.

To further investigate the mechanisms of association between heart rate and NKX2-5 ASE-SNVs identified as candidate causal variants by fgwas (Fig. 6b), we followed up three loci previously associated with heart rate (rs7612445, NC_000003.11:g.179172979G>T; rs8044595, NC_000016.9:g.15906130A>G; and rs6606689, NC_000012.11:g.110975675T>C) and a potential novel locus (rs176107, NC_000005.9:g.89392662A>G) with additional experimental data (Supplementary Note). These data included Hi-C chromatin conformation maps from the same iPSC-CM samples[46] (Supplementary Table 2a), and RNA-seq data from iPSC-CMs from an additional 128 whole-genome-sequenced subjects[26], to examine associations between the putative causal NKX2-5 ASE-SNVs and expression of nearby or distal candidate target genes. For rs7612445 (98% PPA), which altered a T-box motif in the *GNB4* locus, we validated that the two alleles have differential binding using EMSA, and that it is associated with differential expression in iPSC-CMs of several genes, including *GNB4* (heart specific eQTL in GTEx) and *MFN1* (influencing heart rate in zebrafish and Drosophila[15]; Supplementary Fig. 8a-c). rs8044595 (89% PPA) was associated with expression of multiple genes within the same chromatin loop in iPSC-CMs, including a strong candidate *NOMO3* (nodal signaling protein associated with heart defects) (Supplementary Fig. 8d,e). rs6606689 (86% PPA) was associated with *ARPC3* gene expression, an actin cytoskeleton regulator (Supplementary Fig. 8f,g). For rs176107 (35% PPA), Hi-C showed numerous long-range interactions including with the key cardiac TF *MEF2C* (~1.2 Mb distal) and it was also associated with expression of *MEF2C* in iPSC-CMs (Supplementary Fig. 8h,i). Overall, these results uncover plausible molecular mechanism underlying variability in heart rate, both at novel and previously identified GWAS loci.

Validation of NKX2-5 ASE-SNV rs3807989 as a functional variant at the *CAV1* locus

To examine other EKG traits, we applied the fgwas fine-mapping framework to both atrial fibrillation[47] and PR interval[17] GWAS studies (Fig. 7a and Supplementary Fig. 7), and identified 26 and 102 SNPs, respectively, with greater than 30% probability of being causal, of which 8% (2/26) and 14% (14/102) were NKX2-5 ASE-SNVs (Supplementary Table 6). In both the AF and PR interval fgwas analyses, rs3807989 (NC_000007.13:g.116186241A>G) had the highest probability of being causal (>99% PPA) (Fig. 7b,c), and therefore, we experimentally investigated potential mechanisms underlying these

associations. rs3807989, located within the *CAV1* associated interval, has been reported as an eQTL for both *CAV1* and *CAV2* (encoding caveolins, scaffolding proteins involved in various signaling pathways) in multiple tissues[34,48,49], including left atrial samples[17]. This eQTL was reproduced in our 128 iPSC-CMs (Fig. 7d), confirming that there is a clear genetic association between rs3807989 and expression levels of *CAV1* and *CAV2* in cardiomyocytes. To provide evidence that this SNP is directly responsible for differential regulatory activity, we performed EMSA using iPSC-CM nuclear extracts, which demonstrated that oligonucleotide probes for the reference allele (A) bound more strongly than those for the alternate allele (G), consistent with the allelic imbalance that we identified in NKX2-5 ChIP-seq (Fig. 7e). Although rs3807989 was not predicted to directly modify a motif for NKX2-5 or other cardiac TFs, the SNV is located 6 bp from a NKX2-5 motif (Fig. 7f), and could modify a sequence important for the recognition of the binding site, such as those affecting DNA shape[50–52]. Furthermore, we observed consistent allele-specific enhancer activity in iPSC-CMs by luciferase assays (Fig. 7g). Finally, by repressing the rs3807989-containing genomic region using dCas9-KRAB (CRISPRi), we observed a significant reduction in the expression levels of both *CAV1* and *CAV2* in iPSC-CMs (Fig. 7h and Supplementary Fig. 10). Altogether, these results demonstrate that rs3807989 is a regulatory variant that modulates the expression levels of *CAV1* and *CAV2* via differential protein binding, and as such, is highly likely the causal variant underlying the AF and PR interval GWAS signals in the *CAV1* interval.

Discussion

Our study shows that differential binding of NKX2-5 likely underlies the molecular mechanisms of numerous genetic associations with EKG traits across the genome. Additionally, we showed that molecular phenotype data from iPSC-CMs combined with fine-mapping statistical approaches can be used to prioritize putative causal variants underlying genetic associations with cardiac-specific traits. Furthermore, our study demonstrates the effectiveness of using iPSC-derived cells as a model system for understanding the genetic basis of complex human traits and diseases by conducting genome-wide genotype-phenotype analyses as well as interrogating the function of individual variants.

Within ~38,000 NKX2-5 binding sites, we identified 1,941 genetic variants that altered binding of the transcription factor. Because we investigated seven individuals in a three-generational family, the statistical power for identifying ASE-SNVs was increased as there were multiple replicates of allelic imbalance at the same heterozygous SNV. However, we anticipate that analyzing a larger sample size would identify a greater fraction of the NKX2-5 sites affected by genetic variants. For the NKX2-5 sites with differential binding, ~40% had genetic variants that altered the cognate TF motif and/or motifs of functionally related cardiac TFs, suggesting that a large fraction of the observed allelic binding of NKX2-5 was either a direct consequence of the SNV, or an indirect consequence resulting from the differential binding of a known co-factor. ASE-SNVs that were not associated with core cardiac TF motifs could: (i) affect consensus motifs from TFs that were not included in our targeted analysis; (ii) affect important sequences that impact DNA shape or an as of yet unknown regulatory mechanism[50–52]; or (iii) be non-functional. Combinatorial interactions

between key cardiac TFs is known to be an important mechanism for orchestrating the cardiac gene expression program during development[8–11]. While genetic variation has been shown to affect collaborative binding of lineage determining TFs in mice[53], our study is the first that we are aware of to show these effects in humans.

Coding mutations in and non-coding variants near *NKX2-5* have, respectively, been associated with congenital heart defects[12], as well as heart rate, atrial fibrillation and PR interval[13–15], implicating this TF in a range of cardiac disease in both development and adult stages. Here, our analysis of genome-wide NKX2-5 binding enabled us to investigate its role in cardiac phenotypes through a different genetic mechanism, i.e. variation in TF binding sites resulting in differential expression of target genes. We showed that differential NKX2-5 binding was positively correlated with H3K27ac peaks at iPSC-CM enhancers, but not iPSC enhancers, suggesting that NKX2-5 ASE-SNVs altered cardiac specific enhancer activity. These findings are consistent with the fact that we found enrichment for GTEx heart-specific eQTLs in both NKX2-5 and H3K27ac ASE-SNVs in iPSC-CMs. Importantly, out of all the molecular phenotypes examined, NKX2-5 ASE-SNVs were the more strongly enriched within EKG loci, thereby implicating NKX2-5 in the development of these traits, and indicating that NKX2-5 ASE-SNVs could be used to prioritize putative causal variants.

Analyzing GWASs for heart rate, atrial fibrillation and PR interval using a fine-mapping method that integrates functional annotations with GWAS summary statistics (fgwas) revealed several NKX2-5 ASE-SNVs with a high probability of causality at known loci as well as potentially novel sub-threshold GWAS signals. As a proof that this approach was effective to prioritize causal variants, one of the NKX2-5 ASE-SNVs (rs6801957 at the *SCN10A-SCN5A* locus) had been previously investigated in detail and had been shown to be functionally implicated in the association with EKG[3,45]. Further investigation of NKX2-5 ASE-SNVs heart rate loci using Hi-C generated from the same iPSC-CMs and gene expression in iPSC-CMs derived from 128 individuals revealed an association between the putative causal NKX2-5 ASE-SNVs and expression of nearby or distal candidate target genes. As a notable example, one of the prioritized variants (rs176107) at a sub-threshold locus showed long-range (~1.2 Mb) interaction with *MEF2C*, a key cardiac morphogenesis regulator, and was associated with its expression, thus providing a plausible mechanism underlying associations between differential NKX2-5 binding and heart rate.

We further followed up two NKX2-5 ASE-SNVs that were potential causal variants underlying associations with EKG traits with experimental validation including EMSA, luciferase assay, and CRISPRi. These analyses demonstrated that the two common SNPs, rs590041 (associated with P-wave duration) and rs3807989 (associated with PR interval and atrial fibrillation), are functional regulatory variants that influence the expression of *SSBP3* and *CAV1-CAV2* genes, respectively, via differential TF binding. Interestingly, while the rs3807989 stronger TF binding allele was associated with higher gene expression, the rs590041 stronger TF binding allele was associated with reduced gene expression, indicating that NKX2-5 binding is associated with both activating and repressing regulatory elements. Although future experimental studies are needed to elucidate the function of *SSBP3* and *CAV1-CAV2* with respect to the associated EKG phenotypes, our results provide novel

insights into the role differential binding of NKX2-5 and other cardiac TFs play in the genetic underpinnings of EKG traits.

Finally, our study demonstrates that analyzing the allelic binding of master developmental TFs in iPSC-CMs is highly effective to pinpoint genetic variation important for cardiac traits, and suggests that expanding this approach to study other cardiac TF (such as TBX5, GATA4, and MEF2C) in larger sample sizes could potentially identify and characterize many of the regulatory variants that play a role in cardiac traits and diseases.

Methods

Additional details are provided in the Supplementary Note and in the **Reporting Summary**.

Subjects and iPSC derivation

We selected seven individuals that are part of a three-generational family (three genetically unrelated subjects and two parent-offspring quartets) in the iPSCORE resource[28] (Supplementary Table 1). Fibroblasts from skin biopsies of each subject were reprogrammed using non-integrative Sendai virus[61] and analyzed for pluripotency as described in Panopulos et al.[28]. For five individuals, we analyzed one iPSC line ("clone"), and for two individuals we analyzed two iPSC lines (Fig. 1). The nine iPSC lines were harvested in multiple replicates between passages 12 to 40; a total of 35 different iPSC harvests were used in this study (Supplementary Table 2). This study was approved by the Institutional Review Boards of the University of California at San Diego (Project #110776ZF).

