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

Paola Benaglio1, Agnieszka D’Antonio-Chronowska2,9, Wubin Ma3,9, Feng Yang2, William W. Young Greenwald4, Margaret K. R. Donovan4,5, Christopher DeBoever4, He Li2, Frauke Drees2, Sanghamitra Singhal1, Hiroko Matsui2, Jessica van Setten6, Nona Sotoodehnia7,8, Kyle J. Gaulton1, Erin N. Smith1, Matteo D’Antonio6, Michael G. Rosenfeld3,8, Michael G. Rosenfeld3,8* and Kelly A. Frazer1,2*

The cardiac transcription factor (TF) gene NKX2-5 has been associated with electrocardiographic (EKG) traits through genome-wide association studies (GWASs), 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 induced pluripotent stem cell–derived cardiomyocytes from seven related individuals, and identified ~2,000 single-nucleotide variants associated with allele-specific effects (ASE-SNVs) on NKX2-5 binding. NKX2-5 ASE-SNVs were enriched for altered TF motifs, for heart-specific expression quantitative trait loci and for EKG GWAS signals. Using fine-mapping combined with epigenomic data from induced pluripotent stem cell–derived cardiomyocytes, 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.

GWASs for EKG phenotypes have found >500 risk variants, the majority of which are noncoding and enriched in regulatory elements of the genome. Detecting the causal variants and the molecular mechanisms that drive these associations has been challenging; therefore, only a handful of genetic associations with EKG traits have been explained by variants with clear molecular mechanisms. Altered TF binding has been proposed as one of the major mechanisms by which noncoding regulatory variants are causally associated with complex traits. NKX2-5 is an evolutionarily conserved, cardiac-specific TF, which, through cooperative binding with other core cardiac TFs such as TBX5 and GATA4, regulates heart development and is implicated in a spectrum of human congenital heart defects. Moreover, common noncoding variants near NKX2-5, TBX5 and MEIS1 have been associated through GWASs 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 with 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, making them a suitable model to perform expression quantitative trait locus (eQTL) studies. However, there are only a few studies showing similar utility of iPSC-derived cardiomyocytes (iPSC-CMs) to study regulatory variations, with potential limitations being cell type heterogeneity that arises from directed differentiation, as well as the functional immaturity of iPSC-CMs. Thus, while human iPSC-CMs are a promising model system, it is 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 sequencing (RNA-Seq), assay for transposase-accessible chromatin using sequencing (ATAC-Seq) and chromatin immunoprecipitation sequencing (ChIP-Seq) of...
both NKK2-5 and histone modification H3K27ac. After identifying heterogeneous sites that showed allele-specific effects (ASEs), we investigated NKK2-5 ASE-SNVs in detail by examining whether they altered cardiac TF motifs and whether they were enriched for eQTLs and EKG GWAS single-nucleotide polymorphisms (SNPs). By applying a fine-mapping statistical approach to three GWAS studies (heart rate, atrial fibrillation and PR interval), we found the effects that are driven by different genetic backgrounds24. Genetic background underlies the 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 backgrounds41. To identify sources of variability in our iPSC-CM datasets, and to evaluate the contribution of genetic background to this variation compared with the iPSCs, we performed principal component 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 ten principal components. 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 U-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 (that is, lower correlation values) than the iPSCs, probably due to cellular heterogeneity42–44. These analyses show that genetic background was a major driver of variability in our iPSC-CM molecular datasets.

**Results**

**Generation and functional genomic profiling of iPSC-CMs.** We generated iPSC-CMs from seven individuals in a three-generation family that included three genetically unrelated subjects and two parent–offspring quartets (Fig. 1a and Supplementary Table 1). In total, we differentiated nine iPSC lines 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 total, we differentiated nine iPSC lines into 26 iPSC-CM samples: generated iPSC-CMs from seven individuals in a three-generation generation and functional genomic profiling of iPSC-CMs. We showed that variation affecting the binding of NKX2-5 and other cardiac TFs probably serves as a molecular mechanism underlying the 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.
Fig. 2 | Identification of coordinated ASEs in gene expression, H3K27 acetylation, chromatin accessibility and NKX2-5 binding in iPSCs and iPSC-CMs.

**a.** Total numbers of regions and heterozygous SNVs tested for ASE across all individuals and samples in each dataset. **b.** Total numbers of heterozygous SNVs and corresponding regions across all individuals and samples with ASEs at FDR < 0.05. Numbers of ASEs shared between iPSCs and iPSC-CMs are indicated by hatching. **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 ASEs in heterozygous individuals and the effect size of each ASE-SNV, expressed as the slope of linear regression ($\beta$) between gene expression or peak density and the genotypes of all seven individuals. Spearman correlation statistics are indicated. The numbers of SNVs analyzed were: $n = 970$ for iPSCs and $n = 799$ for iPSC-CMs in **d;** $n = 255$ for iPSCs and $n = 550$ for iPSC-CMs in **e;** and $n = 1,714$ in **f.** Scatterplot showing the relationship between effect sizes ($\beta$ values) of ASE-SNVs in NKX2-5 peaks on both NKX2-5 and H3K27ac phenotypes ($n = 854$ SNPs). **g.** Table showing Spearman correlation coefficients of effect sizes between pairs of different molecular phenotypes. Correlations were calculated between $\beta$ values of SNVs that showed ASEs in ChIP-Seq datasets (rows) and $\beta$ values of the same variants for the closest gene or peak in a different molecular phenotype dataset (columns).

**NKX2-5 peaks commonly show ASEs.** 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 ASEs within each individual. First, we merged the sequencing reads of different samples from the same subject and calculated the ASEs. 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) and 123,151 (H3K27ac in iPSC-CMs) heterozygous SNVs within 12,492–57,631 regions (genes or peaks) (Fig. 2a) and identified the fraction of SNVs with significant imbalance at a false discovery rate (FDR) < 0.05 (ASE-SNVs) (Fig. 2b). The different phenotypes showed a difference of >30-fold in the percentage 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 to 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
iPSCs are due to genetic variation, and that, among all molecular samples, and observed a significant (P; linear regression) on the phenotype when including homozygous heterozygous individuals was consistent with the overall effect size the two cell types. We further tested whether the ASE observed in the other cell type (r=0.33; P=4×10^{-12}; and r=0.59; P=1.3×10^{-11}, respectively), with H3K27ac peaks in nearby or overlapping regions in the other cell type, suggesting conserved genetic effects at shared enhancers and promoters. In contrast, while H3K27ac ASE 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=4×10^{-12}; and r=0.41; P=1.7×10^{-7}, respectively, within the same cell type; and r=0.17; P=7×10^{-4} and r=0.06; P=0.43, respectively, for mismatched comparisons; Fig. 2h). These results show that, in both iPSC-CMs and iPSCs, genetic variation underlies coordinated and cell type-specific differences across multiple molecular phenotypes. Of note, while 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 (β; 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 show that the majority of ASEs 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. To examine whether NKX2-5 loci with ASEs were correlated with H3K27ac and gene expression ASEs, we compared the effect sizes (β) of ASE-SNVs identified within ChIP-Seq peaks with the effect sizes of the same SNVs 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=1.7×10^{-77} for NKX2-5 ASE-SNVs (Fig. 2g); and r=0.60; P=1.6×10^{-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.

