

ARTICLE

Received 19 May 2016 | Accepted 16 Aug 2016 | Published 5 Oct 2016

DOI: 10.1038/ncomms12923

OPEN

# Genome-wide compendium and functional assessment of *in vivo* heart enhancers

Diane E. Dickel<sup>1</sup>, Iros Barozzi<sup>1</sup>, Yiwen Zhu<sup>1</sup>, Yoko Fukuda-Yuzawa<sup>1</sup>, Marco Osterwalder<sup>1</sup>, Brandon J. Mannion<sup>1</sup>, Dalit May<sup>1,†</sup>, Cailyn H. Spurrell<sup>1</sup>, Ingrid Plajzer-Frick<sup>1</sup>, Catherine S. Pickle<sup>1</sup>, Elizabeth Lee<sup>1</sup>, Tyler H. Garvin<sup>1</sup>, Momoe Kato<sup>1</sup>, Jennifer A. Akiyama<sup>1</sup>, Veena Afzal<sup>1</sup>, Ah Young Lee<sup>2</sup>, David U. Gorkin<sup>2</sup>, Bing Ren<sup>2,3</sup>, Edward M. Rubin<sup>1,4</sup>, Axel Visel<sup>1,4,5</sup> & Len A. Pennacchio<sup>1,4</sup>

Whole-genome sequencing is identifying growing numbers of non-coding variants in human disease studies, but the lack of accurate functional annotations prevents their interpretation. We describe the genome-wide landscape of distant-acting enhancers active in the developing and adult human heart, an organ whose impairment is a predominant cause of mortality and morbidity. Using integrative analysis of >35 epigenomic data sets from mouse and human pre- and postnatal hearts we created a comprehensive reference of >80,000 putative human heart enhancers. To illustrate the importance of enhancers in the regulation of genes involved in heart disease, we deleted the mouse orthologs of two human enhancers near cardiac myosin genes. In both cases, we observe *in vivo* expression changes and cardiac phenotypes consistent with human heart disease. Our study provides a comprehensive catalogue of human heart enhancers for use in clinical whole-genome sequencing studies and highlights the importance of enhancers for cardiac function.

<sup>1</sup>Functional Genomics Department, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720, USA. <sup>2</sup>Ludwig Institute for Cancer Research, 9500 Gilman Drive, La Jolla, California 92093, USA. <sup>3</sup>Department of Cellular and Molecular Medicine, University of California, San Diego School of Medicine, La Jolla, California 92093, USA. <sup>4</sup>U.S. Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA. <sup>5</sup>School of Natural Sciences, University of California, Merced, Merced, California 95343, USA. † Present address: Department of Family Medicine, Clalit Health Services, The Hebrew University-Hadassah Medical School, 91120 Jerusalem, Israel. Correspondence and requests for materials should be addressed to D.E.D. (email: dedickel@lbl.gov) or to A.V. (email: avisel@lbl.gov) or to L.A.P. (email: lapennacchio@lbl.gov).

Cardiovascular disease is the most common cause of death worldwide<sup>1</sup>. Diseases of the heart include a spectrum of adult-onset conditions, as well as congenital phenotypes that collectively represent the most common category of severe birth defects<sup>2</sup>. Causes of heart disease include environmental risk factors<sup>1</sup>, common variants with moderate effect sizes<sup>3</sup>, and rare and *de novo* mutations that cause familial cases with Mendelian inheritance patterns<sup>3</sup>. In particular for the latter category, candidate gene sequencing has proven powerful for obtaining molecular diagnoses. For example, for familial hypertrophic cardiomyopathy, candidate gene sequencing identifies a clear genetic cause in ~60% of patients<sup>4</sup>. Nevertheless, this approach is by design limited to the coding sequence of candidate genes and fails to identify non-coding mutations. Whole-genome sequencing (WGS) can in principle detect non-coding mutations and is becoming increasingly adopted for patients with unexplained heart disease<sup>5</sup>. However, early WGS studies illustrate major challenges in the interpretation of non-coding variants, and particularly of rare non-coding variants<sup>6</sup>. In the absence of accurate annotations linking non-coding loci to *in vivo* functions, non-coding WGS findings are largely uninterpretable and, thereby, most cases with non-coding mutations remain unresolved.

To address the pressing need for a high-quality, genome-wide annotation of functional non-coding sequences active in the developing and adult heart, in the present study we describe a comprehensive catalogue of more than 80,000 candidate distant-acting cardiac enhancers (Fig. 1). Enhancers are a major category of non-coding regulatory elements that activate gene expression from a distance in a cell type-specific<sup>7</sup> and temporally restricted<sup>8</sup> manner. They are hypothesized to play a major role in development and disease, and sequence variants that alter enhancer function are associated with a variety of human phenotypes (for example, refs 9–11). We derived the heart enhancer compendium from more than three dozen epigenomic data sets mapping enhancer-associated chromatin marks in developing and adult heart tissue from mice and humans. This catalogue of human heart enhancers can be easily and immediately implemented in human disease studies, and to further facilitate its utilization in clinical studies, we provide confidence scores for each predicted enhancer that correlate strongly with *in vivo* validation rates. We find that more than 2,000 human variants implicated in heart-related phenotypes through genome-wide association studies (GWAS), either directly as lead variants or indirectly by linkage disequilibrium (LD), fall into putative heart enhancers. Anticipating downstream validation of WGS studies, which will likely focus first on regulatory sequences near genes already implicated in disease, we experimentally validated putative enhancers and provide *in vivo* characterization of more than 20 novel cardiovascular enhancers near known heart disease genes. Finally, as there remains a limited understanding of the general phenotypic impact of lost or impaired enhancer function, we deleted two enhancers near heart disease genes in mice. In both cases, we observed loss of target gene expression, as well as cardiac phenotypes consistent with heart disease in humans. Our results highlight the functional importance of enhancers for normal heart function, as well as the potential contribution of enhancer mutations to heart disease.

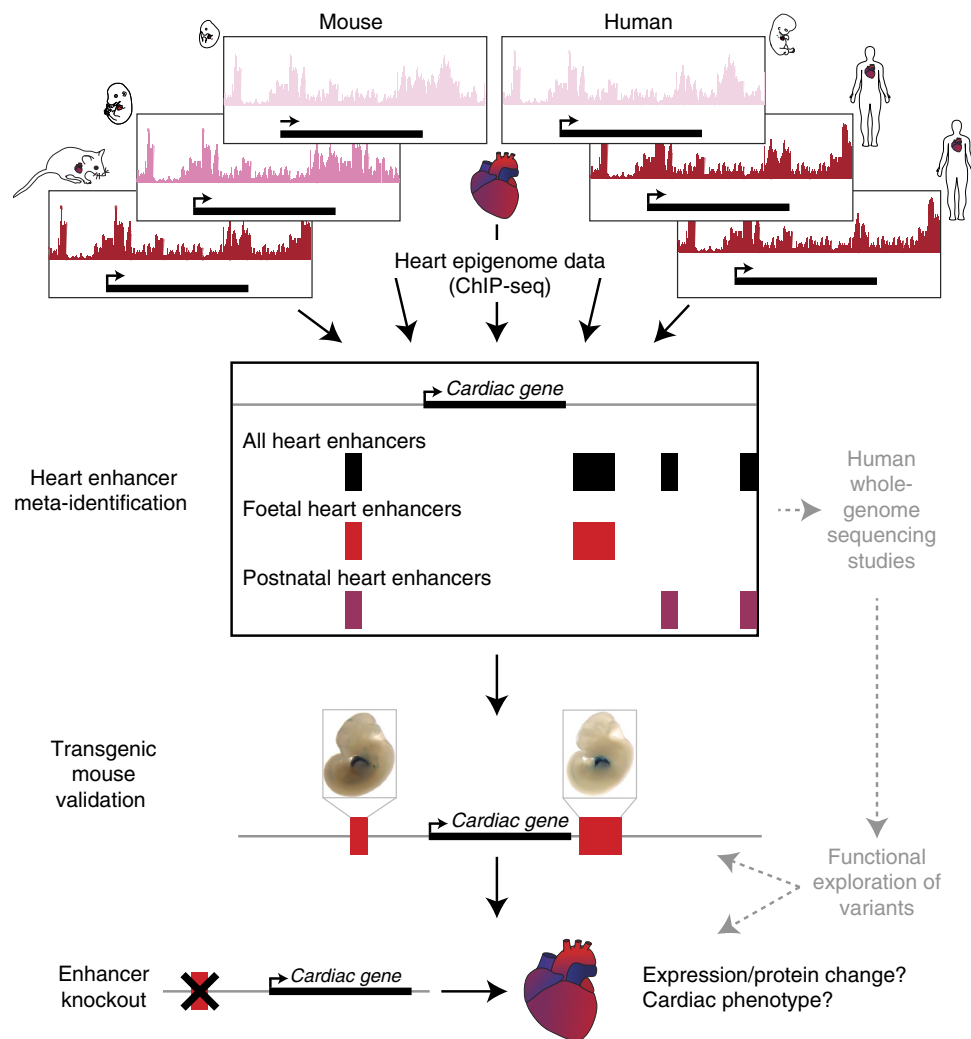
## Results

**Genome-wide mapping of heart enhancers.** Genome-wide profiling of enhancer-associated proteins and histone modifications such as p300/CBP or H3K27ac via chromatin immunoprecipitation (ChIP)-seq directly applied to primary tissue is a powerful approach for the identification of *in vivo* enhancers<sup>12,13</sup>.