Differentiation of iPSCs into cardiomyocytes

The nine iPSCs were each differentiated multiple times using a monolayer protocol[62], resulting in a total of 26 iPSC-CM samples (Supplementary Table 2). Twelve of the iPSC-CM samples were subjected to selection using 4 mM sodium L-lactate media[63] and collected at day 25. Fourteen iPSC-CM samples were collected at day 15, of which one was subjected to lactate purification at day 11. At the day of collection, iPSC-CMs were dissociated using Accutase (Thermo Scientific), pooled, counted and separated into different aliquots. About 6 x 10^7 cells were fixed with formaldehyde and frozen for ChIP-seq. Cells (2 x 10^7) were lysed and stored in RLT plus buffer (Qiagen) for RNA extraction. Nuclei from 2 x 10^5 cells were frozen for ATAC-seq. Differentiation efficiency was measured by the percentage of cells that stained positive for the cardiac marker cardiac troponin T (TNNT2) (Thermo Scientific MA5-12960) using flow cytometry (FACSCanto system, BD Biosciences). The same protocols of dissociation and collection of samples for RNA-seq, ChIP-seq and ATAC-seq were applied to non-differentiated iPSC lines.

Whole-genome sequencing

Genomic DNA was whole genome sequenced as a part of the iPSCORE collection, as described by DeBoever et al.[21]. Briefly, reads were aligned against human genome b37 with decoy sequences[64] using BWA-MEM and default parameters[65]. The resulting BAM files were sorted using Sambamba[66] and duplicate reads were marked using biobambam2[67].

Variant calling was performed using the GATK best-practices pipeline[68,69] on BAM files separated into individual chromosomes.

RNA-seq

We generated and analyzed 56 RNA-seq (iPSCs: 29 independent samples; iPSC-CMs: 26 independent samples and 1 technical replicate). Total RNA was isolated using the Qiagen RNAeasy Mini Kit from frozen RTL plus pellets, and run on a Bioanalyzer (Agilent). Illumina Truseq Stranded mRNA libraries were prepared and sequenced on HiSeq2500, to an average of 40 million 100 bp paired-end reads per sample. RNA-seq reads were aligned using STAR[70] with a splice junction database built from the Gencode v19 gene annotation[71]. Gene-based expression values were quantified using the RSEM package[72] and normalized to transcript per million bp (TPM).

ChIP-seq

We generated and analyzed 48 ChIP-seq of histone modification H3K27ac (iPSCs: 17 samples and 4 technical replicates; iPSC-CMs: 25 samples and 2 technical replicates), and 15 ChIP-seq of NKX2-5 (iPSC-CMs: 12 samples and 3 technical replicates) (Supplementary Tables 2 and 3), using anti-H3K27ac (Abcam ab4729) and anti-NKX2-5 (Santa Cruz Biotechnology, sc-8697x) antibodies. Libraries were sequenced to an average of 35 million 100 bp paired-end reads per sample. ChIP-seq reads were mapped to the hg19 reference using BWA[65]. Duplicate reads, reads mapping to blacklisted regions and read-pairs with mapping quality Q < 30 were filtered. Peak calling was performed using MACS2[73] with reads derived from sonicated chromatin not subjected to IP (i.e. input chromatin) from a pool of samples used as negative control. For each data type, peak coordinates were called from combined BAM files across all samples of either iPSCs or iPSC-CMs. Quantification of the signal at peaks in each sample was performed using featureCounts[74]. Motif enrichment analysis was performed using HOMER[75] and, for NKX2-5, also using MEME ChIP[76].

ATAC-seq

We generated 37 ATAC-seq libraries (iPSCs: 12 samples and 5 technical replicates; iPSC-CMs: 11 samples and 9 technical replicates) using an adapted protocol from Buenrostro et al.[77]. Libraries were sequenced to an average depth of 20 million 100-150 bp paired end reads. ATAC-seq reads were aligned using STAR to hg19 and filtered using the same protocol as for ChIP-seq. In addition, to restrict analysis to regions spanning only one nucleosome, we required an insert size no larger than 140 bp. Peak calling was performed using MACS2 on combined BAM files of either iPSC or iPSC-CM samples.

Analysis of gene expression differences between iPSCs and iPSC-CMs

A matrix of raw gene expression values from 64 RNA-seq samples (29 iPSCs, 27 iPSC-CMs, and 8 RNA-seq samples from Roadmap including H1-hESC, HUES64, iPS-20b, iPS-18, Right Atrium, Right Ventricle, Left Ventricle, and Fetal Heart) was created from the RSEM expected counts, filtered for > 1 TPM on average samples, and rounded to integer values. After filtering, 15,725 genes remained from the initial 57,820. Expression values

were normalized using variance stabilizing transformation (*vst*) implemented in DESeq2[78]. Hierarchical clustering and the heatmap in Supplementary Figure 1 were generated using *vst*-normalized read counts for a panel of 61 selected genes using 'pheatmap' package in R. Analysis and plotting of principal components of all 15,725 genes were performed in R (Fig. 1).

To identify differentially expressed genes (DEGs) between iPSCs and iPSC-CMs, we used a matrix of raw expression counts from 56 RNA-seq (29 iPSCs and 27 iPSC-CMs), filtered for average TPM > 1 (22,447 genes), and applied DESeq2 with default settings to identify DEGs more than 2-fold and at a BH (Benjamini & Hochberg) FDR of 5%.

Normalization and analysis of variability of molecular phenotypes

For RNA-seq, we restricted the analysis to autosomal genes that had on average a minimum of 1 TPM per sample (14,933 and 15,167 genes for iPSCs and iPSC-CMs, respectively) and integer-rounded RSEM expected counts were used as expression levels. For ChIP-seq, we excluded peaks > 5 kb long and those located on sex chromosomes, resulting in 110,345 H3K27ac peaks analyzed in iPSCs and 83,689 H3K27ac peaks and 37,994 NKX2-5 peaks analyzed in iPSC-CMs (Supplementary Table 3). Matrices of raw expression levels or peak coverage for each of the 5 datasets were *vst*-normalized using DESeq2 and analyzed for principal components using R. To investigate the major sources of variability within each dataset, values for the first 10 PCs were correlated with known covariates across samples (for iPSCs: sequencing batch, passage and subject; for iPSC-CMs: TNNT2 expression, protocol of differentiation and subject; for ChIP-seq of both cell types, we also included the fraction of reads mapping to peaks, or FRiP) using ANOVA. We corrected the respective datasets by fitting a model including the covariates that were most associated with the first PC (batch for iPSCs; TNNT2 expression and protocol/batch for iPSC-CMs; and FrIP for all ChIP-seq datasets) using the 'lmFit' function from 'limma' package and calculating the residuals using the 'residuals' function in R. Mean expression and coverage values for each gene/peak were added back to the residuals. Residual-corrected values were used in all subsequent analyses.

To assess the consistency of data generated from cell lines derived from the same individual versus cell lines from different individuals, we selected the 1,000 most variable genes or peaks and computed matrices of Spearman correlation values across all pairs of samples for each molecular phenotype. We then separated correlation values between pairs of samples from the same, different, related or unrelated individuals and calculated the average correlation per sample. Technical replicates were excluded for the comparisons between samples of the same subject. We tested for significant increase in correlation between samples from the same subject using a one-tailed Mann-Whitney test (Fig. 1c-g and Supplementary Fig. 3k-o).

Allelic-specific effect (ASE) analysis

ASE analysis was performed as previously described[21]. To increase sensitivity of ASE and maximize the number of genes/peaks to analyze, reads from all samples from each individual per assay were merged. Heterozygous SNVs were identified by intersecting

variant calls from WGS with either exonic regions from Gencode v19 or regions identified by each ChIP-seq or ATAC-seq dataset. The WASP pipeline[79] was employed to reduce reference allele bias at heterozygous sites. The number of read pairs supporting each allele was counted using the ASEReadCounter from GATK[80]. Heterozygous SNVs were then filtered to keep SNVs where the reference or alternate allele had more than 8 supporting read pairs, the reference allele frequency was between 2-98%, the SNV was located in unique mappability regions according to wgEncodeCrgMapabilityAlign100mer track, and not located within 10 bp of another variant in a particular subject (heterozygous or homozygous alternative)[49,81,82]. ASE *P*-values for each SNV were calculated in each sample using a binomial test method[49,82]. To combine ASE results at each SNV across samples, we performed a meta-analysis on all samples that were heterozygous for a given SNV and for which ASE could be tested. The binomial *P*-values of heterozygous SNVs were combined using the Stouffer *z*-score method[83], using the formula $Z \sim \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}}$, where Z is the *z*-score derived from *P*-values and signed according to the direction of the effect, and k is the number of individuals for each SNV. The combined *z*-scores were transformed to *P*-values and a BH FDR was calculated using 'p.adjust' in R. The alternate allele frequency was averaged across all heterozygous samples.

Correlation of ASEs across all individuals

The direction of ASE effects across all family members (including homozygous individuals) was estimated using the ß coefficient of a linear model testing association between the corrected gene expression or peak coverage (normalized to *z*-scores across individuals) and the genotype of the seven family members (0, 1 or 2, testing only one ASE-SNV per region). Spearman correlation was used to compare ß to the average allele proportion of ASE-SNVs to estimate the consistency of effects (Fig. 2d-f).

Correlation of ASEs across different molecular phenotypes

To test if the direction of ASE of SNVs within ChIP-seq peaks correlated with changes in peak coverage of other ChIP-seq peaks or with gene expression, we performed a linear regression between the ASE-SNV genotypes and each phenotype. ChIP-seq peaks were paired with the closest gene or peak within 500 bp using 'bedtools closest'. Using linear regression, we tested the association between the individual genotypes (0, 1, 2, testing only one ASE-SNV per region) of the ASE-SNVs (FDR < 0.05) and either the corresponding corrected and *z*-score normalized peak coverage or gene expression or those of the closest feature. In both peak/gene and peak/peak pairs, Spearman correlation was calculated between the two slopes (ß) of linear regression (Fig. 2g,h).

Analysis of SNVs altering TFBS motifs

The effect of NKX2-5 ASE-SNVs on TFBS motifs was estimated using position probability matrices (PPMs) of the 12 most enriched families of motifs identified using HOMER (Supplementary Table 4), from a library of known motifs. For NKX2, GATA, TEAD, MEF2, TBX20 and PDX1, we also used PPMs derived from a de-novo analysis. All PPMs are provided in Supplementary Table 4. Position weight matrices (PWMs) were calculated from

the PPMs using a background nucleotide frequency of 0.25 for each base. Using a custom R script, a 40-bp window centered on each SNV tested for ASE was scanned with PWMs for each motif, and the position with the highest score was identified. For SNVs where either the reference or the alternate sequence matched or exceeded the log odds detection threshold reported by HOMER PPMs, the difference between the scores of the two alleles was calculated. In cases where a SNV matched multiple motifs from the same family, we kept only the motif with the highest score for either of the alleles. Fisher's exact test was used to calculate enrichment for motif-altering SNVs in variants with ASE compared to variants without ASE (Fig. 3a). For each of the 12 motifs, we also calculated Spearman correlation between the allelic imbalance proportion of the reference allele and the difference in motif score between the reference and the alternate allele (Fig. 3c,d and Supplementary Fig. 5). Motifs that were altered at NKX2-5 ASE-SNVs are indicated in Supplementary Table 5.