**Fig. 3 | TF binding motifs are altered by SNVs with ASEs in NKX2-5 ChIP-Seq.** a. Odds ratios from a two-sided Fisher’s exact test comparing the proportion of motif-altering SNVs between variants with ASEs (n=1,941) and variants without ASEs (n=19,371) in NKX2-5 ChIP-Seq peaks from combined iPSC-CM samples. Asterisks indicate enrichment at an FDR-corrected P value <0.05. b. Numbers 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 (β) against the motif score (r) for all molecular phenotypes (Fig. 2d–f), with the highest correlation, P<0.05. Numbers of TFBS motifs that were strengthened (odds ratio) and variants without ASEs (proportion of motif-altering SNVs between variants with ASEs (β) 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 (that is, they are not stacked).
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 underlies NKX2-5 ASE-SNVs. To investigate whether NKX2-5 ASE-SNVs affected sequence motifs of TF binding sites (TFBSs), 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 ASEs within NKX2-5 peaks, we calculated the motif position weight matrix (PWM) score of each motif. We then compared SNVs with ASEs versus SNVs without ASEs and observed that those with ASEs 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 12 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 other known cardiac TFs, suggesting that differential allelic binding of NKX2-5 at these sites probably occurred either directly, due to alterations of its own binding sequence, or indirectly, via alterations of TFBSs of co-binding partners.

NKX2-5 ASE-SNVs modulate cardiac-specific gene expression. We examined whether 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 (QTLs) from diverse cell types, including DNase hypersensitivity QTLs (dsQTLs) in lymphoblastoid cell lines (LCLs)34, eQTLs from iPSCs34 and eQTLs from 13 combined studies obtained from haploReg33 (‘combined tissues’) (Fig. 4a–c and Supplementary Table 5). In iPSC-CMs, H3K27ac ASE-SNVs were enriched over SNVs without ASEs for all three types of QTL (Fisher’s exact test, P < 0.05). In contrast, H3K27ac ASE-SNVs in iPSCs were only enriched for ipSC eQTLs. Of note, NKX2-5 ASE-SNVs were significantly depleted for iPSCs and combined tissue eQTLs, suggesting that they exert regulatory functions only in cardiac tissues.

We therefore investigated whether NKX2-5 ASE-SNVs were enriched for heart-specific eQTLs. NKX2-5 and H3K27ac ASE-SNVs were compared with SNVs without ASEs to assess enrichment for tissue-specific eQTLs (defined in the Methods) in 26 tissue types from the Genotype-Tissue Expression (GTEX) project (version 6)44. 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 nine affected the NKX2-5 binding motif and 13 affected one or more of the other cardiac TF motifs in Fig. 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 traits13–15,35,36, we hypothesized that the altered binding of NKX2-5 in other GWAS loci could be causally implicated in these traits. First, we examined whether NKX2-5, H3K27ac or ATAC peaks from iPSC-CMs were enriched for all EKG GW AS SNPs, while H3K27ac and DHS peaks from Roadmap cardiac tissues, which similarly showed high enrichment for all EKG GW AS SNPs, while H3K27ac and DHS peaks from iPSCs did not (Supplementary Fig. 6). These data show that enhancer regions in iPSC-CMs and Roadmap cardiac tissues both show enrichment for EKG trait-specific regulatory variants.

To examine whether differential binding of NKX2-5 might have a role in EKG phenotypes, we determined whether NKX2-5

![Fig. 4 | Enrichment of ChIP-Seq ASE variants for known QTLs.](image-url)

- a-c, Histograms showing the percentage of SNVs with and without ASEs in each ChIP-Seq (from combined iPSC or iPSC-CM samples) and overlapping dsQTLs from LCLs34 (a), eQTLs from iPSCs34 (b) and combined eQTLs identified in different tissues33 (c). Two-sided Fisher’s exact test P values are shown in red or blue for enrichment or depletion, respectively. OR, odds ratio. d, Heat map showing enrichment of ASE variants for tissue-specific eQTLs34 (similar tissues in GTEx were merged; see Methods). Asterisks indicate two-sided Fisher’s exact test FDR-corrected P values < 0.05. The heat map is colored based on −log10(FDR-corrected P values), with a negative sign if the odds ratio was < 1. The complete Fisher’s exact test statistics, including P values, odds ratios and numbers of SNVs analyzed, are reported in Supplementary Table 5.
Fig. 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 an FDR-corrected $P$-value $<0.05$, as calculated using GREGOR. In total, $n = 125$ GWAS traits were tested, of which six were for EKG traits. d. Percentage of NKX2-5 ASE-SNVs overlapping an EKG GWAS SNP versus overlapping a non-GWAS SNP. The two-sided Fisher’s exact test $P$-value and numbers of SNVs are given. e. From top to bottom: regional plot of association values with $P$-wave duration. f. qPCR expression of SSBP3 in iPSC-CMs (ID: iPSCORE_1_5) and were lactate purified.

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 ASEs. Of these, 14 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 ($19\%/121$) using a $t$-test; odds ratio $= 1.88; P = 0.0392$; Fig. 5d). Among these 14 NKX2-5 ASE-SNVs at EKG GWAS loci, seven were evolutionarily 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 the 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.54742618A>T) (Table 1). Two SNPs in the TF SSBP3 locus are in perfect linkage disequilibrium and showed ASES in the same peak; while rs562408 (NC_000001.10:g.54742618A>G) was the lead variant in a P-wave duration GWAS, our data suggested that rs590041 is the probable 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 associated with lower atrial fibrillation (lead)47 Gata Yes – EMSA, luciferase assay and CRISPRi (rs590041)

we experimentally investigated the SNV that showed the strongest evidence for allelic imbalance: rs590041 (NC_000001.10:g.54742618A>T) (Table 1). Two SNPs in the TF SSBP3 locus are in perfect linkage disequilibrium and showed ASES in the same peak; while rs562408 (NC_000001.10:g.54742618A>G) was the lead variant in a P-wave duration GWAS, our data suggested that rs590041 is the probable 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 associated with lower atrial fibrillation (lead)47 Gata Yes – EMSA, luciferase assay and CRISPRi (rs590041)