Initial application of this technique to cardiac tissue samples established the general utility of the method for the identification of heart enhancers, but detected only modestly sized sets of candidate enhancers due to limited sampling<sup>14,15</sup>. To generate a comprehensive genome-wide catalogue of cardiac enhancers in the human genome that can easily be incorporated into human disease studies, we integrated epigenomic data from multiple developmental stages and all major anatomical subregions of the heart. In total, we examined >35 genome-wide p300/CBP and/or H3K27ac profiles from different *ex vivo* cardiac tissue samples, nearly all of which are known to be or presumed to be normal (see Table 1 and Supplementary Table 1 for tissue information and data sources)<sup>8,14–19</sup>. The sampled conditions include prenatal human heart, major anatomical subregions of childhood and adult human hearts, and a closely spaced developmental time series of prenatal and postnatal mouse heart (Table 1). Although the inclusion of mouse samples could create bias towards the prediction of more highly conserved enhancers (see Supplementary Note 1), it allows for better discovery of enhancers active exclusively during prenatal development, a timespan for which there is only a single human data set considered for analysis. To enable integrative analysis across samples and antibodies, all raw data was analysed using a uniform processing pipeline (see the ‘Methods’ section, Supplementary Fig. 1). Because H3K27ac and p300 are associated with both enhancer and promoter sequences<sup>20,21</sup>, we excluded peaks overlapping promoters, defined as those centred within 1.5 kb of a gene transcription start site (TSS; promoters and their scores are included separately in Supplementary Data 1 but were not further evaluated). Peaks identified in mouse samples were mapped to the human genome (see Supplementary Table 2). Merging the peaks from all data sets resulted in a single list of 82,119 unique candidate heart enhancers in the human genome (Supplementary Data 2). Nearly all (>95%) of the putative enhancers were <10 kb in size (Supplementary Fig. 2a), and most (85%) were smaller than 5 kb. A total of 3,677 candidate enhancer regions >10 kb were identified, consistent with previous observations of subsets of very long enhancers that may form central nodes of tissue-specific regulatory networks (Supplementary Note 2)<sup>22</sup>.

Two initial lines of evidence support that this catalogue is comprised of cardiac-specific enhancers. First, gene ontology analysis<sup>23</sup> showed that the identified regions are highly enriched near genes with relevant functions (Supplementary Table 3). For example, 11 out of the 12 most enriched human phenotypes represent cardiovascular conditions (Table 2). In addition, 152 out of 170 (89%) heart enhancers reported in the VISTA Enhancer Browser, a large collection of *in vivo* enhancers validated in transgenic mice<sup>24</sup>, are identified as candidate heart enhancers in the present catalogue (Fig. 2a, Supplementary Fig. 2b, Supplementary Data 2). This high rediscovery rate highlights the sensitivity of the approach and suggests that most known heart enhancers are recovered by the integrative analysis in the present study.

Although p300 and H3K27ac are strong predictors of enhancer activity, there have been recent reports of improved enhancer prediction using various additional criteria (for example, transcription factor (TF) binding and DNase hypersensitivity (DHS))<sup>25</sup> and/or supervised analysis methods trained on experimentally validated enhancer sets<sup>26,27</sup>. To assess whether additional types of data sets or methods would substantially improve enhancer prediction, we first performed a similar integrative analysis using available human and mouse heart DHS and TF data sets (data sets and their references are listed in Supplementary Table 4). Overall, DHS and TF ChIP-seq identify sets of loci that substantially overlap those captured by H3K27ac



**Figure 1 | Generation and validation of a genome-wide cardiac enhancer catalogue.** Integrative analysis of >35 epigenomic data sets from *ex vivo* human and mouse heart tissue resulted in a catalogue of >80,000 putative human heart enhancers. We demonstrate the utility of this catalogue for the discovery of enhancers near heart disease-associated genes by characterizing the *in vivo* activity patterns of 22 novel cardiovascular enhancers in transgenic mouse assays. We also show the functional importance of enhancers by deleting two cardiac enhancers, which resulted in reduced gene expression and impaired cardiac function.

and p300 (Supplementary Fig. 3a,b). Those sites displaying DHS or TF binding in the absence of H3K27ac or p300 are not associated with cardiac-specific function, suggesting they do not identify cardiac enhancers missed by our integrative H3K27ac and p300 analysis (Supplementary Fig. 3c,d). We additionally found that our unsupervised approach performs with an accuracy similar to EMERGE (ref. 26) and EnhancerFinder (ref. 27), supervised methods for enhancer prediction (Supplementary Table 5).

Our analysis, using a large number of data sets and relatively permissive criteria for the identification of candidate heart enhancers, identifies a considerable proportion of the human genome (264 Mb or ~8% of the total genome) as potential enhancers under at least one of the conditions studied. However, individual predicted enhancer sequences vary substantially in the strength of the supporting evidence, which includes the intensity of ChIP-seq signal in individual data sets (peak scores), as well as recurrent observation of the same peak across multiple cardiac source tissues (Fig. 2b, top). We hypothesized that these differences provide a means to distinguish higher- from lower-confidence predictions, allowing for rational experimental prioritization of candidate sequences. Such evidence-based

ranking is critical because WGS studies identify far more rare and *de novo* non-coding variants than can be reasonably followed up experimentally<sup>6</sup>, even with emerging high-throughput methods. We developed confidence scores to assess the support for a putative enhancer based on two criteria: (1) the statistical significance of the ChIP-seq enrichment observed at a locus and (2) the number of conditions under which a putative enhancer is observed (Methods section, Fig. 2b,c and Supplementary Data 2, Supplementary Note 3). For each predicted enhancer sequence we provide a combined score, on a scale from 0 (weakest evidence) to 1 (strongest evidence), reflecting the strength of evidence from all data sets combined. Enhancer usage has previously been shown to be temporally dynamic, with the majority of enhancers not active throughout the entirety of development<sup>8</sup>. Therefore, we have also provided scores from prenatal data alone and postnatal data alone. These stage-specific scores may be more applicable for studies, respectively, of congenital heart diseases, which typically present around birth, and phenotypes, such as coronary artery disease, that have adult onset.

We assessed the validity and utility of the confidence scores by comparing the genome-wide compendium against the collection of >2,000 experimentally tested enhancers available in the

**Table 1 | Data sets included in the integrative epigenomic analysis.**

Species/stage/anatomical region	Enhancer mark	References
<i>Human</i>		
Prenatal (16 week gestation), whole heart	p300/CBP	14
3 years, left ventricle	H3K27ac	16
3 years, right ventricle	H3K27ac	16
30 years, aorta	H3K27ac	16
34 years, aorta	H3K27ac	16
34 years, left ventricle	H3K27ac	16
34 years, right ventricle	H3K27ac	16
34 years, right atrium	H3K27ac	16
45 years, septum	p300/CBP	14
<i>Mouse</i>		
E11.5, whole heart	H3K27ac, p300	8,15, Unpublished (ENCODE)
E13.5, whole heart	H3K27ac	Unpublished (ENCODE)
E14.5, whole heart	H3K27ac	8,17,18
E15.5, whole heart	H3K27ac	Unpublished (ENCODE)
E16.5, whole heart	H3K27ac	Unpublished (ENCODE)
E17.5, whole heart	H3K27ac	8
P0, whole heart	H3K27ac	8, Unpublished (ENCODE)
P5, whole heart	p300	19
P7, whole heart	H3K27ac	8
P21, whole heart	H3K27ac	8
P56, whole heart	H3K27ac, p300	8,17,18

Epigenomic profiles for enhancer-associated marks (H3K27ac and p300/CBP) are from human and mouse heart tissue at a variety of developmental stages. ENCODE indicates previously unpublished data from the ENCYclopedia of DNA Elements project. Developmental stages for mouse are given in embryonic days post-conception (E) or days after birth (P). With the exception of the human septum sample, all tissue samples were reported to be normal, or no explicit statement of health was provided in the referenced publication.

VISTA Enhancer Browser<sup>24</sup>. Many of the previously validated heart enhancers were retrospectively confirmed to be among the highest-ranking scored loci in the present study (Fig. 2d), and validation rates correlated positively with confidence scores (Fig. 2c, also see Supplementary Note 3, Supplementary Fig. 4, Methods section). In total, this retrospective analysis shows that the scoring scheme is a good indicator of the likelihood that a putative enhancer is active *in vivo*.

As an example of how this cardiac compendium can be integrated into human disease studies, we intersected it with variants associated with a variety of heart-related phenotypes reported in the NHGRI-EBI GWAS Catalog<sup>28</sup>. Including all variants in strong LD with the reported lead variants ( $r^2 \geq 0.8$ ), more than 18,000 sequence variants are associated with human heart phenotypes. Approximately 2,300 of these fall within a predicted heart enhancer, with ~900 in predicted enhancers that have a score of at least 0.2 (ranked list and corresponding heart enhancer scores provided in Supplementary Data 3). This includes a variety of loci where protein-altering variants have not been identified to explain the association signal, such as (1) the 6q22 region near *GJA1* that has been strongly implicated in heart rate<sup>29</sup>, (2) the 1p32 region overlapping *PLPP3* implicated in coronary artery disease<sup>30</sup> and (3) the 15q24 locus containing *HCN4* implicated in atrial fibrillation<sup>31</sup> (Supplementary Fig. 5). In the case of the *GJA1* locus, more than 200 single-nucleotide polymorphisms (SNPs) fall within multiple phenotype-associated LD blocks. Intersecting those SNPs with the heart enhancer