Enrichment of ASE-SNVs for known quantitative trait loci

To examine the enrichment of ASE-SNVs in known quantitative trait loci across different tissues, we obtained dsQTLs in LCLs from Degner et al.[32], eQTLs from iPSCs from DeBoever et al.[21], and eQTLs from HaploReg v4.1[33], which contained combined results from 13 different studies including GTEx v.6[82]. To identify tissue-specific eQTLs (Fig. 4d), the 44 tissues from GTEx were classified into 26 groups by merging similar tissues (adipose (n = 2), artery (n = 3), brain (n = 10), cell lines (n = 2), colon (n = 2), esophagus (n = 3), heart (n = 2), skin (n = 2), and the remaining 18 tissues were n = 1). A gene-eQTL combination was defined as tissue-specific if 50% or more of the significant associations were in a single tissue group. All SNVs tested for ASE in ChIP-seq datasets (H3K27ac in iPSCs and H3K27ac and NKX2-5 in iPSC-CMs) were intersected with these annotations, and enrichment between heterozygous SNVs with and without ASE was calculated using Fisher's exact test in R. In cases where multiple SNPs overlapped a peak, we counted only one SNP per peak. The complete Fisher's exact test statistics including P-values, odds ratios and number of SNVs analyzed are reported in Supplementary Table 5.

Enrichment of GWAS-SNPs in regulatory regions in iPSC-CMs

To calculate enrichment for GWAS-SNPs in ChIP-seq and ATAC-seq peaks, we extracted sets of SNPs associated with six EKG traits (heart rate, PR interval, QT interval, QRS duration, atrial fibrillation and P-wave duration) from the GWAS catalog[1] and 119 non-EKG traits that were associated with a similar number of SNPs. We used GREGOR[84] to test each of these 125 SNP sets for enrichment in ChIP-seq and ATAC-seq peaks from iPSCs and iPSC-CMs from this study as well as in peaks from cardiac tissues from Roadmap as a control (Fig. 5a-c and Supplementary Fig. 6). To calculate the enrichment for EKG GWAS-SNPs in NKX2-5 ASE-SNVs, we obtained the SNVs overlapping NKX2-5 peaks and associated with any of the six EKG traits. For the SNVs that could be tested for ASE, we calculated the proportion with and without ASE and tested their relative enrichment using Fisher's exact test (Fig. 5d).

Estimating GWAS enrichment in molecular phenotypes and prioritizing putative causal variants

To determine the enrichment of genetic variants influencing EKG traits within the different iPSC-CM molecular phenotypes and to identify putative causal variants and novel associations, we employed the fgwas framework, as described by Pickrell et al.[38]. We obtained summary statistics from the den Hoed et al.[15] heart-rate GWAS meta-analysis (2,516,407 SNPs analyzed) from LD hub (http://ldsc.broadinstitute.org/ldhub/), the Christophersen et al. atrial fibrillation meta-analysis[47] (11,779,664 SNPs) from the CVD portal (http://broadcvdi.org/), and the van Setten at al. PR-interval[17] GWAS (2,712,310 SNPs) as a collaboration with the authors. For each GWAS, we annotated each variant with the type of molecular phenotype it overlapped: peaks (ATAC-seq, H3K27ac, and NKX2-5 peaks) and/or ASE-SNVs (H3K27ac and NKX2-5), and applied a single annotation model followed by a joint model, where the association enrichment was quantified simultaneously for all five annotations. To prioritize causal variants, we used the enrichment estimates from the joint model as priors to estimate the probability for a variant to be causal (posterior probability of association, PPA) within consecutive 1-Mb windows across the genome. We report all variants with PPA > 0.3 in Supplementary Table 6.

Gene expression analysis of 128 iPSC-CMs

We used RNA-seq of iPSC-CMs from 128 different individuals[26]. Subjects included 43 males and 85 females, between 9 and 88 years of age, of diverse ethnicities (Europeans, n = 78, and Asians, n = 23). iPSCs were differentiated into day-25 cardiomyocytes using the method described above, including a 4 mM sodium L-lactate enrichment step at day 15, and yielded on average 83.9 +/- 13.6 % cTNT-positive populations. RNA-seq was generated and processed as described above. Raw gene expression data were first filtered for genes with TPM ≥ 2 in at least 5% of the samples and then quantile-normalized. From these values we calculated PEER factors[85] and used the residuals of the first 10 factors as normalized gene expression values. We extracted the individuals' genotypes from WGS and performed linear regression for the specific SNV-gene expression associations in R.

Electrophoretic mobility shift assay (EMSA)

EMSAs were performed using the LightShift™ Chemiluminescent EMSA Kit (Thermo Scientific) with biotinylated and non-biotinylated single-stranded oligonucleotides corresponding to 33-34 genomic fragments containing the SNPs rs590041, rs3807989, and rs7512445 (Supplementary Table 7). Both forward and reverse strand were tested; the forward strand was bound in case of rs590041 and rs3807989 and the reverse strand was bound in case of rs7512445. Nuclear extract from day-30-33 iPSC-CMs was extracted using the NE-PER Nuclear and Cytoplasmic Extraction Reagents (Thermo Scientific) with 1X Halt™ Protease Inhibitor Cocktail (Thermo Scientific). The binding reaction was carried in 10 μl volume containing 1 μl of 10X Binding Buffer (100 mM Tris pH 7.5, 500 mM KCl and 10 mM DTT), 2.5% glycerol, 5 mM $MgCl_2$, 0.05% NP40, 50 ng Poly(dI:dC), 1 pmol of biotin-labeled probe, and 15.3-16.8 μg nuclear extract. For competition experiments, a 200-fold molar excess of unlabeled probe was added. Binding reactions were incubated at room temperature for 20 min and loaded onto a 6% polyacrylamide 0.5X TBE gel. After sample

electrophoresis and transfer to a Biodine B Pre-Cut Modified Nylon Membrane, 0.45 μm (Thermo Scientific), DNA was UV-crosslinked for 15 min, and the biotinylated probes were detected using Chemiluminescent Nucleic Acid Detection Module (Thermo Scientific). Membranes were acquired using C-DiGit Blot scanner (LI-COR Biosciences).

Luciferase assay

Candidate functional variants rs590041 (*SSBP3* intron) and rs3807989 (*CAV1* intron) were tested for differential transcriptional activity using luciferase reporter assay. ~1.7-kb regions centered on each SNP were amplified from genomic DNA and cloned into pGL4.23 Firefly Luciferase reporter vectors (Promega) using Kpn I restriction sites, with primers given in Supplementary Table 7. For rs590041, the two allelic variants were obtained using site-directed mutagenesis of a homozygous alternate genomic DNA, while for rs3807989, they were obtained by sub-cloning DNA with heterozygous genotype. Cryopreserved day-25 iPSC-CMs were seeded onto a matrigel-coated 96-w plate at a density of 30-40 x 10^6cells per well and cultured in RPMI + Insulin for 5-10 days prior to transfection, when media was exchanged to Opti-MEM (Life Technologies). Each well was transfected with a mix of 120 ng of Firefly Luciferase reporter vector, 30 ng of Renilla Luciferase control vector (pRL-TK, Promega), and 0.6 μl of Viafect transfection reagent (Promega) in 10 μl of Opti-MEM. We transfected six wells per construct. Luciferase activity was measured 24 hours after transfection using the Dual-Luciferase® Reporter Assay System (Promega).

CRISPRi experiments

Two gRNAs targeting *CAV1* and *SSBP3* regulatory elements were designed using the online software CHOPCHOP (http://chopchop.cbu.uib.no/index.php) and cloned into the lentiviral vector pLKO.1-U6-2sgRNA-ccdB-EF1a-Puromycin. Lentiviral gRNAs or Lenti-dCas9-KRAB-blast plasmids (Addgene #89567) were co-transfected with packaging plasmids (psPAX2 and pMD2.G) into human 293T cells. Culture medium containing lentivirus particles for gRNA and dCas9-KRAB was harvested, mixed well with polybrene (10 μg/ml), and added to a 24-well plate. Day-30 iPSC-CMs (cell lines iPSCORE_1_5 and iPSCORE_75_1) were dissociated and added into the virus-containing media at around 80% confluence. For higher infection efficiency, a new collection of lentiviral particles mixed with polybrene was added to the medium after 24 hours. Medium was exchanged after 24 hours to regular culture medium and changed to selection medium containing 0.2 μg/ml puromycin and 6 μg/ml blasticidin after another 24 hours. Cells were cultured for 6 days when all cells from the non-infected control died, and then harvested. RNA was isolated with Quick-RNA kit (Zymo Research) and reverse-transcribed using SuperScript III Reverse Transcriptase (Life Technologies). qPCR reactions were performed in StepOne™ Real-Time PCR Systems (Applied Biosystems) using 2X Affymetrix qPCR master mix. Relative quantities of gene expression levels were normalized to the *METTL2B* gene. Guide RNAs and primers for qPCR are given in Supplementary Table 7.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported in part by a California Institute for Regenerative Medicine (CIRM) grant GC1R-06673-B, NIH grants HG008118-01 and HL107442-05, and National Science Foundation (NSF) grant no. 1728497. P.B. was supported by the Swiss National Science Foundation (SNSF) Postdoc Mobility fellowships P2LAP3-155105 and P300PA-167612. W.W.Y.G. was supported by the NHLBI of the NIH under award number F31HL142151. C.D. was supported in part by the UCSD Genetics Training Program through an institutional training grant from the National Institute of General Medical Sciences (T32GM008666) and the CIRM Interdisciplinary Stem Cell Training Program at UCSD II (TG2-01154). Library preparation and sequencing services were conducted by K. Jepsen and M. Khosroheidari at the UCSD IGM Genomics Center supported by NIH grant P30CA023100. N.S. was supported by NIH grants HL116747 and HL141989. K.J.G. was supported by the NIH grant R01DK114650 and the ADA grant 1-17-JDF-027. W.M., F.Y. and M.G.R. were supported by NIH grants DK018477 and DK039949; M.G.R. is an HHMI Investigator. We are very thankful to Chia-An Yen and N. Spann for assistance with ChIP-seq experiments and to A. Schmitt for Hi-C data. We thank the contribution of A. Aguirre in performing immunofluorescence. We thank E. Farley and K. Olson for help with reporter assays. We thank many colleagues for helpful comments.