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 | 2.5 × 10^-10  | 0.07 | SSBP3 (intron) | Heart specific | P-wave duration (lead = rs562408) | Tbx5 and Nkx2-5 | – | EMSA, luciferase assay and CRISPRi (rs590041) |
| rs562408 | 7.9 × 10^-4  | 0.05 | SSBP3 (intron) | Heart specific | P-wave duration (lead = rs562408) | Tbx5 and Nkx2-5 | – | EMSA, luciferase assay and CRISPRi (rs590041) |
| rs35176054 | 3.4 × 10^-38  | 0.16 | SSBP3 (intron) | Atrial fibrillation (lead) | Gata | Yes | – |
| rs7612445 | 2.1 × 10^-14  | 0.08 | GNB4 (>3 kb) | Heart specific | Heart rate (lead) | Meis1 and Tbx5 | – | EMSA |
| rs4890490 | 2.1 × 10^-12  | 0.29 | SETBP1 (intron) | – | QRS duration | – | – | – |
| rs4657167 | 3.5 × 10^-12  | 0.74 | NOS1AP (intron) | – | QT interval | – | – | – |
| rs6606689 | 3.8 × 10^-9   | 0.29 | PPTC7 (intron) | Other | Heart rate | – | Yes | – |
| rs7132327 | 4.9 × 10^-4   | 0.68 | TBX3 (>130 kb) | – | PR segment | – | Yes | – |
| rs3807989 | 6.9 × 10^-4   | 0.66 | CAV1 (intron) | Other | PR segment (lead) | – | Yes | EMSA, luciferase assay and CRISPRi |
| rs8044959 | 1.4 × 10^-3   | 0.62 | MYH11 (intron) | – | Resting heart rate | – | – | – |
| rs6932481 | 2.0 × 10^-3   | 0.79 | SAMD3 (intron) | Other | PR interval | – | – | – |
| rs6801957 | 4.2 × 10^-3   | 0.37 | SCN10A (intron) | – | PR segment (lead) | – | Yes | EMSA and reporter assays |
| rs79860508 | 1.0 × 10^-2  | 0.65 | LRCH1 (intron) | Heart specific | PR segment | – | – | – |
| rs10841486 | 1.2 × 10^-2  | 0.28 | PDE3A (>49 kb) | Other | Resting heart rate (lead) | – | Eomes | – |
| rs6569252 | 1.7 × 10^-2  | 0.63 | GJA1 (>7 Mb) | Other | Atrial fibrillation | – | – | – |

Fourteen GWAS loci for EKG traits overlapping NKX2-5 ASE-SNVs, ordered by P-value for allelic imbalance, are listed. 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, closest genes and relative location of the SNV, known association with gene expression (eQTL), tissue (heart specific = restricted to the left ventricle and/or atrial appendage in GTEx; other = any other tissue or cell line), associated EKG GWAS traits, whether 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.

**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—a statistical framework that integrates functional genomics annotations and GWAS summary statistics to identify putative causal variants at known loci, as well as at potentially novel loci. We initially applied a single annotation model to examine a heart rate meta-analysis to determine whether 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 that NKX2-5 ASE-SNVs were the most enriched annotation, followed by NKX2-5 peaks (Supplementary Fig. 7). Next, we applied a joint model, where the association enrichment was quantified simultaneously for all five annotations and refined using tenfold cross-validation, and again found 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-megabase (Mb) windows across the genome. We found 21 variants with a >30% probability of being causal, of which seven (30%) were NXX2-5 ASE-SNVs (Supplementary Table 6), suggesting that altered binding of NXX2-5 accounts for a considerable fraction of the genome-wide genetic contribution underlying variable heart rate. Of these seven NXX2-5 ASE-SNVs (Fig. 6b), four were from 'subthreshold' loci that did not reach genome-wide significance in the heart rate meta-analysis, but was significantly associated in a considerable fraction of the genome-wide genetic contribution underlying variable heart rate. Of these seven NXX2-5 ASE-SNVs (Fig. 6b), four were from 'subthreshold' loci that did not reach genome-wide significance in the heart rate meta-analysis, but was significantly associated in

Validation of the NXX2-5 ASE-SNV rs3807989 as a functional variant at the CAVI locus. To examine other EKG traits, we applied fgwas fine-mapping framework to both atrial fibrillation and PR interval GWAS studies (Fig. 7a and Supplementary Fig. 7), and identified 26 and 102 SNPs, respectively, with a >30% probability of being causal, of which 8% (2/26) and 14% (14/102) were placed within two adjacent 10 Mb regions (Fig. 7a and Supplementary Fig. 7c). To examine other EKG traits, we applied fgwas fine-mapping framework to both atrial fibrillation and PR interval GWAS studies (Fig. 7a and Supplementary Fig. 7), and identified 26 and 102 SNPs, respectively, with a >30% probability of being causal, of which 8% (2/26) and 14% (14/102) were placed within two adjacent 10 Mb regions (Fig. 7a and Supplementary Fig. 7c).
Fig. 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 (gray) and PR interval (orange) in iPSC-CM genomic annotations (y axis). Bars indicate 95% confidence intervals. b, Tables showing the top five SNPs ordered by fgwas PPA and overlapping at least two of the indicated iPSC-CM genomic annotations. c, From top to bottom: regional plot of association values with PR interval17, color coded based on linkage disequilibrium (r²; squared Pearson correlation) values14 (NKX2-5 allelic imbalance (pie chart) for rs3807989 is also shown); fgwas PPAs of the variants in the locus; epigenetic tracks from iPSC-CM combined samples; and 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 in 128 iPSC-CMs from different individuals16. Box plots show median values (thick lines), lower and upper quartiles (box edges), and maximum and minimum values (whiskers). P values of linear regression are shown. e, EMSA with iPSC-CM nuclear extracts, which showed that oligonucleotide probes for the reference allele (A) bound more strongly than those for the alternate allele (G), consistent with the allelic imbalance 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 base pairs (bp) from a NKX2-5 motif (Fig. 7f), and could modify a sequence important for recognition of the binding site, such as a biotinylated antibody. f, Luciferase assays in iPSC-CMs for rs3807989, in which expression levels were enhanced compared to the control gRNA (gCTL) and either a control gRNA (gCTL) or two gRNAs targeting the CAV2 CAV1 CAV1 gene. g, Luciferase assays in iPSC-CMs stably expressing dCas9–KRAB (CRISPRi) (ID: iPSCORE_1_5) and either a control gRNA (gCTL) or two gRNAs targeting the rs3807989 region. h, Luciferase assays in iPSC-CMs stably expressing dCas9–KRAB (CRISPRi) (ID: iPSCORE_1_5) and either a control gRNA (gCTL) or two gRNAs targeting the rs3807989 region. i, Luciferase assays in iPSC-CMs stably expressing dCas9–KRAB (CRISPRi) (ID: iPSCORE_1_5) and either a control gRNA (gCTL) or two gRNAs targeting the rs3807989 region.
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 show 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 to be the causal variant underlying the atrial fibrillation and PR interval GWAS signals in the CAV1 interval.