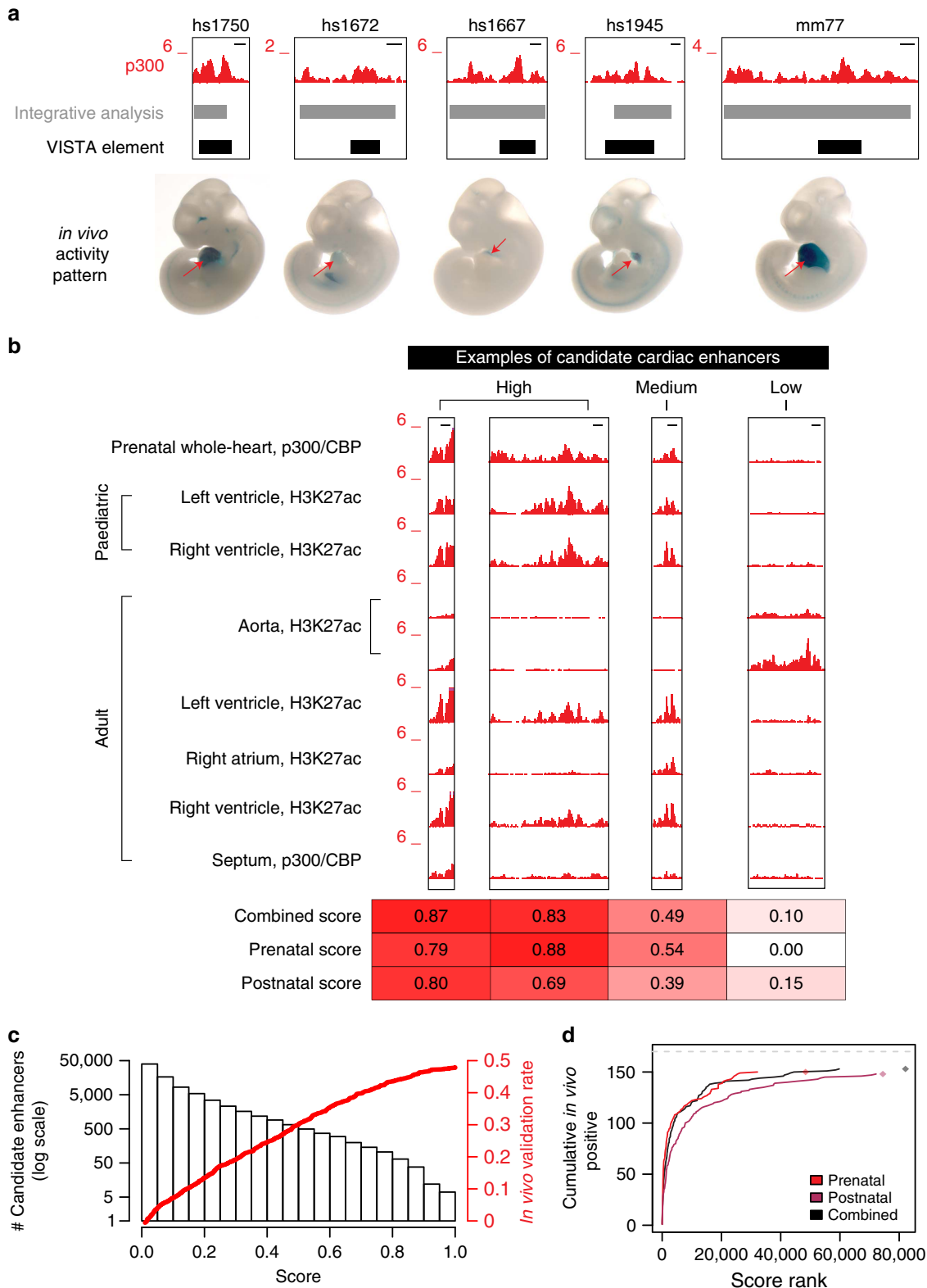
**Table 2 | Top enriched human phenotypes of putative target genes near predicted heart enhancers.**

Top enriched phenotypes	Binomial FDR Q-value	Binomial fold enrichment
Cardiac arrest	$6.81 \times 10^{-89}$	2.0
Sudden cardiac death	$7.84 \times 10^{-89}$	2.0
Sudden death	$3.67 \times 10^{-84}$	3.1
Syncope	$1.22 \times 10^{-64}$	2.2
Atrial fibrillation	$8.88 \times 10^{-62}$	2.0
Abnormal EKG	$4.89 \times 10^{-44}$	2.0
Ventricular tachycardia	$6.28 \times 10^{-44}$	2.6
Aortic dissection	$9.09 \times 10^{-43}$	2.2
Bicuspid aortic valve	$1.60 \times 10^{-39}$	2.2
Pointed chin	$9.10 \times 10^{-35}$	2.0
Aortic aneurysm	$1.13 \times 10^{-34}$	2.3
Prolonged QT interval	$3.91 \times 10^{-34}$	2.2

EKG, electrocardiogram; FDR, false discovery rate corrected. Additional highly enriched terms are included in Supplementary Table 3.

compendium identified 11 variants that fall within predicted enhancers, including one in a very high-scoring element (score = 0.719). This is a number that is tractable to downstream experimental validation. In the case of the *HCN4* locus, the strongest scoring putative enhancer in the region (score = 0.313) includes a SNP (rs17172038) in perfect LD ( $r^2 = 1$ ) with the reported lead variant (rs7164883), and we included this enhancer in the functional validation performed below. Overall, the heart enhancer compendium can be easily intersected with human disease data to identify strong candidates for further experimental validation and should help prioritize enhancer sequence variants found in human resequencing studies.

***In vivo* validation of heart enhancers near disease genes.** Given the large number of rare and *de novo* non-coding sequence variants that are identified in WGS studies, initial analyses will likely focus on characterizing variants in putative enhancers near genes previously implicated in heart disease. To explore the utility of the heart enhancer compendium for identifying *in vivo* enhancers near genes of interest, we examined 58 candidate sequences in transgenic mouse enhancer assays<sup>32,33</sup>. The selected regions spanned the full range of combined cardiac integrative analysis scores and were all located within 100 kb of a heart disease-associated gene (Supplementary Data 4). In addition to the close proximity, there is evidence from the chromatin–chromatin spatial interaction database<sup>34</sup> that most of the tested sites physically interact with the putative target gene (Supplementary Data 4). In total, 22 of the tested sequences drove reproducible reporter gene expression in the developing heart or blood vessels at embryonic day 11.5 (E11.5) (Fig. 3, Supplementary Data 4 and Supplementary Table 6), including the enhancer upstream of *HCN4* that overlaps with a variant associated with atrial fibrillation. This enhancer had strong activity in the ventricles and weaker but reproducible activity in the atria (Fig. 3e). Examples of additional newly identified *in vivo*-validated cardiac enhancers include elements near *TGFB3* (Fig. 3c) and *PRKAG2* (Fig. 3d), two genes previously implicated in heart disease<sup>35,36</sup> but without known associated distant-acting heart enhancers. Furthermore, we characterized new enhancers near genes such as *GATA4* (Fig. 3f, Supplementary Table 6), where several *in vivo* enhancers had been previously identified<sup>15,37</sup>. The overall transgenic validation rate (38% with cardiovascular activity) is somewhat lower than that reported by previous heart enhancer validation efforts. This is in large part due to our inclusion of candidate sequences from across the



**Figure 2 | Integrative analysis of heart ChIP-seq data identifies >80,000 heart enhancers.** (a) Examples of heart enhancers identified through this analysis that had previously been validated in transgenic mouse assays<sup>24</sup>. For each locus, we show the raw p300/CBP ChIP-seq signal from foetal human heart, the locations of the integrative analysis-identified enhancer and the tested element, and a representative E11.5 embryo. The heart is indicated by a red arrow. hs/mm numbers indicate the VISTA identifier. Scale bars, 1 kb. (b) Representative examples of putative enhancers with high, medium and low levels of support. Raw ChIP-seq data from all human heart samples analysed is shown, along with the corresponding confidence scores for each locus. (c) Histogram of combined scores for all 82,119 putative heart enhancers identified by the integrative analysis (left axis). Superimposed (red) is the retrospective heart enhancer validation rate versus combined score for all *in vivo* tested sites that overlap scored regions (see the ‘Methods’ section). (d) Cumulative proportion of *in vivo*-validated heart enhancers captured by score rank. Dashed gray line indicates the total number of validated heart enhancers considered (170). Rhombi indicate the total number of scored loci in each category.

