Biobank-driven genomic discovery yields new insight into atrial fibrillation biology

Jonas B. Nielsen [1,2,23], Rosa B. Thorolfsdottir [3,4,23], Lars G. Fritsche [1,5,6,7,23], Wei Zhou [1,7,23], Morten W. Skov [8,23], Sarah E. Graham [1,2,23], Todd J. Herron [9,23], Shane McCarthy [10,23], Ellen M. Schmidt [11,23], Gardar Sveinbjornsson [3,23], Ida Surakka [1,2], Michael R. Mathis [12], Masatoshi Yamazaki [13], Ryan D. Crawford [7], Maiken E. Gabrielsen [5,6], Anne Heidi Skogholt [5,6], Oddgeir L. Holmen [5,6,14], Maoxuan Lin [1,2], Brooke N. Wolford [1,7], Rounak Dey [11], Håvard Dalen [15,16,17], Patrick Sulem [3], Jonathan H. Chung [10], Joshua D. Backman [10], David O. Arnar [3,4,18], Unnur Thorsteinsdottir [3,4], Aris Baras [10], Colm O'Dushlaine [10], Anders G. Holst [8], Xiaoquan Wen [11], Whitney Hornsby [1], Frederick E. Dewey [10], Michael Boehnke [11], Sachin Kheterpal [12], Bhramar Mukherjee [11], Seunggeun Lee [11], Hyun M. Kang [11], Hilma Holm [3], Jacob Kitzman [2], Jordan A. Shavit [19], José Jalife [1,9,20], Chad M. Brummett [12], Tanya M. Teslovich [10], David J. Carey [21], Daniel F. Gudbjartsson [3,22], Kari Stefansson [3,4], Gonçalo R. Abecasis [6,11]*, Kristian Hveem [5,6,15]* and Cristen J. Willer [1,2,7]*

To identify genetic variation underlying atrial fibrillation, the most common cardiac arrhythmia, we performed a genome-wide association study of >1,000,000 people, including 60,620 atrial fibrillation cases and 970,216 controls. We identified 142 independent risk variants at 111 loci and prioritized 151 functional candidate genes likely to be involved in atrial fibrillation. Many of the identified risk variants fall near genes where more deleterious mutations have been reported to cause serious heart defects in humans (*GATA4, MYH6, NKX2-5, PITX2, TBX5*)[1], or near genes important for striated muscle function and integrity (for example, *CFL2, MYH7, PKP2, RBM20, SGCG, SSPN*). Pathway and functional enrichment analyses also suggested that many of the putative atrial fibrillation genes act via cardiac structural remodeling, potentially in the form of an 'atrial cardiomyopathy'[2], either during fetal heart development or as a response to stress in the adult heart.

We tested association between 34,740,186 genetic variants (minor allele frequency (MAF) >2.5×10^{-5}) and atrial fibrillation, comparing a total of 60,620 cases and 970,216 controls of European ancestry from six contributing studies (The Nord-Trøndelag Health Study (HUNT), deCODE, the Michigan Genomics Initiative (MGI), DiscovEHR, UK Biobank, and the AFGen Consortium) (Supplementary Table 1). We identified 111 genomic regions with at least 1 genetic variant associated with atrial fibrillation ($P < 5 \times 10^{-8}$). Of these, 80 loci have not previously been reported (Fig. 1, Supplementary Fig. 1, and Supplementary Tables 2 and 3). Based on approximate stepwise conditional analyses[3], we identified 31 additional genetic risk variants that demonstrated genome-wide statistically significant association with atrial fibrillation (Supplementary Table 4) that were nearby but independent of the 111 index variants (linkage disequilibrium (LD) $r^2 < 0.10$). We applied the widely used genome-wide association study (GWAS) P value significance threshold of $P < 5 \times 10^{-8}$. Had we applied a more stringent threshold of $P < 5 \times 10^{-9}$ (ref. [4]), we would identify 94 loci, 63 of which have not been previously reported (Supplementary Table 2). We found that the total genome-wide genetic variation captured in this study explained 11.2% (s.e.m. 1.4%) of the variation in atrial fibrillation