Discussion

Our study shows that differential binding of NKX2-5 probably 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 shows 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 the binding of the TF. 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: (1) affect consensus motifs from TFs that were not included in our targeted analysis; (2) affect important sequences that impact DNA shape or an as-of-yet unknown regulatory mechanism; or (3) be non-functional. Combinatorial interactions between key cardiac TFs are known to be an important mechanism for orchestrating the cardiac gene expression program during development11-11. While genetic variation has been shown to affect collaborative binding of lineage-determining TFs in mice15, our study shows these effects in humans.

Coding mutations in (and noncoding variants near) the NKX2-5 gene have, respectively, been associated with congenital heart defects12, as well as heart rate, atrial fibrillation and PR interval13-13, implicating this TF in a range of cardiac diseases 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 (that is, variation in TFBs resulting in the 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 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 subthreshold GWAS signals. As proof that this approach was effective to prioritize causal variants, one of the NKX2-5 ASE-SNVs (rs6801957 at the SCN10A–SCN5A locus) had previously been investigated in detail and had been shown to be functionally implicated in the association with EKG14,14. Further investigation of NKX2-5 ASE-SNV 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 the expression of nearby or distal candidate target genes. As a notable example, one of the prioritized variants (rs176107) at a subthreshold 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 showed 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 roles that differential bindings of NKX2-5 and other cardiac TFs play in the genetic underpinnings of EKG traits.

Finally, our study shows 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 TFs (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.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0499-3.

Received: 21 February 2018; Accepted: 15 August 2019; Published online: 30 September 2019

References

1. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 45, D896–D901 (2017).
2. Gallagher, M. D. & Chen-Plotkin, A. S. The post-GWAS era: from association to function. Am. J. Hum. Genet. 102, 717–730 (2018).
3. Van den Boogaard, M. et al. A common genetic variant within SCN10A modulates cardiac SCN5A expression. J. Clin. Invest. 124, 1844–1852 (2014).
4. Wang, X. et al. Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures. eLife 5, e10557 (2016).
5. Deplante, B., Alpern, D. & Gardeux, V. The genetics of transcription factor DNA binding variation. Cell 166, 538–554 (2016).
6. Pai, A. A., Pritchard, J. K. & Gilad, Y. The genetic and mechanistic basis for regulating transcription factors identifies transcriptional enhancers active in heart. Proc. Natl Acad. Sci. USA 108, 5632–5637 (2011).
7. Muraano, M. T. et al. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. Nat. Genet. 47, 1393–1401 (2015).
8. He, A., Kong, S. W., Ma, Q. & Pu, W. T. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. Proc. Natl Acad. Sci. USA 108, 5632–5637 (2011).
Articles

38. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Nature*, 501, 506–511 (2013).

39. Sun, Y. M. A. H., Brunson, B. G. & Pollard, K. S. A de novo shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs. *Cell Syst.*, 8, 42–62 (2019).

40. Afek, A., Schipper, J. L., Horton, J., Gordan, R. & Lukatsky, D. B. Protein–DNA binding in the absence of specific base-pair recognition. *Proc. Natl. Acad. Sci. USA*, 111, 1740–1745 (2014).

41. Slattery, M. et al. Absence of a simple code: how transcription factors read genome-wide association studies of 18 human traits. *Nature*, 503, 487–492 (2013).

42. Johnson, A. D. et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, 24, 2938–2939 (2008).

43. Hong, K. W. et al. Identification of three novel genetic variations associated with electrocardiographic traits (QRS duration and PR interval) in East Asians. *Hum. Mol. Genet.*, 23, 6659–6667 (2014).

44. Van der Harst, P. et al. 52 genetic loci influencing myocardial mass. *J. Am. Coll. Cardiol.*, 68, 1435–1448 (2016).

45. Evans, D. S. 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.*, 25, 4350–4368 (2016).

46. Ellinor, P. T. et al. Meta-analysis identifies six new susceptibility loci for atrial fibrillation. *Nature*, 464, 670–675 (2010).

47. Johnson, M. et al. Generalization of variants identified by genome-wide association studies for electrocardiographic traits in African Americans. *Ann. Hum. Genet.*, 77, 321–332 (2013).

48. Bezzina, C. R. et al. Common variants at SCNSA-SCN10A and HFE2 are associated with Brugada syndrome, a rare disease with high risk of sudden cardiac death. *Nature*, 45, 1044–1049 (2013).

Acknowledgements

This work was supported in part by California Institute for Regenerative Medicine grant (CIRM) GC10-06673-B, NIH grants HG018188 and HL107442, and the National Science Foundation grant 1728497. B.P. was supported by the Swiss National Science Foundation Postdoctoral Mobility fellowships P2AP3-135105 and P300PA-167612. WWYG was supported by the NHLBI under award number HL142151. C.D. was supported in part by the UCSD Genetics Training Program through an institutional training grant from the NIGMS under award number GM008666 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. Kroshehadi at the UCSD IGM Genomics Center, supported by NIH grant CA023100. N.S. was supported by NIH grants HL116747 and HL114989. K.I.G. was supported by NIH grants DK114650 and ADA grant 1–17JDE027. W.M., F.Y. and M.G.R. were supported by NIH grants DK018477 and DK039949. M.G.R. is a HHMI investigator. We are thankful to C.-A. Yen and N. Spann for assistance with the ChIP-seq experiments, and to A. Schmitt for the Hi-C data. We thank A. Aguirre for performing immunofluorescence. We thank E. Farley and K. Olson for help with reporter assays. We thank many colleagues for helpful comments.

Author contributions

PB designed the study, generated the ChIP-seq and RNA-seq data, and performed the statistical analyses. A.D.A.-C. generated the iPSC-CMs, ChIP-seq, ATAC-seq and RNA-seq data, and performed the EMSA. W.M. generated the constructs for luciferase assay and CRISPRi, and performed the luciferase assays. F.Y. performed the CRISPRi experiments. WWYG implemented the igwas analysis pipeline. C.D. implemented the RNA-seq, ATAC-seq and ASE analysis pipelines. H.I. processed the 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.A. and E.N.S. performed statistical analyses. M.G.R. supervised the 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.