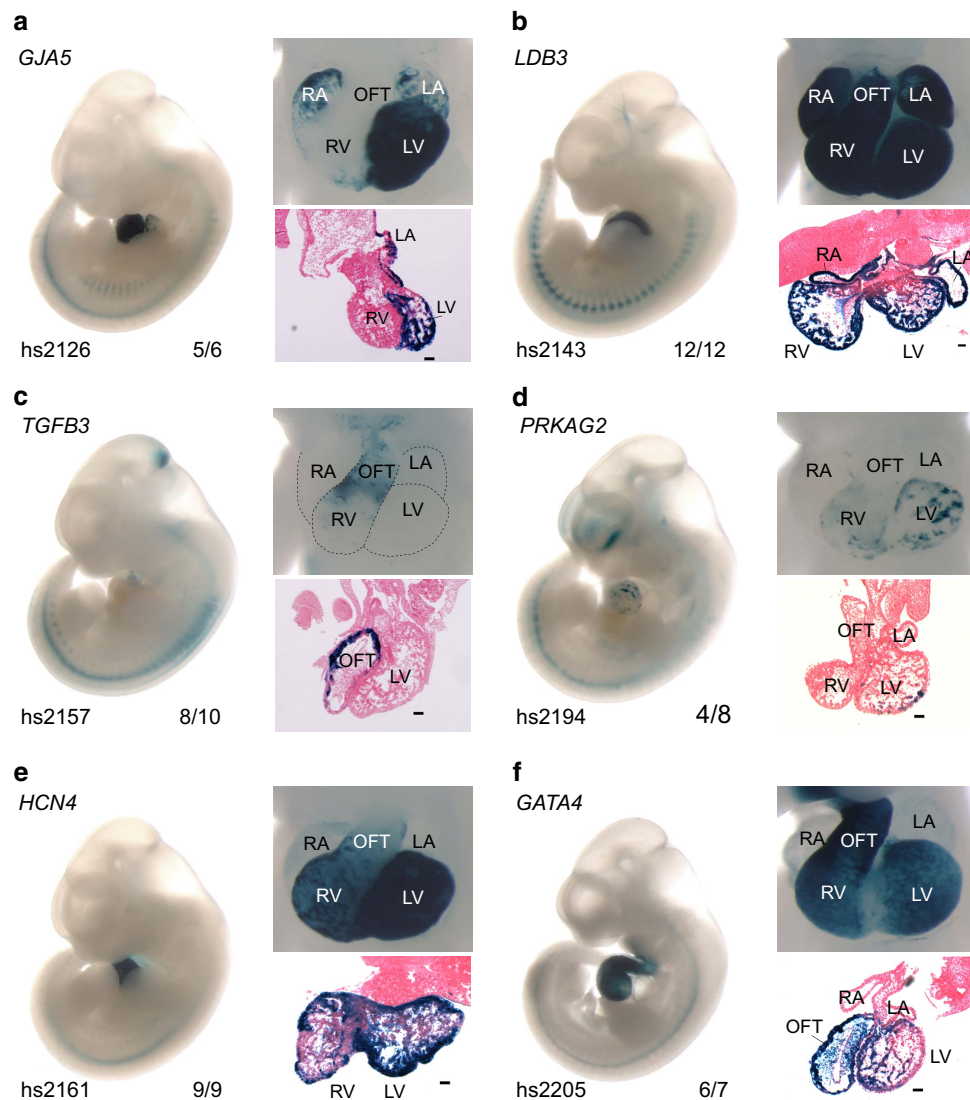


scoring range, in contrast to previous efforts, which tended toward validating loci with high ChIP-seq enrichment scores<sup>14,15</sup>. However, as expected, elements that were confirmed as *in vivo* cardiovascular enhancers had a higher average combined confidence score than those for which no cardiovascular activity was observed (mean combined scores: cardiovascular positive = 0.509, cardiovascular negative = 0.385,  $P = 0.021$ , one-tailed  $t$ -test; Supplementary Data 4). These results illustrate the utility of the genome-wide scored data sets for identification of *in vivo* cardiac enhancers near heart disease genes of interest.

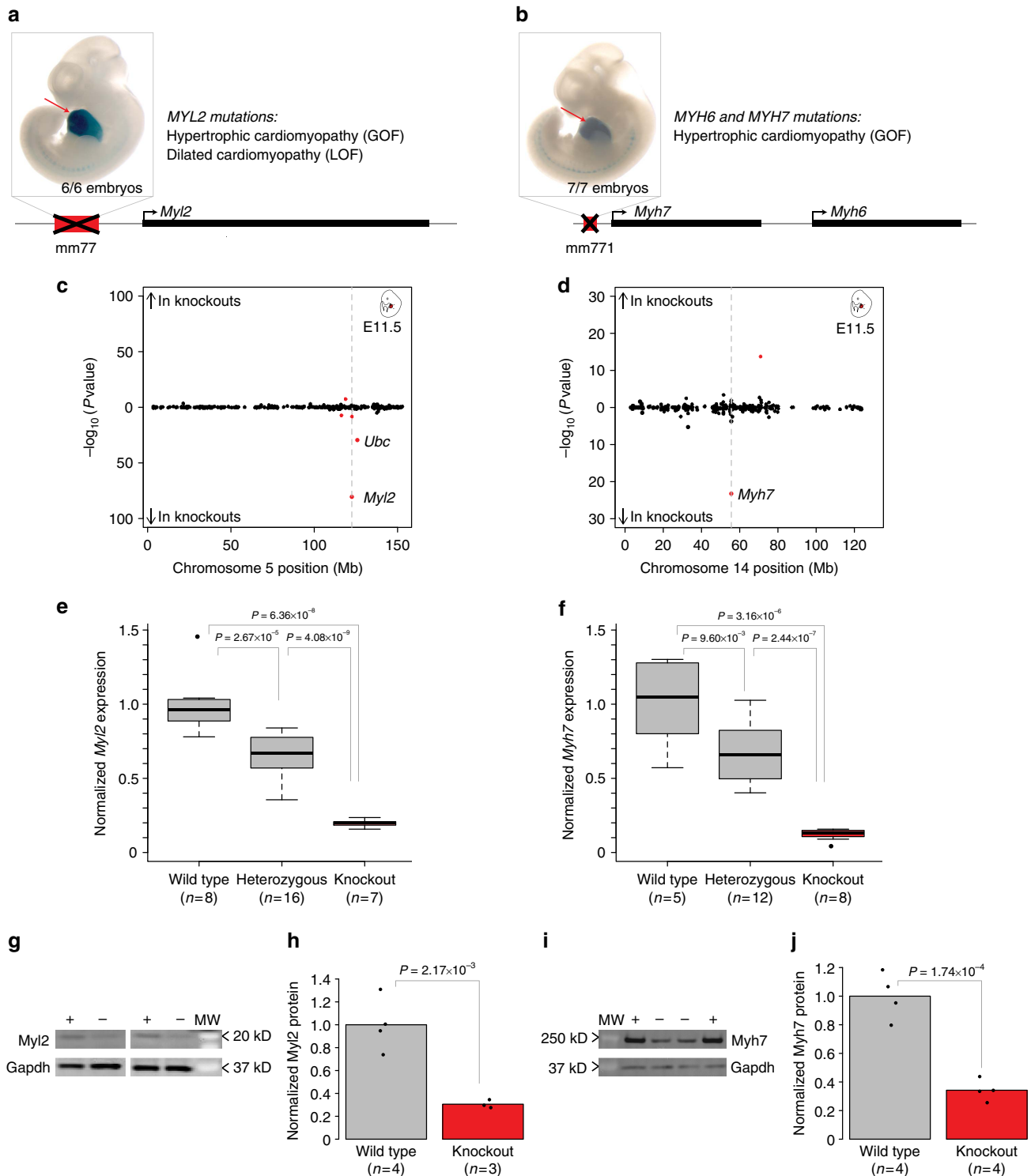
**Heart enhancer deletions result in cardiac dysfunction.** While *in vivo* reporter assay validation is a powerful tool to confirm that an enhancer is sufficient to activate tissue-specific gene expression, it does not illuminate whether a sequence is necessary for proper development or health. Such information is, however, crucial to understanding the phenotypic consequences of non-coding mutations in the human genome. To assess the biological necessity of heart enhancers for proper cardiac development and

function, we created knockout mouse models for two different heart enhancers, mm77 and mm771, upstream of *Myl2* and *Myh7*, respectively. Mutations in either of these genes are associated with hypertrophic cardiomyopathy, and their coding sequences are routinely screened for mutations in the clinic<sup>4</sup>. Both enhancers fall in regions with very high combined confidence scores (0.828 and 0.756, respectively). Furthermore, both drive strong, reproducible reporter gene expression throughout the heart at E11.5 and were initially identified from epigenomic data sets considered in the present integrative analysis (Fig. 4a,b, Supplementary Fig. 6). In both cases, the human and mouse orthologous enhancer sequences drive strong and highly reproducible reporter activity in the heart in mouse transgenic assays, indicating functional conservation of these enhancers in mammals (Supplementary Fig. 6).

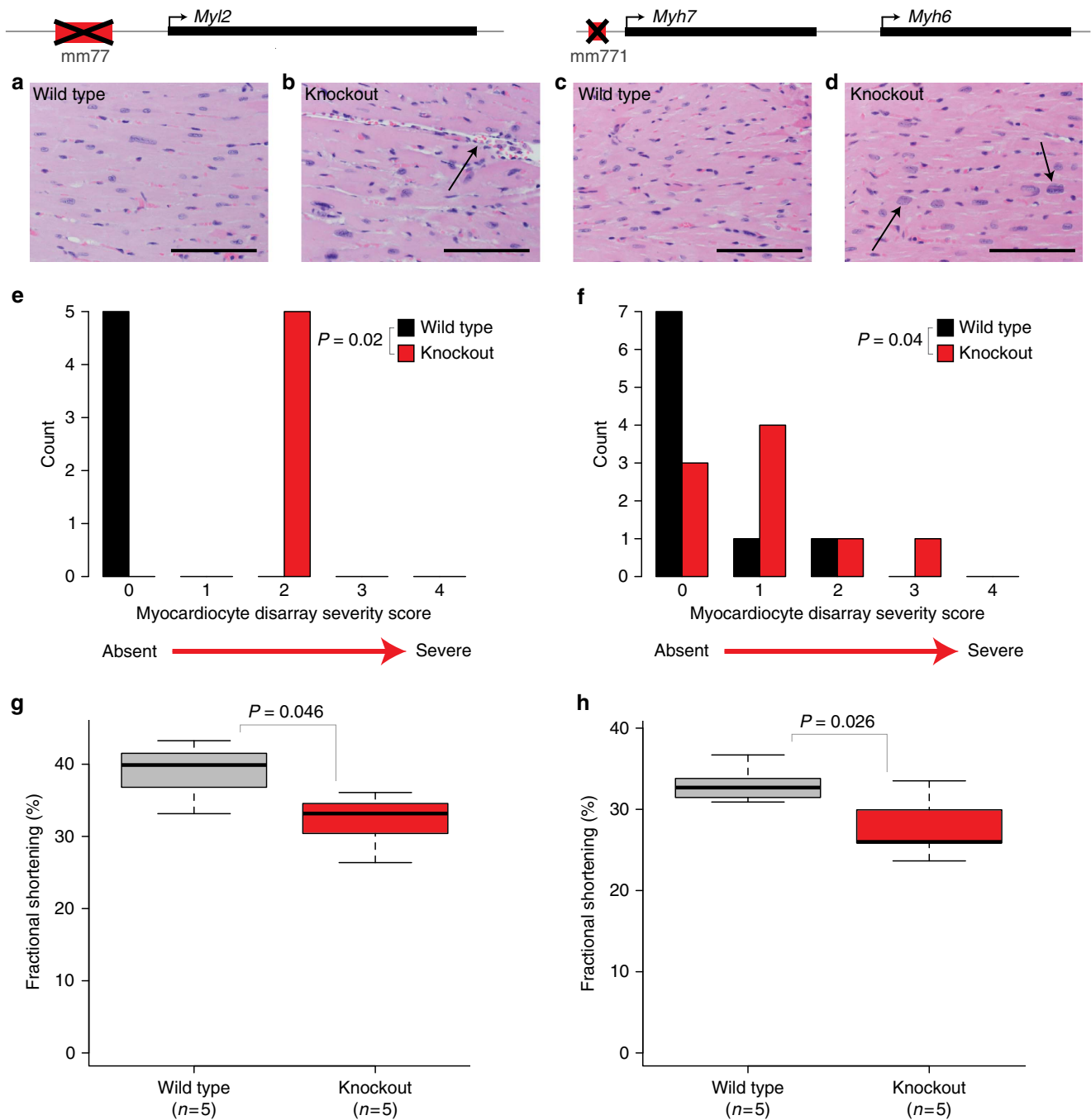
We created two mouse lines, each carrying a deletion of one of these two enhancers (Supplementary Fig. 7). Mice homozygous for either enhancer deletion are born at normal Mendelian ratios and show no gross abnormalities or overt impairments of health (Supplementary Table 7). Adult females heterozygous or