References

1. MacArthur J, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017; 45:D896–D901. [PubMed: 27899670]
2. Gallagher MD, Chen-Plotkin AS. The post-GWAS era: from association to function. Am J Hum Genet. 2018; 102:717–730. [PubMed: 29727686]
3. van den Boogaard M, et al. A common genetic variant within SCN10A modulates cardiac SCN5A expression. J Clin Invest. 2014; 124:1844–1852. [PubMed: 24642470]
4. Wang X, et al. Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures. Elife. 2016; 5:e10557. [PubMed: 27162171]
5. Deplancke B, Alpern D, Gardeux V. The genetics of transcription factor DNA binding variation. Cell. 2016; 166:538–554. [PubMed: 27471964]
6. Pai AA, Pritchard JK, Gilad Y. The genetic and mechanistic basis for variation in gene regulation. PLoS Genet. 2015; 11:e1004857. [PubMed: 25569255]
7. Maurano MT, et al. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. Nat Genet. 2015; 47:1393–401. [PubMed: 26502339]
8. He A, Kong SW, Ma Q, Pu WT. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. Proc Natl Acad Sci USA. 2011; 108:5632–5637. [PubMed: 21415370]
9. Schlesinger J, et al. The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs. PLoS Genet. 2011; 7:e1001313. [PubMed: 21379568]
10. Luna-Zurita L, et al. Complex interdependence regulates heterotypic transcription factor distribution and coordinates cardiogenesis. Cell. 2016; 164:999–1014. [PubMed: 26875865]
11. Ang YS, et al. Disease model of GATA4 mutation reveals transcription factor cooperativity in human cardiogenesis. Cell. 2016; 167:1734–1749.e22. [PubMed: 27984724]
12. Kathiresan S, Srivastava D. Genetics of human cardiovascular disease. Cell. 2012; 148:1242–1257. [PubMed: 22424232]
13. Pfeufer A, et al. Genome-wide association study of PR interval. Nat Genet. 2010; 42:153–159. [PubMed: 20062060]
14. Verweij N, et al. Genetic determinants of P wave duration and PR segment. Circ Cardiovasc Genet. 2014; 7:475–481. [PubMed: 24850809]
15. den Hoed M, et al. Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. Nat Genet. 2013; 45:621–631. [PubMed: 23583979]
16. Nielsen JB, et al. Genome-wide study of atrial fibrillation identifies seven risk loci and highlights biological pathways and regulatory elements involved in cardiac development. Am J Hum Genet. 2018; 102:103–115. [PubMed: 29290336]
17. van Setten J, et al. PR interval genome-wide association meta-analysis identifies 50 loci associated with atrial and atrioventricular electrical activity. Nat Commun. 2018; 9

18. Panopoulos AD, et al. Aberrant DNA methylation in human iPSCs associates with MYC-binding motifs in a clone-specific manner independent of genetics. Cell Stem Cell. 2017; 20:505–517.e6. [PubMed: 28388429]
19. Carcamo-Orive I, et al. Analysis of transcriptional variability in a large human iPSC library reveals genetic and non-genetic determinants of heterogeneity. Cell Stem Cell. 2017; 20:518–532.e9. [PubMed: 28017796]
20. Kilpinen H, et al. Common genetic variation drives molecular heterogeneity in human iPSCs. Nature. 2017; 546:370–375. [PubMed: 28489815]
21. DeBoever C, et al. Large-scale profiling reveals the influence of genetic variation on gene expression in human induced pluripotent stem cells. Cell Stem Cell. 2017; 20:533–546.e7. [PubMed: 28388430]
22. Banovich NE, et al. Impact of regulatory variation across human iPSCs and differentiated cells. Genome Res. 2018; 28:122–131. [PubMed: 29208628]
23. Pashos EE, et al. Large, diverse population cohorts of hiPSCs and derived hepatocyte-like cells reveal functional genetic variation at blood lipid-associated loci. Cell Stem Cell. 2017; 20:558–570.e10. [PubMed: 28388432]
24. Schwartzentruber J, et al. Molecular and functional variation in iPSC-derived sensory neurons. Nat Genet. 2018; 50:54–61. [PubMed: 29229984]
25. He JQ, Ma Y, Lee Y, Thomson JA, Kamp TJ. Human embryonic stem cells develop into multiple types of cardiac myocytes: action potential characterization. Circ Res. 2003; 93:32–39. [PubMed: 12791707]
26. D'Antonio-Chronowska A, et al. Human iPSC gene signatures and X chromosome dosage impact response to WNT inhibition and cardiac differentiation fate. bioRxiv. 2019; doi: 10.1101/644633
27. Burridge PW, et al. Chemically defined generation of human cardiomyocytes. Nat Methods. 2014; 11:855–860. [PubMed: 24930130]
28. Panopoulos AD, et al. iPSCORE: a resource of 222 iPSC lines enabling functional characterization of genetic variation across a variety of cell types. Stem Cell Reports. 2017; 8:1086–1100. [PubMed: 28410642]
29. Kilpinen H, et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science. 2013; 342:744–747. [PubMed: 24136355]
30. Dupays L, et al. Sequential binding of MEIS1 and NKX2-5 on the Popdc2 gene: a mechanism for spatiotemporal regulation of enhancers during cardiogenesis. Cell Rep. 2015; 13:183–195. [PubMed: 26411676]
31. Prall OW, et al. An Nkx2-5/Bmp2/Smad1 negative feedback loop controls heart progenitor specification and proliferation. Cell. 2007; 128:947–959. [PubMed: 17350578]
32. Degner JF, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. 2012; 482:390–394. [PubMed: 22307276]
33. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012; 40:D930–D934. [PubMed: 22064851]
34. GTEx Consortium. et al. Genetic effects on gene expression across human tissues. Nature. 2017; 550:204–213. [PubMed: 29022597]
35. Roselli C, et al. Multi-ethnic genome-wide association study for atrial fibrillation. Nat Genet. 2018; 50:1225–1233. [PubMed: 29892015]
36. Nielsen JB, et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. Nat Genet. 2018; 50:1234–1239. [PubMed: 30061737]
37. Christophersen IE, et al. Fifteen genetic loci associated with the electrocardiographic P wave. Circ Cardiovasc Genet. 2017; 10:e001667. [PubMed: 28794112]
38. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet. 2014; 94:559–573. [PubMed: 24702953]
39. Eppinga RN, et al. Identification of genomic loci associated with resting heart rate and shared genetic predictors with all-cause mortality. Nat Genet. 2016; 48:1557–1563. [PubMed: 27798624]

40. Butler AM, et al. Novel loci associated with PR interval in a genome-wide association study of 10 African American cohorts. Circ Cardiovasc Genet. 2012; 5:639–646. [PubMed: 23139255]
41. Sano M, et al. Genome-wide association study of electrocardiographic parameters identifies a new association for PR interval and confirms previously reported associations. Hum Mol Genet. 2014; 23:6668–6676. [PubMed: 25055868]
42. Arking DE, et al. Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. Nat Genet. 2014; 46:826–836. [PubMed: 24952745]
43. Holm H, et al. Several common variants modulate heart rate, PR interval and QRS duration. Nat Genet. 2010; 42:117–122. [PubMed: 20062063]
44. Ritchie MD, et al. Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. Circulation. 2013; 127:1377–1385. [PubMed: 23463857]
45. van den Boogaard M, et al. Genetic variation in T-box binding element functionally affects SCN5A/SCN10A enhancer. J Clin Invest. 2012; 122:2519–2530. [PubMed: 22706305]
46. Greenwald WW, et al. Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. Nat Commun. 2019; 10
47. Christophersen IE, et al. Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. Nat Genet. 2017; 49:946–952. [PubMed: 28416818]
48. Ramasamy A, et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat Neurosci. 2014; 17:1418–1428. [PubMed: 25174004]
49. Lappalainen T, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501:506–511. [PubMed: 24037378]
50. Samee MAH, Bruneau BG, Pollard KS. A de novo shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs. Cell Syst. 2019; 8:27–42.e6. [PubMed: 30660610]
51. Afek A, Schipper JL, Horton J, Gordan R, Lukatsky DB. Protein-DNA binding in the absence of specific base-pair recognition. Proc Natl Acad Sci USA. 2014; 111:17140–17145. [PubMed: 25313048]
52. Slattery M, et al. Absence of a simple code: how transcription factors read the genome. Trends Biochem Sci. 2014; 39:381–399. [PubMed: 25129887]
53. Heinz S, et al. Effect of natural genetic variation on enhancer selection and function. Nature. 2013; 503:487–492. [PubMed: 24121437]
54. Johnson AD, et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008; 24:2938–2939. [PubMed: 18974171]
55. Hong KW, et al. Identification of three novel genetic variations associated with electrocardiographic traits (QRS duration and PR interval) in East Asians. Hum Mol Genet. 2014; 23:6659–6667. [PubMed: 25035420]
56. van der Harst P, et al. 52 genetic loci influencing myocardial mass. J Am Coll Cardiol. 2016; 68:1435–1448. [PubMed: 27659466]
57. Evans DS, et al. Fine-mapping, novel loci identification, and SNP association transferability in a genome-wide association study of QRS duration in African Americans. Hum Mol Genet. 2016; 25:4350–4368. [PubMed: 27577874]
58. Ellinor PT, et al. Meta-analysis identifies six new susceptibility loci for atrial fibrillation. Nat Genet. 2012; 44:670–675. [PubMed: 22544366]
59. Jeff JM, et al. Generalization of variants identified by genome-wide association studies for electrocardiographic traits in African Americans. Ann Hum Genet. 2013; 77:321–332. [PubMed: 23534349]
60. Bezzina CR, et al. Common variants at *SCN5A-SCN10A* and *HEY2* are associated with Brugada syndrome, a rare disease with high risk of sudden cardiac death. Nat Genet. 2013; 45:1044–1049. [PubMed: 23872634]
61. Ban H, et al. Efficient generation of transgene-free human induced pluripotent stem cells (iPSCs) by temperature-sensitive Sendai virus vectors. Proc Natl Acad Sci USA. 2011; 108:14234–14239. [PubMed: 21821793]