**Additional information**
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0499-3.

Correspondence and requests for materials should be addressed to M.G.R. or K.A.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019
Methods

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

Subjects and iPSC derivation. We selected seven individuals who were part of a three-generational family (three genetically unrelated subjects and two parent–offspring quartets) in the iPSCORE resource 

Supplementary Table 1.  Fibroblasts from skin biopsies of each subject were reprogrammed using non-integrative Sendai virus 

and analyzed for pluripotency as described by Panopoulou et al. 

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 and 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, San Diego (project number 110777ZEF).

Differentiation of iPSCs into cardiomyocytes. The nine iPSCs were each differentiated multiple times using a monolayer protocol, 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 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 Fisher Scientific), pooled, counted and separated into different aliquots. About 6 × 10^7 cells were fixed with formaldehyde and frozen for ChIP-Seq. Cells (2 × 10^7) were lysed and stored in RLT Plus buffer (Qiagen) for RNA extraction. Nuclei from 2 × 10^6 cells were frozen for ATAC-Seq. The differentiation efficiency was measured by the percentage of cells that stained positive for the cardiac marker cardiac troponin T (TNNT2; MA5-12960; Thermo Fisher Scientific) 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 (WGS). Genomic DNA was whole-genome sequenced as part of the iPSCORE collection, as described by DeBoever et al. 

Briefly, reads were aligned against human genome b37 using BWA-MEM and default parameters. The resulting BAM files were sorted using Sambamba, and duplicate reads were marked using biobambam2 (ref. 67). Variant calling was performed using the GATK best-practices pipeline 

and normalized to transcripts per million (TPM).

RNA-Seq. We generated and analyzed 36 RNA-Seq samples (iPSC-CMs: 29 independent samples; iPSC-CMs: 26 independent samples and one technical replicate). Total RNA was isolated using the Qiagen RNeasy Mini Kit from frozen RLT plus pellets, and run on a Bioanalyzer (Agilent). Illumina TruSeq Stranded mRNA libraries were prepared and sequenced on a HiSeq 2500 system, to an average of 40 million 100-bp paired-end reads per sample. RNA-Seq reads were aligned using STAR 

with a splice junction database built from the GENCODE v19 gene annotation. Gene-based expression values were quantified using the RSEM package 

and normalized to transcripts per million bp (TPM).

ChIP-Seq. We generated and analyzed 48 ChIP-Seq samples of histone modification H3K27ac (iPSCs: 17 samples and one technical replicate; iPSC-CMs: 25 samples and two technical replicates) and 15 ChIP-Seq samples of NKX2-5 (iPSC-CMs: 12 samples and three technical replicates) (Supplementary Tables 2 and 3), using anti-H3K27ac (ab4729; Abcam) and anti-NKX2-5 antibodies (sc-8697; Santa Cruz Biotechnology). 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. Duplicate reads, mapping to blacklisted regions and read pairs with mapping quality Q < 30 were filtered. Peak calling was performed using MAC2S (ref. 1), with reads derived from sonicated chromatin not subjected to immunoprecipitation (that is, input chromatin) from a pool of samples used as a 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 

Motif enrichment analysis was performed using Homer and, for NKX2-5, also using MEME-Chip.

ATAC-Seq. We generated 37 ATAC-Seq libraries (iPSCs: 12 samples and five technical replicates; iPSC-CMs: 11 samples and nine technical replicates) using an adapted protocol from Buenrostro et al. 

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 140bp. Peak calling was performed using MAC2S 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 eight RNA-Seq samples from Roadmap, including H1-iPSC, HUES64, iP5-20b, iP5-18, 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 (ref. 1). Hierarchical clustering and the heat map in Supplementary Fig. 1 were generated using vst-normalized read counts for a panel of 61 selected genes using the ‘heatmap’ 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 between iPSCs and iPSC-CMs, we used a matrix of raw expression counts from 56 RNA-Seq samples (29 iPSCs and 27 iPSC-CMs), filtered for an average TPM value of >1 (22,447 genes), and applied DESeq2 with default settings to identify genes that were differentially expressed more than twofold and at a Benjamini and 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 one 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 kilobases (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 five 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 ten principal components were correlated with known covariates across samples (for iPSCs, sex, differentiation batch, passage and subject; for iPSC-CMs, TNNT2 expression, protocol of differentiation and subject; for ChIP-Seq of both cell types, we also associated the fraction of reads mapping to peaks) using analysis of variance. We corrected the respective datasets by fitting a model including the covariates that were most associated with the first principal component (batch for iPSCs; TNNT2 expression and protocol/batch for iPSC-CMs; and fraction of reads mapping to peaks for all ChIP-Seq datasets) using the lmFit function from the 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 1,003 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 a significant increase in correlation between samples from the same subject using a one-tailed Mann–Whitney U-test (Fig. 1c–g and Supplementary Fig. 3k–o).

A SE analysis. ASE analysis was performed as previously described. 

To increase the sensitivity of ASEs 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 exomic or genome-wide SNVs from GENCODE v19 or regions identified by each ChIP-Seq or ATAC-Seq dataset. The WASP pipeline 

employed to reduce reference allele bias at heterozygous sites. The number of read pairs supporting each allele was counted using the ASEReadCounter from GATK. 

Heterozygous SNVs were then filtered to keep SNVs where the reference or alternate allele had more than five supporting read pairs, the reference allele frequency was between 2 and 98%, and the SNV was located in unique mapability regions according to wgEncodeCrgMapabilityAlign100mer track, and not located within 10bp of another variant in a particular subject (heterozygous or homozygous alternative) 

. ASE p-values for each SNV were calculated in each sample using a binomial test method. 

To combine ASE results at each SNV across samples, we performed a meta-analysis on data that was derived from two SNVs and for which ASEs could be tested. The binomial p-values of heterozygous SNVs were combined using the Stouffer z-score method 

, using the formula 

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 -scores were transformed to p-values and a Benjamini and Hochberg 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 β coefficient of a linear model testing the association between the corrected gene expression or peak coverage (normalized 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 β with the average allele proportion of ASE-SNVs, to estimate the consistency of effects (Fig. 2d–i).
Correlation of ASEs across different molecular phenotypes. To test whether the direction of ASEs 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 analysis using ChIP-nexus, which accounted for multiple SNV genotypes and each phenotype. ChIP-Seq peaks were paired with the closest gene or peak within 500bp using bedtools closest. Using linear regression, we tested the association between the individual genotypes (0, 1 or 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 of those of the closest feature. In both peak–gene and peak–peak pairs, Spearman correlation was calculated between the two slopes (β) 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 transcription factors identified using GOMER (Supplementary Table 4), from a library of known motifs. For NKX2, GATA, TEAD, MEFI2, TEX20 and PDX1, we also used PPMs derived from a de novo analysis. All PPMs are provided in Supplementary Table 4. 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 in each SNV tested for ASEs 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 an SNV matched multiple motifs from the same family the motif of the highest score, was used for the alternate allele.