**Figure 3 | *In vivo* activity of human heart enhancers near heart disease genes.** (a–f) Whole mount (left), heart close up (top right) and histological section (bottom right, scale bar, 100  $\mu$ m) of E11.5 mouse embryos expressing a reporter *LacZ* gene (dark blue) under the control of a heart enhancer. VISTA Enhancer Browser identifier (hs number), nearest heart disease-associated gene, and the reproducibility of heart enhancer activity are indicated. LA, left atrium; LV, left ventricle; OFT, outflow tract; RA, right atrium; RV, right ventricle.



**Figure 4 | Cardiac enhancers are required for normal expression of heart disease-associated genes.** Knockout analysis of enhancers mm77 (a,c,e,g,h) and mm771 (b,d,f,i,j). (a,b) Representative transgenic reporter assay results. Heart activity is indicated by red arrows. Numbers indicate reproducibility over total transgenic embryos. Gene models not drawn to scale. LOF: loss-of-function, GOF: gain-of-function. (c,d) Chromosome-wide mRNA expression changes in E11.5 whole heart. Points indicate individual genes, with red indicating statistically significant differences after FDR correction ( $P < 0.01$  using an  $FDR < 5\%$ , see the 'Methods' section for details). Dashed grey line indicates position of enhancer. Mb: megabases. (e,f) Normalized mRNA levels measured by quantitative RT-PCR in E11.5 hearts. Boxplots indicate median and quartile values for each data set; points indicate outliers. (g-j) Representative western blot images (g,i) and normalized protein levels (h,j) in wild-type (+) and homozygous null (-) samples. Gapdh was used as a loading control, MW, molecular weight. (g,i). Bars indicate group means, and points represent biological replicates (h,j). For (e,f,h,j): values were normalized to the wild-type mean, and  $P$  values were calculated using a one-tailed  $t$ -test.



**Figure 5 | Loss of heart enhancers results in cardiac abnormalities.** (a–d) Representative images of haematoxylin and eosin stained heart tissue from mice wild-type (a,c) or homozygous null (b,d) for the mm77 (a,b) or mm771 (c,d) enhancers. (b) Characteristic myocardocyte disarray (arrow) observed in a homozygous  $\Delta mm77$  mouse. (d) Myocardocyte karyomegaly (arrows) observed in a homozygous  $\Delta mm771$  mouse. Scale bars, 100  $\mu\text{m}$ . (e,f) Severity of myocardocyte disarray observed in the hearts of mice wild-type or homozygous null for the mm77 (e) or mm771 (f) enhancer. Cardiac tissue was scored by a genotype-blind pathologist from 0 (absent) to 4 (severe), and  $P$  values calculated by one-tailed paired Wilcoxon rank-sum test. (g,h) Cardiac fractional shortening for mice wild-type and homozygous null (knockout) for the mm77 (g) or mm771 (h) enhancer. For (g,h) Boxplots indicate median, range and quartile values for each data set, and  $P$  values were calculated by one-tailed paired  $t$ -test.

homozygous null for mm771 have 5–10% lower body weight than wild-type females ( $P < 0.05$ , one-tailed  $t$ -test), but no growth phenotypes were observed for mm771 males or for mm77-null mice of either gender (Supplementary Fig. 8).

To assess gene expression changes resulting from the loss of either enhancer, we performed RNA sequencing on heart tissue from E11.5 and adult mice from each enhancer deletion line. In mice homozygous null for mm77, *Myl2* RNA expression is reduced by 75–80% compared with wild-type levels at both E11.5

and in adulthood (both  $P < 4.9 \times 10^{-19}$ , see the ‘Methods’ section, Fig. 4c, Supplementary Fig. 9, Supplementary Data 5), establishing that *Myl2* is a direct regulatory target and that the enhancer is required for normal expression of *Myl2* in both embryonic and adult heart. In addition to *Myl2*, embryonic and adult mm77-null animals show a deficit of *Ubc*, a ubiquitin gene more than 3 Mb away from the enhancer (both  $P < 5.3 \times 10^{-11}$ ). Embryonic  $\Delta mm771$  mice show downregulation of *Myh7* by ~85% compared with wild-type ( $P = 5.7 \times 10^{-24}$ ) but show no



change in *Myh6* or any other gene in *cis*, indicating that the enhancer specifically controls the expression of *Myh7* (Fig. 4d, Supplementary Data 5). *Myh7* is not expressed in postnatal mouse heart<sup>38</sup>, and, consequently, no significant changes to its expression were observed in adult  $\Delta mm771$  mice (Supplementary Fig. 10a, Supplementary Data 5). In all cases, the expression changes observed by RNA-seq were confirmed by targeted quantitative reverse transcriptase PCR (qRT-PCR; Fig. 4e,f, Supplementary Figs 9c and 10b). For the affected myosin genes, gene expression changes are dose dependent, with heterozygotes having a reduction in myosin gene expression by  $\sim 35\%$  compared with wild-type (mm77:  $P = 2.7 \times 10^{-5}$ , mm771:  $P = 9.6 \times 10^{-3}$ , one-tailed *t*-test, Fig. 4e,f). We next examined whether these mRNA expression changes result in decreased myosin protein abundance. Western blot analysis on whole-heart tissue established that both enhancer deletions result in  $\sim 70\%$  reduction of cardiac protein levels (*Myl2*,  $P = 2.1 \times 10^{-3}$ ; *Myh7*,  $P = 1.7 \times 10^{-4}$ , one-tailed *t*-test; Fig. 4g–j, Supplementary Fig. 11). Cumulatively, these results show that mm77 and mm771 are required for normal levels of *Myl2* and *Myh7* protein in the heart.

To evaluate potential cardiac dysfunction resulting from the loss of the *Myl2* or *Myh7* enhancer, we performed histological and pathological analysis on both lines (Supplementary Data 6). In heart tissue from both  $\Delta mm77$  and  $\Delta mm771$  mice, we observed cases of myocardiocyte disarray and karyomegaly (cell nucleus enlargement; Fig. 5a–d). To quantify these observations, we used histological severity scores assigned by a genotype-blind pathologist. Compared with wild-type littermates, we observed substantially increased rates and severity of myocardiocyte disarray both in  $\Delta mm77$  ( $P = 0.02$ , paired one-tailed *t*-test) and  $\Delta mm771$  ( $P = 0.04$ ) mice, as well as karyomegaly in  $\Delta mm77$  ( $P = 0.02$ , Fig. 5e,f, Supplementary Fig. 12). In contrast, no reproducible histological or pathological abnormalities were observed in any other major organ system (Supplementary Data 6). These results support that the deleted enhancers are important for establishing appropriate levels of myosin gene products and for maintaining healthy cellular morphology in the heart.

To further assess the effect of heart enhancer loss on cardiac physiological function, we used echocardiography. Homozygous  $\Delta mm77$  ( $P = 0.046$ , paired one-tailed *t*-test) and  $\Delta mm771$  ( $P = 0.026$ ) mice had modest but significant decreases in fractional shortening (Fig. 5g,h, Supplementary Fig. 13a,b), as well as decreases in ejection fraction for  $\Delta mm771$  ( $P = 0.025$ , Supplementary Fig. 13c–f). These results are consistent with an early-stage cardiomyopathy phenotype, specifically dilated cardiomyopathy (Supplementary Note 4, Supplementary Fig. 14). This conclusion is further supported by adult  $\Delta mm77$  mice showing upregulation of *Nppa* (RNA-seq:  $P = 8.29 \times 10^{-10}$ , see the ‘Methods’ section) and *Nppb* (RNA-seq:  $P = 2.59 \times 10^{-07}$ ) (Supplementary Figs 9b and 15), genes whose upregulation is a common biomarker of heart failure<sup>39</sup>. Overall, the loss of either of the cardiac enhancers results in a large decrease of the mRNA and protein products of a neighbouring cardiac myosin gene, which leads to cardiac cellular abnormalities and reduced heart function. These results illuminate a potential role for enhancer mutations in human heart disease.

## Discussion

Technological advances have enabled routine WGS for the study of human disease. However, initial WGS analyses have focused primarily, or exclusively, on the  $< 2\%$  of the genome that encodes proteins<sup>6,40</sup>. While the strong contribution of non-coding sequences in many human disease traits is now widely recognized<sup>41</sup>, the skew towards coding sequence in the analysis

of WGS data persists because of a lack of annotations and analytical tools to assess whether rare non-coding mutations are associated with phenotypes<sup>6</sup>. In the present study, we begin to fill this void by generating a comprehensive compendium of more than 80,000 enhancers predicted to be active in the developing and adult human heart. Since these sequences represent functional units with defined boundaries, this allows for the binning of rare non-coding variants for association testing, analogous to aggregating coding variants by gene<sup>42</sup>. To facilitate the prioritization of non-coding regions for experimental follow-up studies, each predicted enhancer is provided along with a confidence score summarizing the strength of the supporting epigenomic evidence. While the present study focuses on a single organ system of major epidemiological importance, the approach described here is applicable to nearly all human organ systems. Thus, we expect that similar resources for other phenotype-relevant tissues will serve as a critical foundation for the analysis of genetic data from many additional classes of human disease.