62. Lian X, et al. Directed cardiomyocyte differentiation from human pluripotent stem cells by modulating Wnt/beta-catenin signaling under fully defined conditions. Nat Protoc. 2013; 8:162–75. [PubMed: 23257984]
63. Tohyama S, et al. Distinct metabolic flow enables large-scale purification of mouse and human pluripotent stem cell-derived cardiomyocytes. Cell Stem Cell. 2013; 12:127–137. [PubMed: 23168164]
64. Auton A, et al. A global reference for human genetic variation. Nature. 2015; 526:68–74. [PubMed: 26432245]
65. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
66. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015; 31:2032–2034. [PubMed: 25697820]
67. Tischler G, Leonard S. biobambam: tools for read pair collation based algorithms on BAM files. Source Code for Biology and Medicine. 2014; 9
68. DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43:491–498. [PubMed: 21478889]
69. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–1303. [PubMed: 20644199]
70. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29:15–21. [PubMed: 23104886]
71. Harrow J, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012; 22:1760–1774. [PubMed: 22955987]
72. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011; 12:323. [PubMed: 21816040]
73. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9:R137. [PubMed: 18798982]
74. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014; 30:923–930. [PubMed: 24227677]
75. Heinz S, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010; 38:576–589. [PubMed: 20513432]
76. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics. 2011; 27:1696–1697. [PubMed: 21486936]
77. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013; 10:1213–1218. [PubMed: 24097267]
78. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15:550. [PubMed: 25516281]
79. van de Geijn B, McVicker G, Gilad Y, Pritchard JK. WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods. 2015; 12:1061–1063. [PubMed: 26366987]
80. Van der Auwera GA, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013; 43:11.10.1–33. [PubMed: 25431634]
81. Mayba O, et al. MBASED: allele-specific expression detection in cancer tissues and cell lines. Genome Biol. 2014; 15:405. [PubMed: 25315065]
82. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015; 348:648–660. [PubMed: 25954001]
83. Whitlock MC. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J Evol Biol. 2005; 18:1368–1373. [PubMed: 16135132]
84. Schmidt EM, et al. GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. Bioinformatics. 2015; 31:2601–2606. [PubMed: 25886982]

85. Stegle O, Parts L, Durbin R, Winn J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol. 2010; 6:e1000770. [PubMed: 20463871]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Editorial summary

Analysis of iPSC-derived cardiomyocytes identifies variants associated with allele-specific effects on NKX2-5 binding. Fine-mapping and functional studies suggest that such variants underlie cardiac-specific expression quantitative trait loci and associations with electrocardiographic traits.

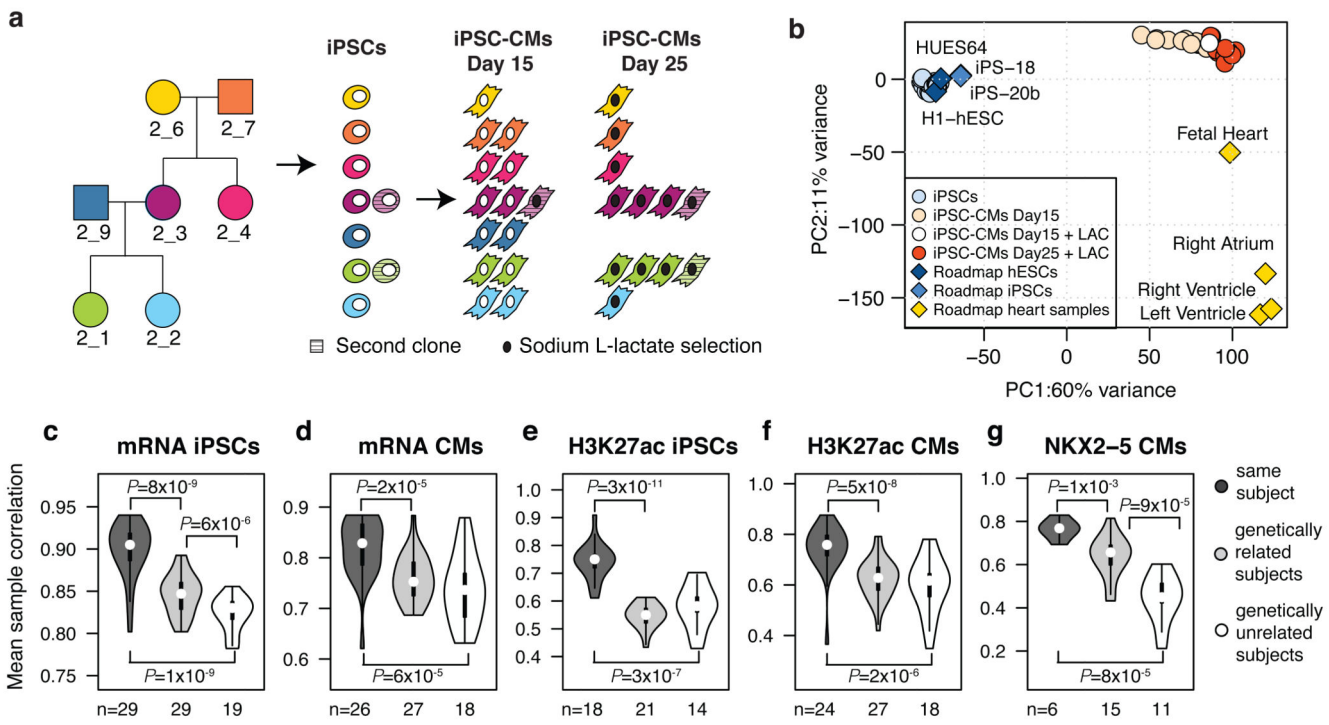

Figure 1. Generation and characterization of iPSCs and iPSC-CMs by gene expression and epigenetic profiling.

a, Pedigree showing the relationships of the seven individuals and summary of derived cell types analyzed. **b**, Principal component 1 and 2 of RNA-seq (15,725 genes) from iPSCs (29 samples from 7 individuals), iPSC-CMs (27 samples from 7 individuals), Roadmap stem cell lines (H1, HUES64, iPS-20b and iPS18), and human tissues (right ventricle, left ventricle, right atrium, and fetal heart). **c**-**g**, Distributions of the average Spearman correlation coefficients between pairs of samples across the 1,000 most variable genes (**c**,**d**) or peaks (**e**-**g**) for the indicated molecular phenotypes. Median (white dot), interquartile range (thick bar), and the rest of the distributions (line) are shown within each violin plot, with each sample size reported below. *P*-values of significant ($P < 0.05$) one-tailed Mann-Whitney tests are shown.

Europe PMC Funders Author Manuscripts

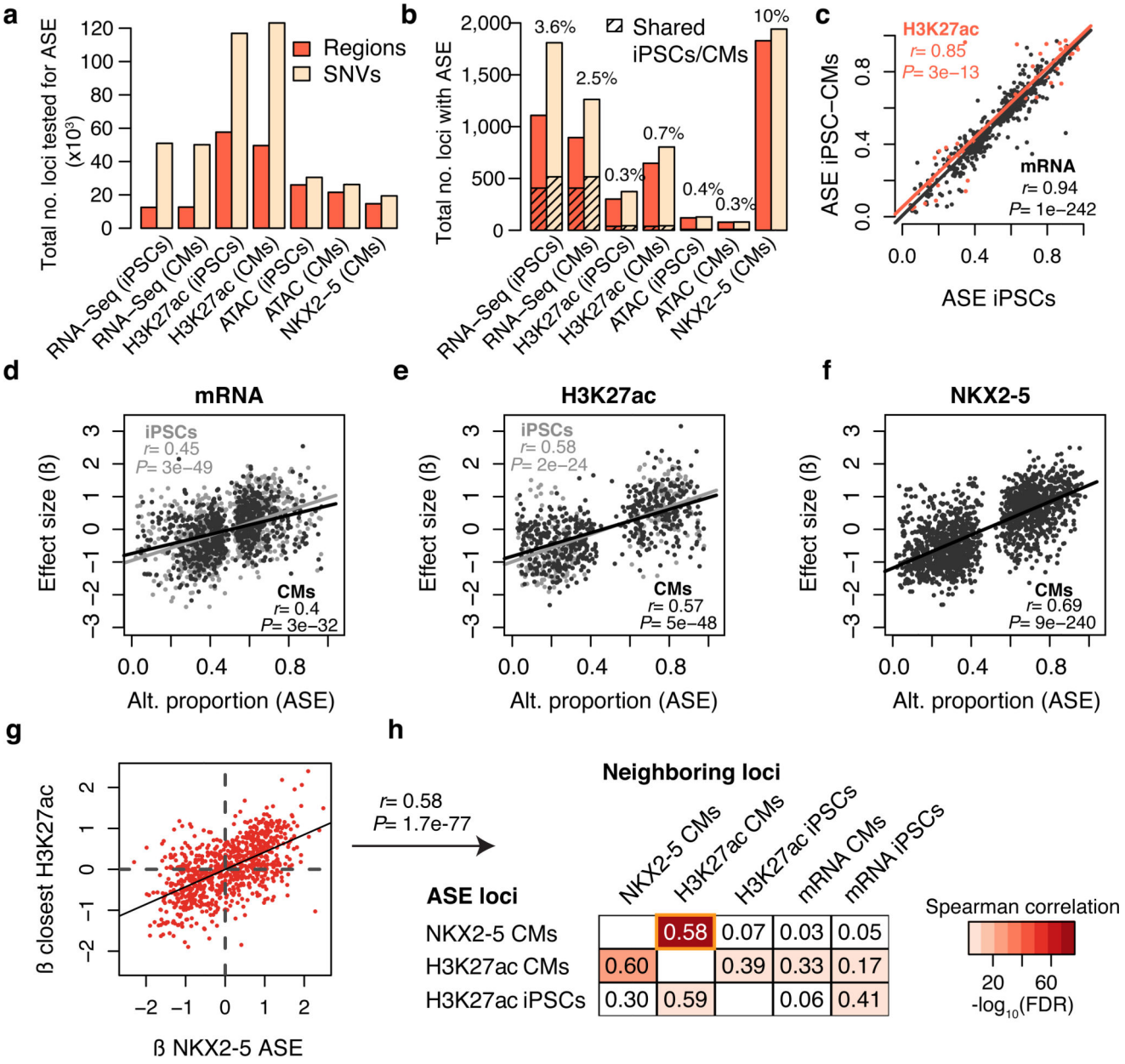

Figure 2. Identification of coordinated allele-specific effects (ASE) in gene expression, H3K27 acetylation, chromatin accessibility, and NKX2-5 binding in iPSCs and iPSC-CMs.

a, Total number of regions and heterozygous SNVs tested for ASE across all individuals and samples in each dataset. **b**, Total number of heterozygous SNVs and corresponding regions across all individuals and samples with ASE at FDR < 0.05. The number of ASE shared between iPSCs and iPSC-CMs is indicated by hatches. **c**, Scatterplot of the alternate allele proportion at shared ASE-SNVs between iPSCs and iPSC-CMs for RNA-seq (n = 516 SNVs) and H3K27ac (n = 43 SNVs). Spearman correlation statistics are indicated. **d-f**, Scatterplots of the mean proportion of the alternate allele of SNVs with ASE in heterozygous individuals and the effect size of each ASE-SNV, expressed as the slope of

linear regression (ß) between gene expression or peak density and genotypes of all seven individuals. Spearman correlation statistics are indicated. The number of SNVs analyzed in **d** are: 970 for iPSCs and 799 for iPSC-CMs; in **e**: 255 for iPSCs and 550 for iPSC-CMs; in **f**: 1,714. **g**, Scatterplot showing relationship between effect sizes (ß's) of ASE-SNVs in NKX2-5 peaks on both NKX2-5 and H3K27ac phenotypes (n = 854 SNPs). **h**, Table showing Spearman correlation coefficients of effect sizes between pairs of different molecular phenotypes. Correlations were calculated between ß's of SNVs that showed ASE in ChIP-seq datasets (rows) and ß's of the same variant for the closest gene or peak in a different molecular phenotype dataset (columns).