[1]Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor, MI, USA. [2]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA. [3]deCODE genetics/Amgen, Inc., Reykjavik, Iceland. [4]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [5]HUNT Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, Levanger, Norway. [6]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health, Norwegian University of Science and Technology, Trondheim, Norway. [7]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [8]Laboratory of Molecular Cardiology, Department of Cardiology, The Heart Centre, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark. [9]Department of Internal Medicine, Center for Arrhythmia Research, University of Michigan, Ann Arbor, MI, USA. [10]Regeneron Genetics Center, Tarrytown, NY, USA. [11]Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [12]Department of Anesthesiology, University of Michigan, Ann Arbor, MI, USA. [13]Medical Device Development and Regulation Research Center, The University of Tokyo, Tokyo, Japan. [14]Department of Cardiology, St. Olav's University Hospital, Trondheim, Norway. [15]Department of Medicine, Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway. [16]Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway. [17]Department of Cardiology, St. Olav's University Hospital, Trondheim University Hospital, Trondheim, Norway. [18]Department of Medicine, Landspitali – National University Hospital, Reykjavik, Iceland. [19]Department of Pediatrics and Communicable Diseases, University of Michigan, Ann Arbor, MI, USA. [20]Fundación Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid, Spain. [21]Geisinger Health System, Danville, PA, USA. [22]School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland. [23]These authors contributed equally: Jonas B. Nielsen, Rosa B. Thorolfsdottir, Lars G. Fritsche, Wei Zhou, Morten W. Skov, Sarah E. Graham, Todd J. Herron, Shane McCarthy, Ellen M. Schmidt, Gardar Sveinbjornsson. *e-mail: goncalo@umich.edu; kristian.hveem@ntnu.no; cristen@umich.edu
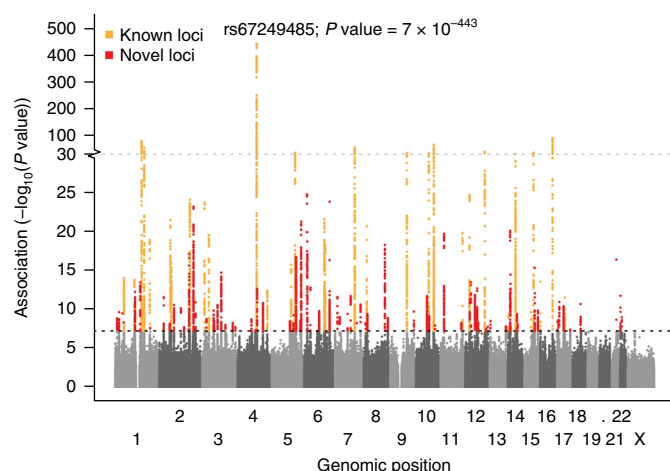
Fig. 1 | Manhattan plot showing known (orange) and novel (red) loci associated with atrial fibrillation. A total of 34,740,186 genetic variants (each represented by a dot) were tested, comparing 60,620 atrial fibrillation cases and 970,216 controls free of atrial fibrillation. The *x* axis represents the genome in physical order, and the *y* axis represents *P* values ($-\log_{10}(P \text{ value})$) of association. The black horizontal dotted line represents a Bonferroni-corrected threshold of statistical significance corresponding to 1,000,000 independent tests ($P < 5 \times 10^{-8}$).

(h^2_{SNP} heritability). This is consistent with a recent report of 11.4%[5]. When combining the 111 locus index variants and the 31 additional genetic risk variants, we found that they explained 4.6% of the variation in atrial fibrillation.