Non-coding sequence annotation is a critical first step for interpreting whole-genome sequence data but is generally insufficient to conclusively implicate specific mutations as causal in disease. One powerful complementary approach is through the use of genome engineering in animal models to formally test the importance of a sequence change *in vivo*. In this study, we illustrate how enhancers identified using genome-wide approaches can be prioritized and assessed for normal organismal function. Focusing on two heart enhancers located near disease-associated genes, we show in mouse models how their loss of function results in cardiac phenotypes. While a few enhancers previously knocked-out in mice have the potential to result in cardiac phenotypes (for example, refs 43,44), to our knowledge, the *Myl2* and *Myh7* enhancers are the first examples whose loss has been shown to result in a phenotype consistent with heart disease. With the recent advances in genome editing technology<sup>45</sup>, it appears likely that *in vivo* engineering approaches will be increasingly used for large-scale modelling of both coding and non-coding human mutations. Overall, this study highlights the important role enhancers play in cardiac health and provides a valuable compendium of human heart enhancers that can be easily integrated into cardiovascular disease studies.

## Methods

**Chip-seq data analyses for heart enhancer prediction.** Short read data of published data sets were retrieved from the Gene Expression Omnibus<sup>46</sup> or Human Roadmap Epigenome website (<http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/>). Unpublished ENCODE data was accessed through the ENCODE Data Coordination Center website (<https://www.encodeproject.org/search/>). Supplementary Table 1 lists details about each sample included in this meta-analysis. In the case of reads stored in the Short Read Archive format<sup>47</sup>, fastq files were obtained using the *fastq-dump* script available in the sratoolkit (v2.4.5).

Alignments to either the mouse (mm10) or human (hg19) reference genomes were performed using Bowtie v0.12.7 (ref. 48). Only reads with a unique match to the genome and showing two or fewer mismatches ( $-m 1 -v 2$ ) were retained. Peak calling was performed using MACS v1.4 (ref. 49) with the following parameters:  $-gsize = mm -bw = 300 -nomodel -shiftsize = 100$ . Whenever available, the experiment-matched input DNA was used as control. When this was not available (Supplementary Table 1), peaks were called without using an input DNA as control with MACS v1.4 parameter *-nolambda*.

The resulting lists of peaks were annotated to the nearest RefSeq (ref. 50) gene TSS in either mm10 or hg19 using HOMER<sup>51</sup>. Enriched regions with their annotated centre within 1.5 kb from any TSS were considered promoters (Supplementary Data 1) and separated from the putative enhancers for the scoring described below. In case of very large regions of enrichment, the centre might happen to lie outside the  $\pm 1.5$  kb from annotated TSSs. To avoid considering them, *subtractBed* (ref. 52) was used to avoid any overlapping sub-interval. mm10 v3 and hg19 v19 Basic Gencode annotations<sup>53</sup> were downloaded from the UCSC Genome Browser<sup>54</sup> on 3 June and 12 July 2015, respectively. Filtering for the TSS-proximal regions of Gencode annotated transcripts also ensured the exclusion of potential promoters of non-coding genes from the final list.

In some cases, multiple H3K27ac data sets were available from independent sources for the same mouse developmental time point, and these samples were treated as biological replicates. Biological replicates were combined using MSPC (ref. 55), using the following parameters  $-r$  biological  $-s$  1E-10  $-W$  1E-6. The confirmed peaks were assigned the best  $P$  value, as defined by MACS, among the overlapping peaks.

**Meta-analysis and annotation of heart enhancers.** The obtained lists of putative enhancer regions for the mouse samples (mm10) were mapped to the human genome (hg19) using liftOver<sup>54</sup> with a requirement that  $\geq 50\%$  of the bases in each region map to human. Beside, each region was required to both uniquely map to hg19, and to uniquely map back to the original region in mm10 (similar to ref. 56). Human, as well as human-lifted mouse, regions were then merged together using mergeBed<sup>52</sup>. The resulting regions were re-annotated to the lowest  $P$  value among those shown by the overlapping regions; only those regions showing at least one peak with a  $P$  value  $\leq 1e-10$  (as defined by MACS) in at least one condition were considered for further analysis.

To exclude potential technical artifacts, published H3K27ac data sets from human cell lines<sup>57</sup> were used to produce a blacklist of regions systematically enriched across all ChIP-seq data sets. A blacklist of human regions generated by the ENCODE Consortium itself was also used<sup>57</sup> (<https://sites.google.com/site/anshulkundaje/projects/blacklists>).

Gene ontology enrichment was performed for the complete list of putative heart enhancers using GREAT<sup>23</sup>.

The positions of the resulting regions were annotated relative to the TSS of: (1) the nearest gene; (2) the nearest gene showing one or more heart-related phenotypes in the Mouse Genome Database<sup>58</sup>; and (3) the nearest gene annotated with a heart-related phenotype in the Human-Phenotype Ontology<sup>59</sup>. These annotations were performed using a custom script and lists of RefSeq genes downloaded from the UCSC Genome Browser 4 May 2015. The manually curated lists of heart-related terms are included in Supplementary Data 7. The putative enhancer regions were also annotated to variants significantly associated to human phenotypic traits—as listed in the GWAS catalogue<sup>28</sup>. Bedmap from the BEDOPS suite<sup>60</sup> was used to perform this annotation step.

**Scoring scheme for putative heart enhancers.** Given one condition (for example, E11.5 mouse H3K27ac), for each given putative enhancer region  $r$  and its MACS-generated  $P$  value  $p$ , the following score  $S$  was calculated:

$$S(r, p) = -\log_{10} \left( \frac{\text{number of enhancer regions with a } P \text{ value } \leq p}{\text{total number of enhancer regions}} \right)$$

Each score represents the probability of observing an equal or better ChIP-seq enrichment (that is, an equal or lower  $P$  value), under a specific condition. For each region, the scores were summed up either across all conditions (to obtain the final combined score) or only across foetal or postnatal conditions (to get the pre- and postnatal scores, respectively). Scores were normalized on a scale of 0.0–1.0 such that the highest scoring enhancer in each score class (combined, prenatal and postnatal) was set to 1.0.

**Intersecting enhancer predictions with VISTA.** Regions experimentally tested for *in vivo* enhancer activity in mouse transgenic assays were downloaded from the VISTA enhancer browser (<http://enhancer.lbl.gov/>)<sup>24</sup> on 24 September 2015. Only VISTA elements that could be mapped to hg19 were considered, and those that overlapped promoters or blacklisted sequences (as defined in the meta-analysis section above) were excluded. The overlap among VISTA elements and the putative heart enhancers resulting from the meta-analysis was assessed using coverageBed<sup>52</sup>. Only regions and VISTA elements intersecting at least 500 bps were considered for further analyses. The overlapping VISTA elements (both positive and negative for enhancer activity) were then ranked based on the score, and a cubic spline ( $df = 3$ ) was fit using the *splineFit* function available in R. This way, a curve estimating the value of *in vivo* validation rate across the whole spectrum of scores over the validated elements was obtained. These values were then used to extrapolate the curve that directly relates the *in vivo* validation rate to the score of all regions in the meta-analysis. This procedure was repeated separately for each of the three scores.

**Comparison with other enhancer prediction methods.** We compared our unsupervised enhancer prediction against two previously reported methods, EnhancerFinder<sup>27</sup> and EMERGE<sup>26</sup>. Unless otherwise noted (Supplementary Table 5), area under the curve for receiver operating characteristic curves for EnhancerFinder and EMERGE results are those reported in the corresponding papers for the prediction of mammalian heart enhancers. All other area under the curve values were generated as described in the EMERGE paper<sup>26</sup>.

**Intersecting heart enhancer compendium with GWAS catalogue.** We retrieved all phenotype-associated SNPs in the NHGRI-EBI GWAS catalogue<sup>28</sup> implicated in heart-related traits (search term ‘heart’). The SNP Annotation and Proxy Search (SNAP)<sup>61</sup> tool was used to identify all SNPs in strong LD using the following

parameters: 1,000 Genomes Pilot 1 data set,  $r^2 \geq 0.8$ , the Northern Europeans from Utah (CEU) population panel, 500 kb distance limit.

**Transgenic mouse assays.** Enhancer and allelic variant names (mm and hs numbers) used in this study are the unique identifiers used in the VISTA Enhancer Browser (<http://enhancer.lbl.gov/>). Enhancer sequences were amplified from human (hs numbers) or mouse (mm numbers) genomic DNA and cloned into an hsp68-lacZ expression vector as previously described<sup>33</sup>. Genome coordinates and primer sequences for all elements are listed in Supplementary Data 4 and Supplementary Table 8. Transgenic mouse assays were performed as previously described<sup>32,33</sup>, and results for mm77 (ref. 15) and the human ortholog of mm771 (hs1670)<sup>62,63</sup> were previously reported. To determine if tested enhancers physically contacted the putative target genes, we queried all available human and mouse data sets in the chromatin–chromatin spatial interaction database<sup>34</sup>. To avoid considering interactions that could have been annotated just due to physical proximity, enhancer-promoter pairs closer than 10 kb were not annotated (indicated as ‘Close’ in Supplementary Data 4).

**Generation of enhancer knockout mice.** Enhancer null lines were generated via homologous recombination (Supplementary Fig. 7a) as previously described<sup>62</sup>. Primer sequences used for generating, validating (Supplementary Fig. 7b,c) and genotyping (Supplementary Fig. 7d,e) targeted ES cell lines and mice are listed in Supplementary Table 9. The mm771 deletion removed 332 bp (mm10 chr14:54,996,893–54,997,224) and the mm77 deletion removed 2,517 bp (mm10 chr5:122,092,252–122,094,768) of non-coding sequence.