Europe PMC Funders Author Manuscripts

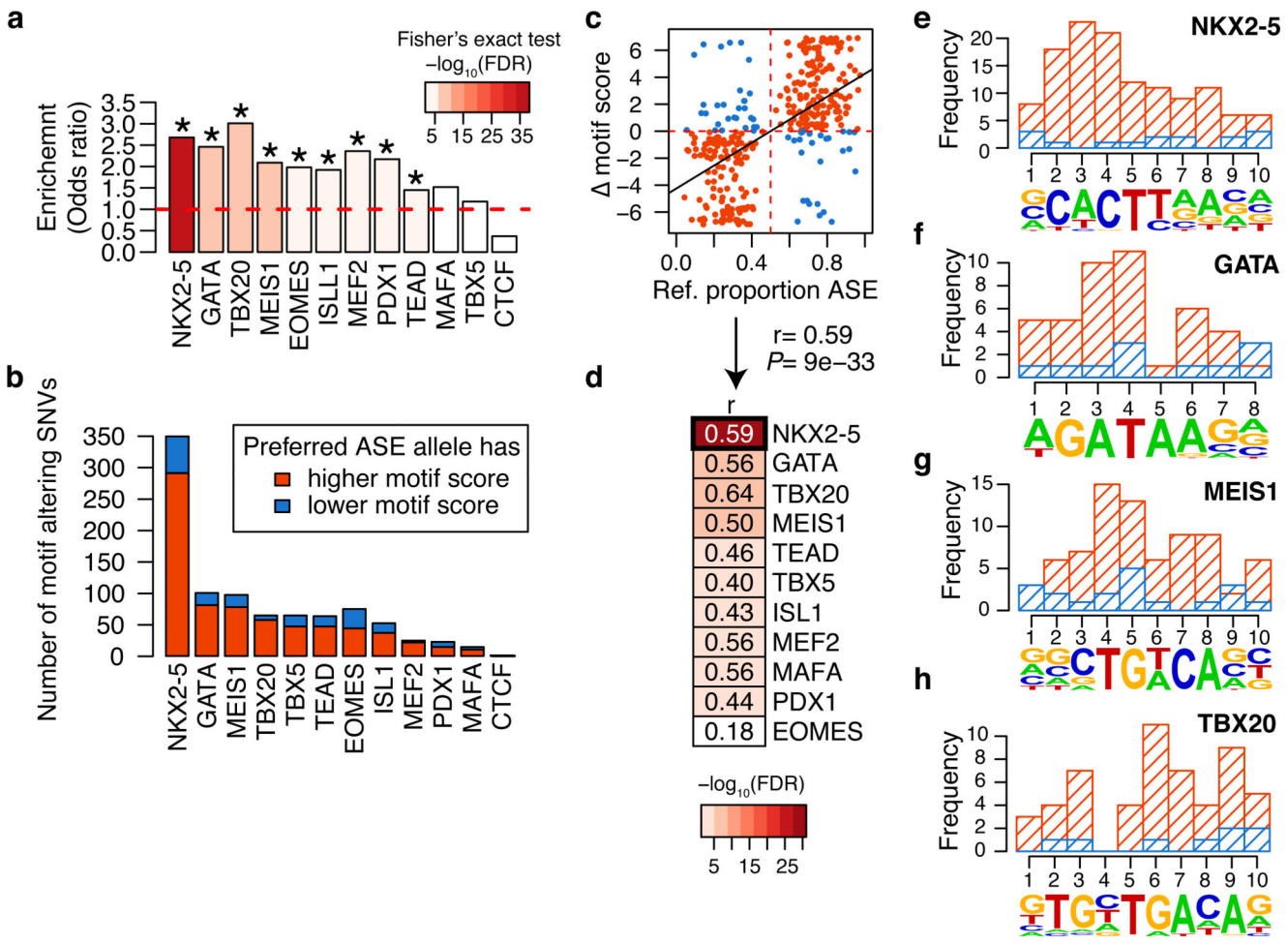

Figure 3. Transcription factor binding motifs are altered by SNVs with ASE in NKX2-5 ChIP-seq.

a, Odds ratios from two-sided Fisher's exact test comparing the proportion of motif-altering SNVs between variants with ASE (n = 1,941) and variants without ASE (n = 19,371) in NKX2-5 ChIP-seq peaks from combined iPSC-CM samples. Asterisks indicate enrichment at FDR corrected P-value < 0.05. **b**, Number of TFBS motifs that were strengthened (red) or weakened (blue) by the preferred allele of ASE-SNVs identified in NKX2-5 ChIP-seq. **c**, Scatterplot of the reference allele proportion at ASE-SNVs (n = 341) and the difference of NKX2-5 motif score between reference and alternate alleles. Spearman correlation coefficient and P-value are indicated at the bottom. Dots are color-coded as in **b**. **d**, Summary table of Spearman correlation statistics calculated as in **c** for all motifs tested (see Supplementary Fig. 4 for the other scatterplots). **e-h**, Frequency of ASE-SNVs altering different positions within the motifs of NKX2-5 (**e**), GATA (**f**), MEIS1 (**g**), and TBX20 (**h**). NKX2-5, GATA and TBX20 PWMs were obtained using de-novo motif finding. Bars are color-coded as in **b**. Blue bars overlap the red ones (i.e. they are not stacked).

Europe PMC Funders Author Manuscripts

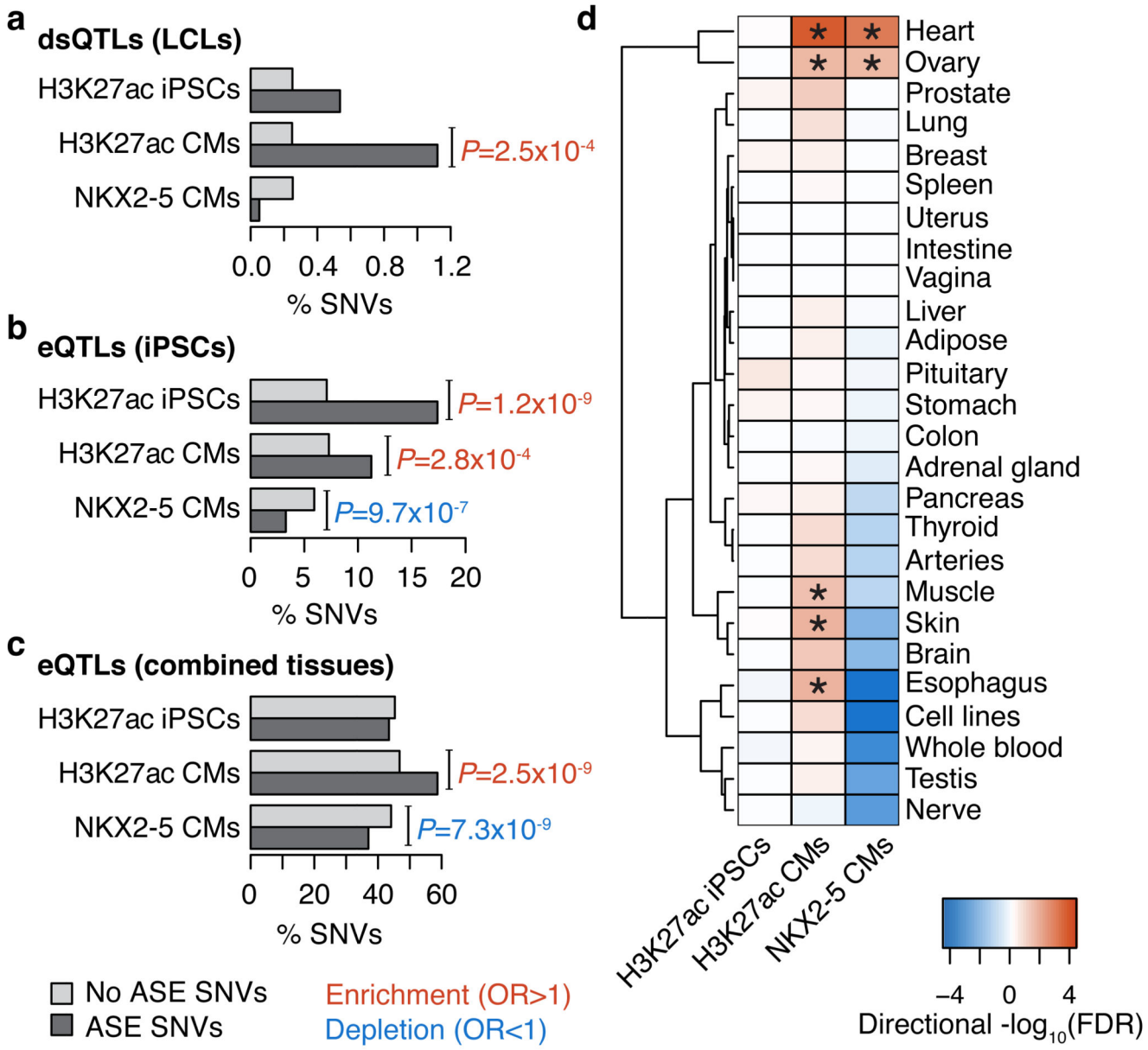

Figure 4. Enrichment of ChIP-seq ASE variants for known QTLs.

a-c, Histograms showing the percentage of SNVs with and without ASE in each ChIP-seq (from combined iPSC or iPSC-CM samples) and overlapping dsQTLs from LCLs[32] (**a**), eQTLs from iPSCs[21] (**b**), and combined eQTLs identified in different tissues[33] (**c**). Two-sided Fisher's exact test *P*-values are shown in red or blue for enrichment or depletion, respectively. **d**, Heatmap showing enrichment of ASE variants for tissue-specific eQTLs[34] (similar tissues in GTEx were merged; see Methods). Asterisks indicate two-sided Fisher's exact test FDR corrected *P*-value < 0.05. Heatmap is colored based on -$\log_{10}$ of FDR corrected *P*-values, with negative sign if odds ratio was < 1. The complete Fisher's exact test statistics including *P*-values, odds ratios and number of SNVs analyzed are reported in **Supplementary Table 5**.