Enrichment of ASE-SNVs for known QTLs. To determine whether SNVs with ASEs compared with variants without ASEs (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 and Supplementary Fig. 5). Motifs that were altered at NKX2-5 ASE-SNVs are indicated 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. We narrowed the search to 119 non-EKG traits that were associated with a similar number of SNPs. We used GREGOR84 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 of ChIP-Seq GWAS SNPs in NKX2-5 ASE-SNVs, we obtained the SNPs overlapping NKX2-5 peaks and associated with any of the six EKG traits. For the SNPs that could be tested for ASEs, we calculated the proportion with and without ASEs and tested their relative enrichment using Fisher’s exact test (Fig. 5d).

Estimating GWAS enrichment in molecular phenotypes and prioritizing putative causal variants. To test 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 gwas framework, as described by Pickrell et al.40. We obtained summary statistics from the den Hoed et al.11 heart rate GWAS meta-analysis (2,516,407 SNPs analyzed) from LD Hub (http://ldsc.broadinstitute.org/lidhub), the Christophersen et al. atrial fibrillation meta-analysis (11,779,664 SNPs) from the CVD project (http://cavd.cgu.org) and the van Setten et al.13 PR interval 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 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 (PPA) within the same 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 individuals2. Subjects included 43 males and 85 females, between 9 and 88 years of age, of diverse ethnicities (Europeans (n=78) and Asians (n=50) were differentiated with a day 25 cardiomyocyte isolation method described above, including a 4 mM sodium l-lactate enrichment step at day 15, and yielded on average 83.9±13.6% cardiac troponin T-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 PKE parameters and used the residuals of the first ten factors as normalized gene expression values. We extracted the individuals’ genotypes from WGS and performed linear regression for the specific SNP–gene expression associations in R.

EMSA. EMSAs were performed using the LightShift Chemiluminescent EMSA Kit (Thermo Fisher Scientific) with biotinylated and non-biotinylated single-stranded oligonucleotides corresponding to 33 or 34 genomic fragments containing the SNPs rs590041, rs3807989 and rs7512445 (Supplementary Table 7). Both forward and reverse strands were tested. The forward strand was bound in the case of rs590041 and rs7907989, and the reverse strand was bound in the case of rs7512445. Nuclear extract from day-30–33 iPSC-CMs was extracted using the NE-PER Nuclear and Cytoplasmic Extraction Reagents (Thermo Fisher Scientific) with 1× Halt Protease Inhibitor Cocktail (Thermo Fisher Scientific). The binding reaction was carried in 10 µl volume containing 1 µl of 10× Binding Buffer (100 mM Tris (pH 7.5), 500 mM KCl and 10 mM dithiothreitol), 2.5% glycerol, 5 mM MgCl2, 0.05% NP-40, 100 ng Poly[d(C-R)], 100 ng Poly[d(A-T)] and 50–150 µM radiolabeled DNA. For competition experiments, a 20-fold molar excess of unlabeled probe was added. Binding reactions were incubated at room temperature for 20 min and loaded onto a 6% polycrylamide 0.5X TBE gel. After sample electrophoresis and transfer to a 0.45-µm Biodine B pre-cut modified nylon membrane (Thermo Fisher Scientific), DNA was ultraviolet-crosslinked for 15 min, and radiolabeled DNA was detected using a Chemiluminescent Detection Module (Thermo Fisher Scientific). Membranes were acquired using a C-DiGit Blot scanner (LI-COR Biosciences).

Luciferase assay. The candidate functional variants rs590041 (SSBP3 intron) and rs3807989 (PDK1 intron) were tested for differential transcriptional activity by luciferase reporter assay. Regions of ~1.7 kb 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 homozgyous copy of a genome DNA, while for rs3807989, we were obtained by subcloning DNA with a heterozygous genotype. Cryopreserved day-25 iPSC-CMs were seeded onto a Matrigel-coated 96-well plate at a density of 30–40×103 cells per well and cultured in RPMI + insulin for 5–10 d before transfection, when the media was exchanged to Opti-MEM (Life Technologies). Each well was transfected with a mix of 120 ng Firefly Luciferase reporter vector, 30 ng Renilla Luciferase control vector (pRL-TK; Promega) and 0.6 µl Viafect transfection reagent (Promega) in 10 µl Opti-MEM. We transfected six wells per construct. Luciferase activity was measured 24 h after transfection using the Dual-Luciferase Reporter Assay System (Promega).

CrisPRi experiments. Two guide RNAs (gRNAs) targeting CAV1 and SSBP3 regulatory elements were designed using the online software CHOPCHOP (http://chopchop.cbu.uib.no) and cloned into the lentivector plK01:U6-29ng-cdcb-FE1a-Puromycin. Lentiviral gRNAs or Lenti-dCas9-KRAB-blast plasmids (89567; Addgene) were co-transfected with packaging plasmids (psPAX2 and pMD2.G) into human 293T cells. Culture medium containing lentiviruses particles for gRNA and dCas9-KRAB was harvested, mixed well with polybrene (10 µg ml<sup>−1</sup>), and added to a 24-well plate. Day-30 iPSC-CMs (cell lines iPSCORE_1_5 and iPSCORE_75_1) were dissociated and added to the virus-containing media at around 80% confluence. For a higher infection efficiency, a new collection of lentiviral particles mixed with polybrene was added to the medium after 24 h. Transfected cells was exchanged after 24 h to regular culture medium and selected to medium containing 0.2 µg ml<sup>−1</sup> puromycin and 6µg ml<sup>−1</sup> blasticidin after another 24 h. Cells were cultured for 6 d, when all cells from the noninfected control died, and then harvested. RNA was isolated with a Quick-RNA kit (Zymo Research) and reverse transcribed using SuperScript III Reverse Transcriptase (Life Technologies). Quantitative PCR (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. gRNAs and primers for qPCR are given in Supplementary Table 7.