**Gene expression analysis for enhancer knockouts.** RNA was isolated from whole hearts microdissected from multiple litters of embryonic mice using the Ambion RNAqueous Total RNA Isolation Kit (Life Technologies) according to manufacturer instructions. Matched genomic DNA was collected for each embryo from limb or tail tissue as previously described<sup>62</sup>. Each embryo’s enhancer genotype was assessed using the mm77 or mm771 genotyping primers listed in Supplementary Table 9. For adult mice, RNA was extracted from the ventricular portion of the heart using physical homogenization and the TRIzol Reagent (Life Technologies).

For RNA-seq, RNA samples were DNase-treated with the TURBO DNA-free Kit (Life Technologies), and RNA quality was then assessed using a 2100 Bioanalyzer (Agilent) with an RNA 6,000 Nano Kit (Agilent). RNA sequencing libraries were made using the TruSeq Stranded Total RNA with Ribo-Zero Human/Mouse/Rat kit (Illumina) or the TruSeq Stranded mRNA Sample Prep Kit (Illumina) according to manufacturer instructions. RNA-seq libraries were subjected to an additional purification to remove remaining high molecular weight products as follows: sample volume was increased to 100  $\mu$ l by addition of 1X TE buffer or Illumina Resuspension Buffer and then incubated with 60  $\mu$ l Agencourt AMPure XP beads for 4 min. The beads were pelleted by incubation on a magnet, and the entire supernatant was transferred to a tube containing 50  $\mu$ l of fresh AMPure XP beads and incubated for 4 min. After pelleting the new beads with a magnet, the supernatant was discarded, the beads washed twice with 80% ethanol and the DNA was eluted in 30  $\mu$ l Illumina Resuspension buffer. The resulting RNA-seq libraries were diluted 10  $\times$  and their quality and concentration were assessed using a 2100 Bioanalyzer with the High Sensitivity DNA Kit (Agilent) and a Qubit Fluorometer with the Qubit dsDNA HS Assay Kit (Life Technologies). RNA-seq libraries were pooled four libraries per lane and sequenced via single end 50 bp reads on a HiSeq 2000 (Illumina).

RNA-seq data was analysed as follows: CASAVA v1.8.0 (Illumina) was used to demultiplex data, and reads with CASAVA ‘Y’ flag (purity filtering) were discarded. After quality filtering and adaptor trimming using cutadapt.v1.1 (ref. 64) with parameter ‘-m 25 -q 25’, between 37 and 54 million reads were obtained for each sample. Mouse genome sequence (mm9) and gene annotation were retrieved from the iGenomes repository ([https://support.illumina.com/sequencing/sequencing\\_software/igenome.html](https://support.illumina.com/sequencing/sequencing_software/igenome.html)). Tophat v2.0.6 (ref. 65) was used to align the reads to the mouse reference genome and transcriptome, then reads mapping to UCSC known genes were counted by HTSeq (ref. 66). Differential gene expression analysis between wild type and knockout for mm77 or mm771 enhancers was performed using edgeR (ref. 67), and genes whose expression was extremely low in all samples (fragments per kilobase of transcript per million mapped reads (FPKM)  $< 1$ , calculated by Cufflinks v2.2.1 (ref. 68)) were discarded for further analysis.  $P$  values for all RNA-seq experiments are those reported by edgeR after false discovery rate (FDR) correction (FDR  $< 5\%$ ).

For quantitative RT-PCR (qPCR) measuring *Myl2*, *Myh7* and *Ubc*, RNA was reverse transcribed using SuperScript III (Life Technologies) with random hexamer or poly-dT priming according to manufacturer instructions. For qPCR measuring *Nppa* and *Nppb*, RNA was first treated with RNase-free DNase (Promega) and then reverse transcribed using SuperScript III with poly-dT priming. qPCR was performed on a LightCycler 480 (Roche) using TaqMan-style reactions containing Master Mix (LightCycler 480 Probes Master Mix (Roche) or Probe Fast Universal qPCR Master Mix (Kapa Biosystems)), PrimeTime qPCR Assay primer/probe mix (Integrated DNA Technologies) for the test gene, primer/probe mix to the actin control gene, 0.5–1  $\mu$ l of the reverse transcriptase reaction and RNAase-free water.

qPCR assays were performed in triplicate for each sample, and genomic DNA amplification was excluded for all samples by a lack of substantial amplification in reverse transcriptase-negative qPCR reactions or by the absence of a genomic DNA amplification band when reverse transcriptase-positive reactions were run on an agarose gel. PrimeTime qPCR Assay primer/probe mixes used for qPCR assays are provided in Supplementary Table 10. Reactions were analysed as previously described<sup>69</sup>, using actin as the reference gene.

**Protein expression analysis for enhancer knockouts.** Myosin was extracted from individual embryonic or adult mouse whole hearts as previously described<sup>70</sup> with minor modifications. Protein was quantified by Bradford Assay. Proteins were separated on 8–16% gradient Tris-glycine gels by electrophoresis and transferred to polyvinylidene difluoride membranes using standard western blotting procedures. Membranes were blocked with 5% bovine serum albumin, incubated with primary antibodies overnight, washed, incubated with secondary antibodies for 1 h, washed and imaged using a VersaDoc Molecular Imager (Bio-Rad). Antibody details are provided in Supplementary Table 11. Quantification of band intensities was carried out using the gel analyzer feature of ImageJ (ref. 71), with GAPDH serving as a loading control. Uncropped western blot images are provided in Supplementary Fig. 11.

**Echocardiography, necropsy, pathology and histology.** Echocardiography was performed by the University of California Davis Mouse Metabolic Phenotyping Center's Cardiovascular Biology and Pathology Core. Echocardiograms were performed on conscious animals to assess the systolic function using M-mode and two-dimensional measurements as previously described<sup>72</sup>.

Gross necropsy, histology and pathology were performed by the University of California Davis Comparative Pathology Lab using standard techniques. Briefly, animals were euthanized by carbon dioxide asphyxiation. Organs were collected, weighed and fixed with paraformaldehyde. Tissues were paraffin embedded, and 5 µm sections were stained with haematoxylin and eosin for pathological evaluation.

Cardiac histology sections were assessed by a trained veterinary pathologist, who was blinded to genotype status. The severity of left ventricular hypertrophy, myocardiocyte disarray, myocardiocyte karyomegaly and interstitial fibrosis present in each heart sample were scored as follows:

- 0 = not observed
- 1 = minimal (1–2 foci or <10% of heart involved)
- 2 = mild (3–6 foci or 10–40% of heart involved)
- 3 = moderate (6–10 foci or 40–60% of heart involved)
- 4 = severe (10+ foci or >60% of heart involved)

**Animal approval and experimental design.** All animal work was reviewed and approved by the Lawrence Berkeley National Laboratory Animal Welfare and Research Committee or the University of California, Davis Institutional Animal Care and Use Committee.

R was used to compute the statistics and to generate plots for this work.

**Transgenic mouse assays.** Transgenic assays were performed in *Mus musculus* FVB strain mice. Sample sizes were selected empirically based on our previous experience of performing transgenic mouse assays for >2,000 total putative enhancers (for example, refs 12,14,15,33,62). Mouse embryos were only excluded from further analysis if they did not carry the reporter transgene or if they were not at the correct developmental stage. As all transgenic mice were treated with identical experimental conditions, and as there were no groups of animals directly compared in this section of the study, randomization and experimenter blinding were unnecessary and not performed.

**Enhancer knockouts.** Enhancers were deleted in *Mus musculus* W4 (129S6 strain) mouse embryonic stem cells (Taconic). Resulting mice were crossed into the C57BL/6J strain. All mice and mouse embryos described in the enhancer knockout section of this paper resulted from heterozygous x heterozygous crosses to allow for the comparison of matched littermates of different genotypes. With the exception of qPCR experiments (where entire litters were analysed) and the RNA-seq performed on adult mm77 mice (where sufficient matched littermates were not available), all experiments employed a matched littermate selection strategy. For every homozygous null animal selected, a homozygous wild-type animal from the same litter was selected for comparison. For all postnatal mice, littermate pairs were selected to have matching genders. Embryonic samples used for qPCR, RNA-seq and western blotting were dissected blind to genotype.

The underlying hypothesis of the enhancer knockout section was that loss of either cardiac enhancer would decrease neighbouring myosin gene and protein levels. Based on the *Myl2* gene-null phenotype previously observed in humans and mice (see Supplementary Note 4), we expected this decrease in gene expression to result in reduced cardiac function and possibly mild cardiac hypertrophy in homozygous null mice relative to matched wild-type littermates. Therefore, statistical significance of results was assessed by one-tailed *t*-test or Wilcoxon rank-sum test for metrics such as gene/protein expression, heart weight/body weight ratios, left ventricular mass and cardiac function (ejection fraction and fractional shortening) comparisons. Remaining metrics were assessed using

two-tailed tests. Because of the matched littermate selection scheme used, paired tests were used to assess significance of echocardiography and pathology results.

**qPCR.** Embryonic samples were collected from at least two independent litters for each line. Embryos were excluded from any further analysis if they were in the process of being resorbed, not at the correct developmental stage, or insufficient quantities of RNA were isolated from them. Otherwise, all embryonic samples collected for each qPCR experiment were analysed. Adult samples for qPCR validation of *Nppa*, *Nppb* and *Ubc* were technical replicates of those used for RNA-seq.

**RNA-seq.** Samples were chosen for RNA-seq based on RNA sample quantity and quality, and the availability of matched littermate pairs. To avoid batch effects, RNA-seq libraries from all samples within the same experiment were made in the same batch. Libraries were pooled together such that matched littermate pairs were run on the same flow cell lane.