Europe PMC Funders Author Manuscripts

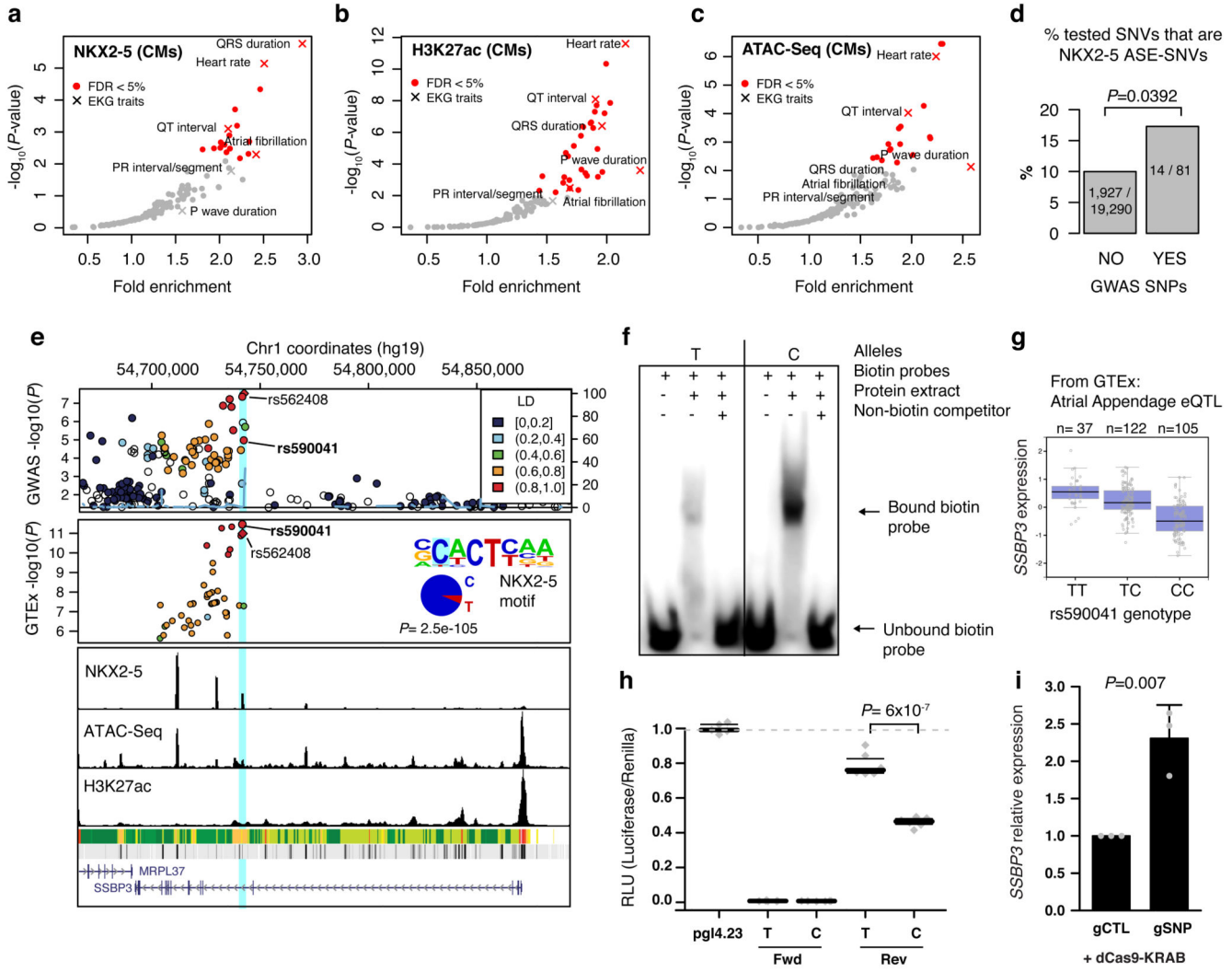

Figure 5. Enrichment of NKX2-5 SNVs at GWAS loci and validation of rs590041 as a regulatory variant in the *SSBP3* locus for p-wave duration.

a-c, Volcano plots showing -$\log_{10}$ *P*-values and fold enrichment for GWAS loci in NKX2-5 (**a**), H3K27ac (**b**), and ATAC-seq (**c**) peaks from combined iPSC-CM samples. Red symbols indicate significant enrichment at FDR corrected *P*-value < 0.05, calculated using GREGOR. In total $n = 125$ GWAS traits were tested, of which 6 were for EKG traits. **d**, Percentage of NKX2-5 ASE-SNVs overlapping an EKG GWAS-SNP versus overlapping a non-GWAS-SNP. Two-sided Fisher's exact test *P*-value and the number of SNVs are given. **e**, Top panel: regional plot of association *P*-values with P-wave duration[37], color coded based on r^2 values[54]. Second panel: regional plot of eQTLs for *SSBP3* in atrial appendage samples from GTEx. NKX2-5 allelic imbalance (pie chart) for rs590041 is shown. Panels three through five: epigenetic tracks from iPSC-CM combined samples. Bottom panel: UCSC genome browser tracks for Roadmap fetal heart ChromHMM, DHS, and gene annotations. **f**, EMSA with nuclear extract from iPSC-CM using probes containing two allelic variants of rs590041. Similar results were obtained in two independent experiments. The full scans of the blots are shown in Supplementary Figure 9. **g**, Screenshot from the GTEx portal (https://

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

gtexportal.org) showing association between rs590041 genotypes and expression levels of *SSBP3* in heart atrial appendage samples. **h**, Luciferase assay in iPSC-CMs for rs590041, in both forward and reverse orientations. RLUs are normalized to cells transfected with the empty vector (pgl4.23). Lines indicate median, lower and upper quartiles of 6 transfection replicates per plasmid. *P*-values from two-tailed *t*-tests are shown, comparing expression from the two alleles. **i**, qPCR expression of *SSBP3* in iPSC-CMs (id: iPSCORE_1_5) stably expressing dCas9-KRAB (CRISPRi) and either a control guide RNA (gCTL) or two guide RNAs targeting the region encompassing rs590041. Bars and error bars represent the mean and the standard deviation from three qPCR measurements, respectively; two-tailed *t*-test *P*-value is shown. Similar results were obtained in an independent cell line (Supplementary Fig. 10). All iPSC-CMs used in **f**, **h**, and **i** were lactate-purified.

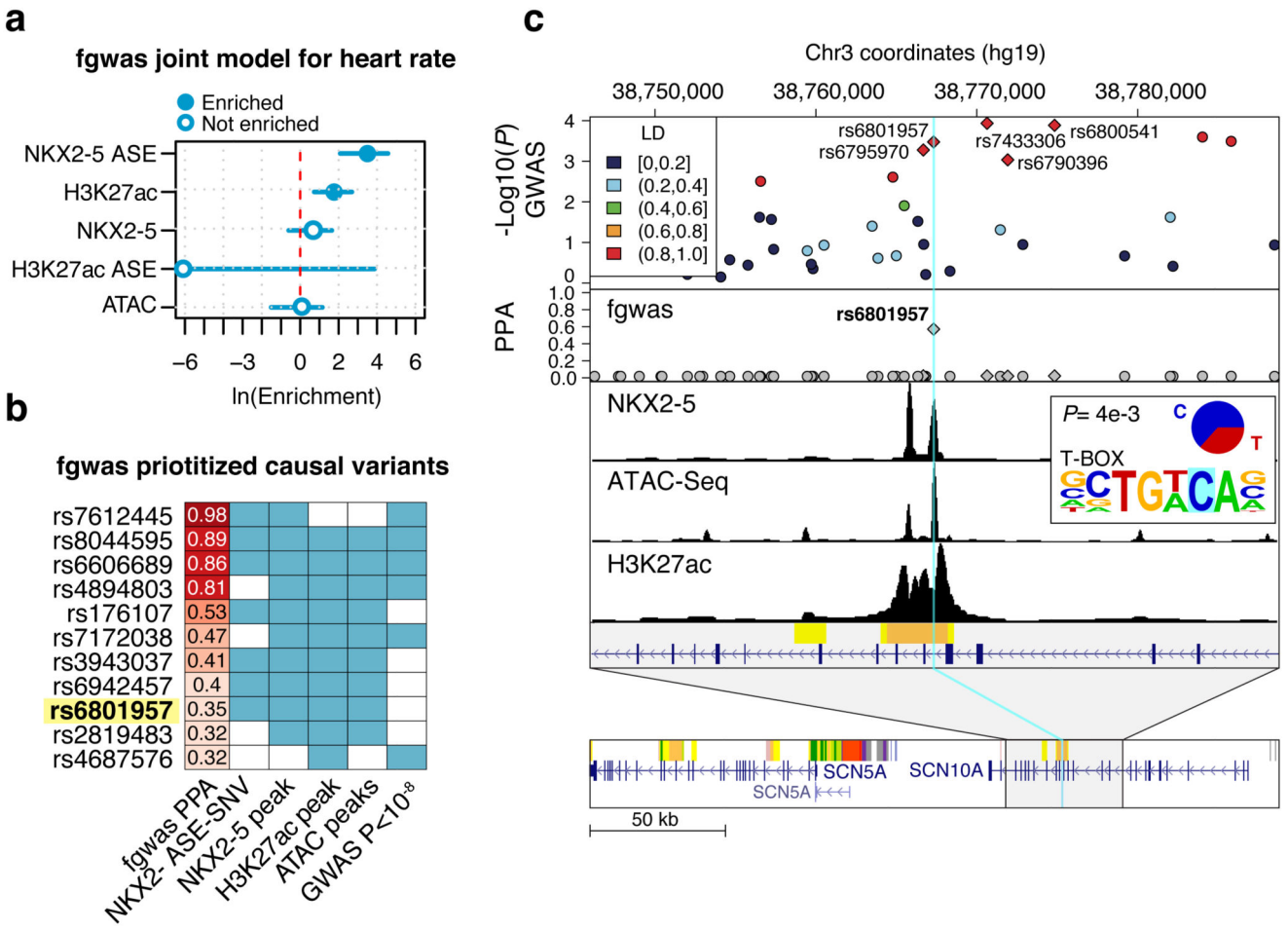

Figure 6. Prioritization of candidate causal variants at heart rate loci using fgwas.

a, fgwas natural log fold enrichment of GWAS-SNPs for heart rate[15] in iPSC-CM genomic annotations (*y*-axis). The bars indicate 95% confidence intervals. **b**, Table showing 11 SNPs with > 0.3 fgwas posterior probability of causality (PPA) and that overlapped at least two of the indicated iPSC-CM genomic annotations. SNPs that showed genome-wide significance ($P < 10^{-8}$) for each trait in the corresponding studies are indicated, while those with $P > 10^{-8}$ are sub-threshold, and thus novel, GWAS loci. **c**, Functional annotation of rs6801957 associated with heart rate[15]. Top panel: regional plot of association *P*-values; SNPs are color coded based on r^2 values from the 1000 Genome Project CEU population[54]; lead GWAS variants from other studies in the locus are indicated by a diamond. Second panel: fgwas PPA of the variants in the locus. Panels three through five: epigenetic tracks from iPSC-CM combined samples. Bottom panels: Roadmap fetal heart ChromHMM and genes from UCSC genome browser. Inner panel: allelic imbalance (pie chart) of NKX2-5 ASE with FRD-corrected *P*-values, and altered TF motif.