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

Data availability. All iPSC lines are available through the WiCell Research Institute (www.wicell.org; NHLBI Next Gen Collection). All genomic data are available through the database of Genotypes and Phenotypes (accessions phs009024 (RNA-Seq, ChIP-Seq).
ATAC-Seq and Hi-C) and phs001325 (whole-genome-sequenced SNV and copy number variation genotypes)) and National Center for Biotechnology Information BioProject PRJNA285375. Processed data files are available through Gene Expression Omnibus accessions GSE123540 and GSE133833.

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

**References**
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* 108, 14234–14239 (2011).
62. Lian, X. et al. Directed cardiomyocyte differentiation from human pluripotent stem cells by modulating Wnt/β-catenin signaling under fully defined conditions. *Nat. Protoc.* 8, 162–175 (2013).
63. Tohyama, S. et al. Distinct metabolic flow enables large-scale purification of mouse and human pluripotent stem cell-derived cardiomyocytes. *Cell Stem Cell* 12, 127–137 (2013).
64. Auton, A. et al. A global reference for human genetic variation. *Nature* 526, 68–74 (2015).
65. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760 (2009).
66. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 31, 2032–2034 (2015).
67. Tischler, G. & Leonard, S. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.* 9, 13 (2014).
68. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498 (2011).
69. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303 (2010).
70. Dobin, A. et al. STAR: ultrafast universal RNA-Seq aligner. *Bioinformatics* 29, 15–21 (2013).
71. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774 (2012).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection Software was not used for data collection

Data analysis

- The following software were used: BWA-MEM (0.7.15); GATK(3.4-46); STAR (2.5.0a); RSEM(1.2.20); MACS2 (2.1.0); WASP (0.2.2); HOMER (4.7); fgwas; PLINK(1.07); GREGOR(v1.4.0); HICCUPS (1.6); Fit-Hi-C (1.0.1), DESeq2 (R 3.2.2), featureCounts (1.5.0), MEME-CHIP (online version), bedtools (2.25.0), Python (2.7.11), CHOPCHOP (online version, 2.0.0), Nexus Copy Number software (7.5).
- All statistical analyses were performed using R, version 3.2.2.
- Custom codes and Jupyter notebooks are available at https://github.com/frazer-lab/NKX2-5_ASE_iPSC-CM

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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 GSEXXXXXX.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

**Sample size**
No statistical methods were used to predetermine sample sizes. We selected seven individuals of Asian and European descent in the iPSCORE resource that are part of a three-generational family. Our study design included three genetically unrelated subjects and two parent-offspring quartets, which enabled us to examine the inheritance of genetic effects. The primary analyses in this manuscript identify differences in molecular phenotypes between two alleles within an individual (ASE – allele-specific effects). Therefore, the power for these analyses was primarily dependent on the read depth of RNA-Seq, ChIP-Seq or ATAC-Seq at heterozygous SNVs, rather than the number of subjects. To improve power we then combined the results from the same SNV across individuals using a meta-analysis.

**Data exclusions**
We used established quality control metrics and filtering criteria for all molecular data:
For all sequencing data, blacklisted regions from ENCODE, reads mapping in chromosome other than chr1-chr22, chrX, chrY, and read-pairs with mapping quality Q<30 were filtered.
For ChIP-Seq and ATAC-Seq, peaks were filtered for q-value <0.01, and samples with a FrIP <4% were excluded.

**Replication**
We compared the ASE (allele-specific effects) between the different molecular data, which showed good correlation, indicating that the genetic effects are reproducible.
Using our approach we identified at least one variant (at the SCN10A locus) that was previously described to be functional and showing allele-specific binding.
EMSA experiment for variants at the CAV1 and GNBP loci were performed twice and showed consistent results in both independent experiments, as shown in the paper’s figures; for the variant at the SSBP3 locus, we performed 4 EMSA experiments, of which 3 showed consistent allelic differences (two of those are shown in the paper) and one showed no difference.
CRISPRi was performed in two independent cell lines and the results were reproducible.
Luciferase assays had 6 replicates each variant, consisting of transfection of 6 different wells.

**Randomization**
As we are testing for genetic associations, there are no experimental groups in this study.

**Blinding**
As there are no experimental groups in this study, blinding was not applicable for this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☑ | Antibodies |
| ☑ | Eukaryotic cell lines |
| ☑ | Palaeontology |
| ☑ | Animals and other organisms |
| ☑ | Human research participants |
| ☑ | Clinical data |

### Methods

| n/a | Involved in the study |
|---|---|
| ☑ | ChIP-seq |
| ☑ | Flow cytometry |
| ☑ | MRI-based neuroimaging |

### Antibodies

**Antibodies used**
- anti-TNNT2 (Thermo Scientific MAS-12960, clone 13-11, dilution 1:200, monoclonal)
- anti-MYL7 (Synaptics Systems 311011, clone 56FS, dilution 1:200, monoclonal)
- anti-H3K27ac (Abcam ab4729, lots GR183922-2 (1.75μg per IP), GR184333-2 (1μg per IP), or GR00324078 (1μg per IP), polyclonal)
- anti-NKX2-5 (Santa Cruz Biotechnology, sc-8697x, lot C0113, 5μg per IP, polyclonal)

**Validation**
- anti-TNNT2: from company website: MAS-12960 targets Troponin T Cardiac Isoform in IF/ICC, IHC (P), and IM applications and shows reactivity with Avian, Canine, Chicken, Fish, Guinea Pig, Human, mouse, Porcine, Rabbit, and Rat samples. Referenced in 122 publications.
- anti-MYL7: from company website: Tested applications: WB, ICC, IHC, IHC-P/FFPE, Reacts with: human (Q01449), rat, mouse
Human research participants

Policy information about studies involving human research participants

Population characteristics

All subject information is provided in Supplementary Table 1. In summary, a family of seven individuals spanning three generations was utilized, with all individuals from EAS, EUR, or EAS/EUR descent, spanning ages 18-77. Five members of this family segregate the long-QT syndrome type 2 mutation KCNH2 p.Trp1001* (rs121912509, c.3003G>A), and the genotypes are reported. However, the disease phenotype is not analyzed in this study. These individuals are explained in Panopolous et al, as cited in the references. Individual covariates (sex, ethnicity, disease state) were not used in the analysis of NKX2-5 allele-specific effect.

Recruitment

Recruitment for these individuals is explained fully in Panopolous et al, as cited in the references. Specifically, these 7 individuals were recruited through both the Twin Sibling Pedigree cohort (TSP; a population-based twin registry spanning counties in Southern California) and open enrollment through the Clinical and Translational Research Institute (CTRI) at the University of California at San Diego (UCSD). Each of the subjects first consented to the study and filled out a questionnaire. These data were transcribed to a database and subjects were de-identified with a new sample ID. Ethnicity was reported as a free-response answer and translated into one of six recorded ethnicity groupings (African American, European, Hispanic, Indian, Middle Eastern, Asian). A seventh category was used when more than one ethnicity was reported; that individual was recorded as “Multiple ethnicities reported.” We do not report any recruitment or self-selection bias that could have influenced the results of this study.