**Western blots.** One potential homozygous null mm77 animal was excluded from analysis due to inconclusive genotyping results, resulting in one unmatched wild-type sample in this analysis. To avoid batch effects, homozygous null and homozygous wild-type samples were alternated for protein extraction, and all samples within an experiment were run on the same gel and blot.

**Echocardiography, Histology and Pathology:** to control for any physiological effects due to strain background, age or gender, littermate pairs were selected for further phenotyping such that one mouse in each pair was homozygous wild-type and one mouse was homozygous null and both mice in each pair were the same gender. To minimize physiological effects due to body weight, littermate pairs were selected to minimize differences in body weights between paired samples. Detailed information (age, sex, body mass and so on) about mice selected for this phenotyping is provided in Supplementary Data 6. All echocardiography, necropsy, pathology and histology assays were performed blinded to genotype, and mice were randomized for echocardiography. Sample sizes were selected for echocardiography based on similar previously reported studies<sup>73</sup>. Left ventricular mass could not be ascertained by echocardiography for one mouse homozygous null for the mm771 enhancer.

**Data availability.** RNA-seq files are available in the NCBI GEO database with the accession code GSE75907. All other data that support the findings of this study are available from the corresponding author upon request.

## References

- Mozaffarian, D. *et al.* Heart disease and stroke statistics—2015 update: a report from the American Heart Association. *Circulation* **131**, e29–322 (2015).
- Yang, Q. *et al.* Racial differences in infant mortality attributable to birth defects in the United States, 1989–2002. *Birth Defects Res. A Clin. Mol. Teratol.* **76**, 706–713 (2006).
- Ganesh, S. K. *et al.* Genetics and genomics for the prevention and treatment of cardiovascular disease: update: a scientific statement from the American Heart Association. in *Circulation* **128**, 2813–2851 (2013).
- Hershberger, R. E. *et al.* Genetic evaluation of cardiomyopathy—a Heart Failure Society of America practice guideline. *J. Card. Fail.* **15**, 83–97 (2009).
- Vassy, J. L. *et al.* The MedSeq Project: a randomized trial of integrating whole genome sequencing into clinical medicine. *Trials* **15**, 85 (2014).
- UK10K Consortium, *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
- Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
- Nord, A. S. *et al.* Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155**, 1521–1531 (2013).
- Lettec, L. A. *et al.* A long-range *Shh* enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735 (2003).
- Bauer, D. E. *et al.* An erythroid enhancer of *BCL11A* subject to genetic variation determines fetal hemoglobin level. *Science* **342**, 253–257 (2013).
- Oldridge, D. A. *et al.* Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism. *Nature* **528**, 418–421 (2015).
- Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
- Cotney, J. *et al.* Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. *Genome Res.* **22**, 1069–1080 (2012).
- May, D. *et al.* Large-scale discovery of enhancers from human heart tissue. *Nat. Genet.* **44**, 89–93 (2012).
- Blow, M. J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* **42**, 806–810 (2010).
- Roadmap Epigenomics Consortium, *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
- Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
- Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116–120 (2012).



19. van den Boogaard, M. *et al.* Genetic variation in T-box binding element functionally affects SCN5A/SCN10A enhancer. *J. Clin. Invest.* **122**, 2519–2530 (2012).
20. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.* **40**, 897–903 (2008).
21. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
22. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
23. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
24. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–D92 (2007).
25. Dogan, N. *et al.* Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics Chromatin* **8**, 16 (2015).
26. van Duijvenboden, K., de Boer, B. A., Capon, N., Ruijter, J. M. & Christoffels, V. M. EMERGE: a flexible modelling framework to predict genomic regulatory elements from genomic signatures. *Nucleic Acids Res.* **44**, e42 (2016).
27. Erwin, G. D. *et al.* Integrating diverse datasets improves developmental enhancer prediction. *PLoS Comput. Biol.* **10**, e1003677 (2014).
28. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
29. Eijgelsheim, M. *et al.* Genome-wide association analysis identifies multiple loci related to resting heart rate. *Hum. Mol. Genet.* **19**, 3885–3894 (2010).
30. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43**, 333–338 (2011).
31. Ellinor, P. T. *et al.* Meta-analysis identifies six new susceptibility loci for atrial fibrillation. *Nat. Genet.* **44**, 670–675 (2012).
32. Kothary, R. *et al.* Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice. *Development* **105**, 707–714 (1989).
33. Pennacchio, L. A. *et al.* *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
34. Xie, X. *et al.* CCSi: a database providing chromatin-chromatin spatial interaction information. *Database (Oxford)* **2016**, bav124 (2016).
35. Gollob, M. H. *et al.* Identification of a gene responsible for familial Wolff-Parkinson-White syndrome. *N. Engl. J. Med.* **344**, 1823–1831 (2001).
36. Bertoli-Avella, A. M. *et al.* Mutations in a TGF- $\beta$  ligand, TGFB3, cause syndromic aortic aneurysms and dissections. *J. Am. Coll. Cardiol.* **65**, 1324–1336 (2015).
37. Schachterle, W., Rojas, A., Xu, S.-M. & Black, B. L. ETS-dependent regulation of a distal Gata4 cardiac enhancer. *Dev. Biol.* **361**, 439–449 (2012).
38. Lyons, G. E., Schiaffino, S., Sassoon, D., Barton, P. & Buckingham, M. Developmental regulation of myosin gene expression in mouse cardiac muscle. *J. Cell Biol.* **111**, 2427–2436 (1990).
39. Yasue, H. *et al.* Localization and mechanism of secretion of B-type natriuretic peptide in comparison with those of A-type natriuretic peptide in normal subjects and patients with heart failure. *Circulation* **90**, 195–203 (1994).
40. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
41. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
42. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
43. Hu, J. *et al.* Endothelin signaling activates Mef2c expression in the neural crest through a MEF2C-dependent positive-feedback transcriptional pathway. *Development* **142**, 2775–2780 (2015).
44. Han, Y., Slivano, O. J., Christie, C. K., Cheng, A. W. & Miano, J. M. CRISPR-Cas9 genome editing of a single regulatory element nearly abolishes target gene expression in mice. *Arterioscler. Thromb. Vasc. Biol.* **35**, 312–315 (2014).
45. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
46. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2013).
47. Leinonen, R., Sugawara, H. & Shumway, M. International nucleotide sequence database collaboration. The sequence read archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
48. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
49. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
50. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).
51. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
52. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
53. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
54. Rosenbloom, K. R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43**, D670–D681 (2015).
55. Jalili, V., Matteucci, M., Maseroli, M. & Morelli, M. J. Using combined evidence from replicates to evaluate ChIP-seq peaks. *Bioinformatics* **31**, 2761–2769 (2015).
56. Cotney, J. *et al.* The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell* **154**, 185–196 (2013).
57. ENCODE Project Consortium, *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
58. Eppig, J. T. *et al.* The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res.* **43**, D726–D736 (2015).
59. Köhler, S. *et al.* The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42**, D966–D974 (2014).
60. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).
61. Johnson, A. D. *et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938–2939 (2008).
62. Attanasio, C. *et al.* Fine tuning of craniofacial morphology by distant-acting enhancers. *Science* **342**, 1241006 (2013).
63. Dickel, D. E. *et al.* Function-based identification of mammalian enhancers using site-specific integration. *Nat. Methods* **11**, 566–571 (2014).
64. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal* **17**, 10–12 (2011).
65. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
66. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
67. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
68. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
69. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2<sup>-</sup>(Delta Delta C(T)) Method. *Methods* **25**, 402–408 (2001).
70. Bär, A. & Pette, D. Three fast myosin heavy chains in adult rat skeletal muscle. *FEBS Lett.* **235**, 153–155 (1988).
71. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
72. Li, N. *et al.* Beneficial effects of soluble epoxide hydrolase inhibitors in myocardial infarction model: insight gained using metabolomic approaches. *J. Mol. Cell Cardiol.* **47**, 835–845 (2009).
73. Chin, B. B. *et al.* Left ventricular functional assessment in mice: feasibility of high spatial and temporal resolution ECG-gated blood pool SPECT. *Radiology* **245**, 440–448 (2007).

## Acknowledgements

This work was supported by National Institutes of Health grants R24HL123879, U01DE024427, R01HG003988, U54HG006997 and UM1HL098166. M.O. was supported by a Swiss National Science Foundation (SNSF) fellowship. Research was conducted at the E.O. Lawrence Berkeley National Laboratory and performed under Department of Energy Contract DE-AC02-05CH11231, University of California. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, which is supported by National Institute of Health Grants S10RR029668 and S10RR027303. Histology, pathology and echocardiography were performed at the University of California Davis Mouse Metabolic Phenotyping Center, which is supported by National Institute of Diabetes and Digestive and Kidney Diseases grant U24DK092993. We would like to thank N. Chiamvimonvat (University of California Davis) for help in performing and interpreting the echocardiography testing.

## Author contributions

D.E.D. and I.B. developed the cardiac compendium, and I.B. performed the integrative and retrospective analyses. D.E.D., M.O., B.J.M., D.M., C.H.S., I.P.-F., C.S.P., T.H.G., M.K., J.A.A. and V.A. carried out transgenic validation. D.E.D., Y.Z. and Y.F.-Y. performed the enhancer knockout studies. A.Y.L., D.U.G. and B.R. generated the

unpublished ENCODE ChIP-seq data using tissue provided by A.V. and L.A.P; D.E.D., E.M.R., A.V. and L.A.P. planned the study. D.E.D., A.V. and L.A.P. wrote the manuscript with input from the remaining authors.

### Additional information

**Supplementary Information** accompanies this paper at <http://www.nature.com/naturecommunications>

**Competing financial interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**How to cite this article:** Dickel, D. E. *et al.* Genome-wide compendium and functional assessment of *in vivo* heart enhancers. *Nat. Commun.* **7**, 12923 doi: 10.1038/ncomms12923 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2016