Europe PMC Funders Author Manuscripts

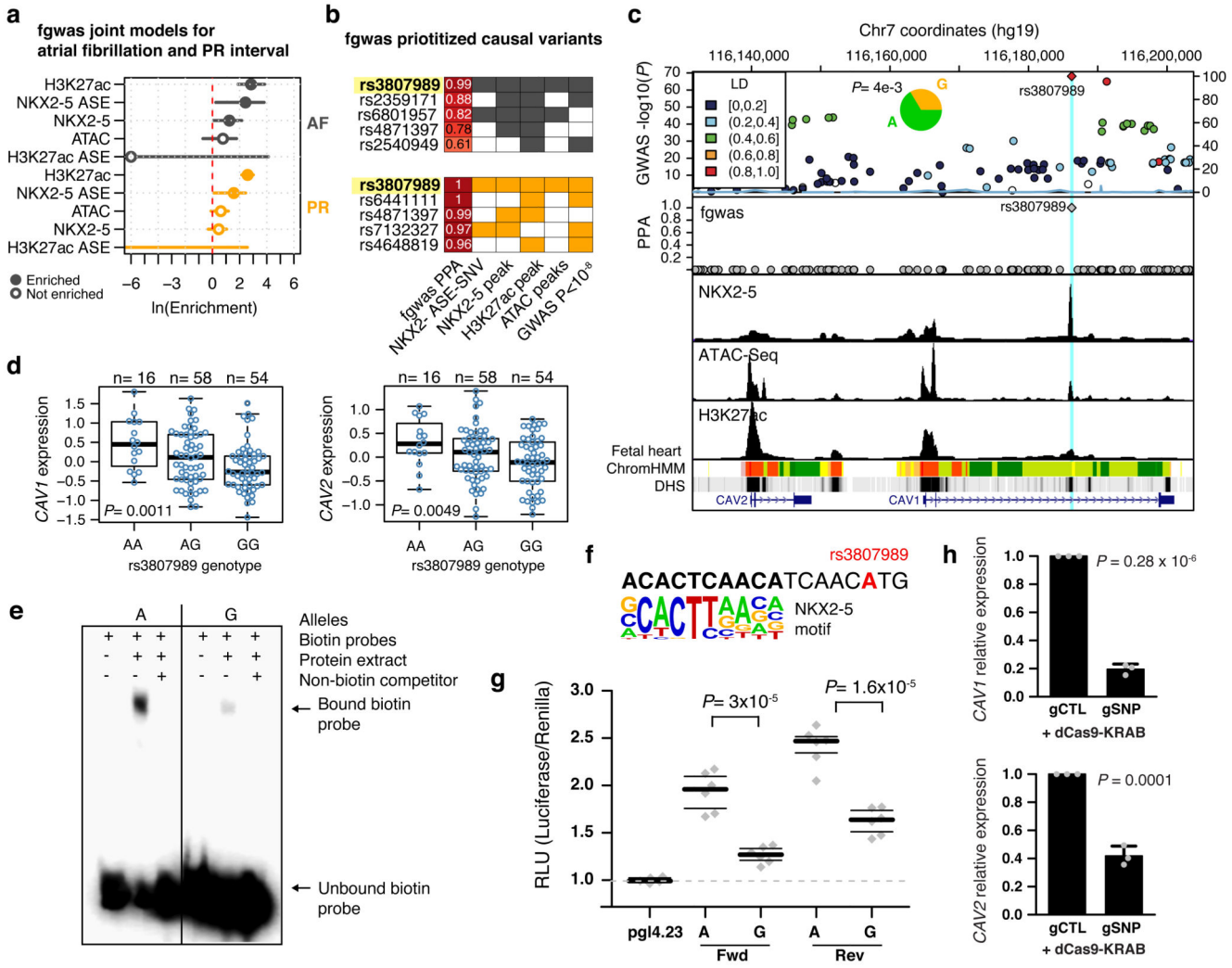

Figure 7. Functional characterization of rs3807989 as candidate causal variants for PR interval and atrial fibrillation.

a, fgwas natural log fold enrichment of GWAS-SNPs for atrial fibrillation (AF) and PR interval (PR) in iPSC-CM genomic annotations (*y*-axis). Bars indicate 95% confidence intervals. **b**, Tables showing the top 5 SNPs ordered by fgwas posterior probability of causality (PPA) and overlapping at least two of the indicated iPSC-CM genomic annotations. **c**, Top panel: regional plot of association *P*-values with PR interval[17], color coded based on r^2 values[54]. NKX2-5 allelic imbalance (pie chart) for rs3807989 is shown. Second panel: fgwas PPA of the variants in the locus. Panels three through five: epigenetic tracks from iPSC-CM combined samples. Bottom panel: UCSC genome browser tracks for Roadmap fetal heart ChromHMM, DHS, and gene annotations. **d**, Association between rs3807989 genotypes and gene expression of *CAV1* and *CAV2* genes in 128 iPSC-CMs from different individuals[26]. Boxplot elements: median (thick line), lower and upper quartiles (box), maximum and minimum (whiskers). *P*-value of linear regression is shown. **e**, EMSA with iPSC-CMs nuclear extract using probes containing two allelic variants of rs3807989. A second blot from an independent experiment with similar results and full scans of the blots

Europe PMC Funders Author Manuscripts

are shown in Supplementary Figure 9. **f**, Position of rs3807989 with respect to the NKX2-5 motif. **g**, Luciferase assays in iPSC-CMs for rs3807989, in both forward and reverse orientations. RLUs are normalized to cells transfected with the empty vector (pGL4.23). Plot lines indicate median, lower and upper quartiles of 6 transfection replicates per plasmid. *P*-values from two-tailed *t*-tests are shown. **h**, qPCR expression of *CAV1* and *CAV2* genes in iPSC-CMs stably expressing dCas9-KRAB (CRISPRi) (id: iPSCORE_1_5) and either a control guide RNA (gCTL) or two guide RNAs targeting the region encompassing rs3807989. Bars and error bars represent the mean and the standard deviation from three qPCR measurements, respectively; two-tailed t-test *P*-value is shown. The result was replicated in an independent cell line (Supplementary Fig. 10). All iPSC-CMs used in **d-h** were lactate-purified.

Table 1
Allelic binding of NKX2-5 at GWAS loci for EKG traits

dbSNP ID	ASE FDR	ASE reference allele ratio	Gene locus	eQTL	GWAS traits	Altered motifs	Conserved	Functional validation
rs590041	2.5E-105	0.07	*SSBP3* (intron)	Heart-specific	P wave duration (**Lead** = **rs562408**)[37]	Tbx5, Nkx2.5	-	EMSA, luciferase assay, CRISPRi
rs562408	7.9E-04	0.05				-	-	-
rs35176054	3.4E-18	0.16	*SH3PXD2A* (intron)	-	Atrial fibrillation (**Lead**)[47]	Gata,	Yes	-
rs7612445	2.1E-15	0.08	*GNB4* (>3 kb)	Heart-specific	Heart rate (**Lead**)[15,39]	Meis1, Tbx5	-	EMSA
rs4890490	2.1E-12	0.29	*SETBP1* (intron)	-	QRS duration[55–57]	-	-	-
rs4657167	3.5E-12	0.74	*NOS1AP* (intron)	-	QT interval[42]	-	-	-
rs6606689	3.8E-09	0.29	*PPTC7* (intron)	Other	Heart rate[15]	-	Yes	-
rs7132327	4.9E-04	0.68	*TBX3* (>130 kb)	-	PR segment[14] PR interval[13] QRS duration (**Lead**)[56]	-	-	-
rs3807989	6.9E-04	0.66	*CAV1* (intron)	Other	PR segment (**Lead**)[14] PR interval (**Lead**)[13,41,43] Atrial fibrillation (**Lead**)[58]	-	Yes	EMSA, luciferase assay, CRISPRi
rs8044595	1.4E-03	0.62	*MYH11* (intron)	-	Resting heart rate[39]		-	-
rs6932481	2.0E-03	0.79	*SAMD3* (intron)	Other	PR interval[59]	-	-	-
rs6801957	4.2E-03	0.37	*SCN10A* (intron)	-	PR segment (**Lead**)[14] PR interval (**Lead**)[13,40,41] QT interval (**Lead**)[42] P wave duration (**Lead**)[14] QRS duration (**Lead**)[43,44] Brugada syndrome[60] Resting heart rate[39]	Meis1	Yes	EMSA, reporter assays, from [45]
rs7986508	1.0E-02	0.65	*LRCH1* (intron)	Heart-specific	PR segment[14]	-	-	,-
rs10841486	1.2E-02	0.28	*PDE3A* (>49 kb)	Other	Resting heart rate (**Lead**)[39]	Eomes	-	-
rs6569252	1.7E-02	0.63	*GJA1* (>7 Mb)	-	Atrial fibrillation[47]	-	-	

Fourteen GWAS loci for EKG traits overlapping NKX2-5 ASE-SNVs, ordered by *P*-value for imbalance. For each SNV, we indicate the dbSNP ID (build 137), ASE corrected *P*-value (FDR) combined across heterozygous samples from the seven individuals, ASE reference allele ratio, the closest genes and relative location of the SNV, known association with gene expression (eQTL) and in which tissue (heart-specific = restricted to left ventricle and/or atrial appendage in GTEx; other = any other tissue or cell line), associated EKG GWAS traits and if the SNV is the lead variant, altered motifs, conservation in mammals and experiments performed for functional validation in this or previous studies. Additional annotations are reported in **Supplementary Table 5**.

Europe PMC Funders Author Manuscripts