Ethics oversight

This study was approved by the Institutional Review Boards of the University of California at San Diego (Project #110776Z).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

ChIP-seq

Data deposition

- Confirm that both raw and final processed data have been deposited in a public database such as GEO.
- Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication. BAM files for all ChIP-seq data are available through dbGAP (phs000924, BioProject PRJNA285375), as it contains identifiable information.

BigWig files for H3K27ac were deposited in GEO (GSE125540), as a part of a parallel publication utilizing the same data (Greenwald, W.W. and Li, H. et al. Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. Nat Commun, 2019).

The NKX2-5 ChIP-Seq BigWig files were deposited in GEO, accession GSE133833

Files in database submission

NKX2-5_ChIPSeq_iPSC-CM_iPSCORE_2_1_FS009
NKX2-5_ChIPSeq_iPSC-CM_iPSCORE_2_3_FS010
NKX2-5_ChIPSeq_iPSC-CM_iPSCORE_2_3_FS003
NKX2-5_ChIPSeq_iPSC-CM_iPSCORE_2_6_FS018_A
NKX2-5_ChIPSeq_iPSC-CM_iPSCORE_2_2_FS005_A
NKX2-5_ChIPSeq_iPSC-CM_iPSCORE_2_4_FS015_A
NKX2-5_ChIPSeq_iPSC-CM_iPSCORE_2_9_FS008
NKX2-5_ChIPSeq_iPSC-CM_iPSCORE_2_1_FS016
NKX2-5_ChIPSeq_iPSC-CM_iPSCORE_2_7_FS011
NKX2-5_ChIPSeq_iPSC-CM_iPSCORE_2_4_FS015_B
NKX2-5_ChIPSeq_iPSC-CM_iPSCORE_2_3_FS024
NKX2-5_ChIPSeq_iPSC-CM_iPSCORE_2_9_FS007
NKX2-5_ChIPSeq_iPSC-CM_iPSCORE_2_6_FS018_B
NKX2-5_ChIPSeq_iPSC-CM_iPSCORE_2_7_FS014
NKX2-5_ChIPSeq_iPSC-CM_iPSCORE_2_2_FS005_B
NKX2-5_ChIPSeq_iPSC-CM_Pool_Input control
036253ba-25d4-4b93-9127-9d21527a082_iPSC-CM (ChIP-seq)
03a95c32-f31e-4846-b4e7-6991a7bdc686_iPSC (ChIP-seq)
0441cd83-9d1a-416b-a82d-856ff0d4f6e4_iPSC-CM (ChIP-seq)
079927aa-29e1-462c-8b15-af1a3d8f1b20_iPSC (ChIP-seq)
093b028a-b5e8-496e-b2cb-aaee6d4d3a6b_iPSC (ChIP-seq)
190c066f-17f7-48af-92f1-a50607fe9a3f_iPSC-CM (ChIP-seq)
175c52e-c463-47ab-b388-0fd7f5f5e24_iPSC-CM (ChIP-seq)
20f37c64-0cb6-4cb7-9501-612d1d01b84_iPSC-CM (ChIP-seq)

(Q9QVP4). No signal: chicken. Referenced in 36 publications.

anti-H3K27ac: from company website: Suitable for: IHC-Fr, ICC/IF, WB, IHC-P, ChIPseq, Chip/Chip, ChIP, PepArr. Reacts with: Mouse, Rat, Chicken, Cow, Human, Arabidopsis thaliana, Drosophila melanogaster, Monkey, Zebrafish, Plasmodium falciparum, Rice, Cyanidioschyzon merolae. Referenced in 766 publications.

anti- NKX2-5: from company website: recommended for detection of Nkx-2.5 and, to a lesser extent, Nkx-2.3 of mouse, rat and human origin by WB, IP, IF and ELISA; also reactive with additional species, including and canine, bovine and porcine. Referenced in 30 publications. Nkx-2.5 (N-19) has been discontinued and replaced by Nkx-2.5 (A-3): sc-376565

Additionally, anti-NKX2-5 (lot C0113) was further validated in our laboratory by Chip followed by Western Blot in iPSC-CMs.
Methodology

Replicates
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) as detailed in Supplementary Tables 2 and 3.

The median pairwise Spearman correlation coefficient across all samples and across replicate samples of the same individual were, respectively:
- 0.97 and 0.98 for H3K27ac in iPSCs
- 0.93 and 0.93 for H3K27ac in iPSC-CMs
- 0.79 and 0.79 for NKX2-5

Sequencing depth
All details on individual sequencing metrics are provided in Supplementary table 2 and summarized in Supplementary table 3. For NKX2-5 the average depth was 26 M uniquely mapped reads while for H3K27ac was 27 M for iPSC-CMs, and 40 M for iPSCs. We used in 100 PE reads in most of cases, and 150 PE in few cases.

Antibodies
- anti-H3K27ac (Abcam ab4729, lots GR183922-2 (1.75 μg), GR184333-2 (1 μg), or GR00324078 (1 μg))
- anti- NKX2-5 (Santa Cruz Biotechnology, sc-8697x, lot C0113, 5 μg)

Peak calling parameters
Peak calling was performed using MACS2 ('macs2 callpeak -f BAMPE -g hs --SPMR --verbose 3 --cutoff-analysis --call-summits -q 0.01') with reads derived from sonicated chromatin not subjected to IP (i.e. input chromatin) from a pool of samples used as negative control. Peak coordinates were called from combined samples of either iPSCs or iPSC-CMs, generated by pooling the BAM files of each data type across all samples of the given cell type.

Data quality
All peaks provided and utilized in the analyses were FDR (q) < 0.01
Motif enrichment analysis was performed using HOMER 'findMotifsGenome.pl' and, for NKX2-5, also using MEME ChIP.
Motif analysis of the NKX2-5 ChIP-Seq confirmed a significant enrichment (binomial test, q-value <0.0001) for the NKX2-5 homeobox motif, as well as for the motifs of other heart development TFs (GATA4, TBX5, TBX20, MEF2A/C and MEIS1).

Software
- Alignment: BWA
- Peak calling: MACS2
- Quantification of signal: featureCounts
Motif enrichment: HOMER and MEME
Allele-specific effect analysis: WASP, MBASED, GATK
Postprocessing analyses and data manipulation: R custom codes found at https://github.com/frazer-lab