Of the 35 loci previously reported for atrial fibrillation (Supplementary Table 3), we identified genome-wide significant association ($P < 5 \times 10^{-8}$) at 31 (89%) after excluding results from the previously published AFGen Consortium (Supplementary Table 5)[6]. The four loci not captured comprised three loci discovered in East Asian populations (*KCNIP1*, *NEBL*, *CUX2*) and one missense variant (*PLEC*) for which we did not have data[7]. To further test the validity of our findings, we performed a heterogeneity test for the 111 index variants across the 6 contributing studies. Of the 111 index variants, only 2 index variants demonstrated evidence for heterogeneity of effect size across the 6 contributing studies when correcting for multiple testing ($P < 0.05/111 = 4.5 \times 10^{-4}$) (Supplementary Table 2). Both of these index variants represent loci that have previously been established as associated with atrial fibrillation across multiple studies (near *PRRX1*, *PITX2*) (Supplementary Table 3). These findings demonstrate a high external validity of our results.

To understand the biology underlying the 111 atrial fibrillation-associated loci, we employed a number of approaches, including 'Data-driven Expression Prioritized Integration for Complex Traits' (DEPICT)[8] to identify cell types and tissues in which atrial fibrillation genes are likely to be expressed. Based on 37,427 human microarray expression samples from 209 different tissues and cell types, we observed a statistically significant enrichment for atrial ($P = 2.4 \times 10^{-5}$), atrial appendage ($P = 2.8 \times 10^{-5}$), heart ($P = 5.2 \times 10^{-5}$), and ventricular tissues ($P = 1.1 \times 10^{-4}$) (Fig. 2a and Supplementary Table 6). We further applied DEPICT to detect gene sets that were enriched for genes at atrial fibrillation-associated loci. Of the 14,461 gene sets we tested, 889 were enriched (false discovery rate (FDR) < 0.05; Fig. 2b and Supplementary Table 7). The highlighted gene sets point to biological processes related to cardiac development and morphology along with structural remodeling of the myocardium. These findings are consistent with recent reports that have linked atrial fibrillation with rare coding variants

in the sarcomere genes *MYH6* and *MYL4* and in the multidomain cytoskeletal linking protein *PLEC*, along with more common coding variants in *TTN*, essential for the passive elasticity of heart and skeletal muscle[7,9–11].

Although we could identify protein-altering variants at 21 loci, comprising either the index variant ($n = 2$ loci) or a variant in high LD (r^2) with the index variant ($n = 19$ loci; Supplementary Table 8), we noted that most associated risk variants are in the non-coding genome. To assess the potential function of associated non-coding variants, we tested for enrichment of atrial fibrillation-associated variants with a variety of regulatory features, including DNase I hypersensitive sites, histone methylation marks, transcription factor binding sites, and chromatin states in a variety of cell and tissue types available from Roadmap Epigenomics[12] using 'Genomic Regulatory Elements and Gwas Overlap algoRithm' (GREGOR)[13]. This method tests whether the number of atrial fibrillation-associated index variants, or their LD proxies, overlap with the corresponding regulatory feature more often than expected when compared to control sets. Of 785 combinations of regulatory features and tissues examined (Supplementary Table 9), we found that atrial fibrillation-associated variants were most strongly associated with features in adult and fetal heart: active enhancers as indicated by H3K27ac in right atrium ($P = 2 \times 10^{-33}$; 2.9 × enrichment); H3K27ac in left ventricle ($P = 3 \times 10^{-33}$; 2.6 × enrichment); and in fetal heart tissue we found strong enrichment with H3K4me1 ($P = 9 \times 10^{-27}$; 2.0 × enrichment) and open chromatin as measured by DNase hypersensitivity ($P = 2 \times 10^{-26}$; 2.1 × enrichment) (Fig. 2c, Supplementary Fig. 2 and Supplementary Table 9). This suggests that some atrial fibrillation loci are important in transcriptional regulation in the adult heart, in development of the fetal heart, or both.

To further enhance the biological understanding of the atrial fibrillation-associated loci, we identified candidate functional genes. There were 3,048 genes or transcripts for which the transcribed region overlapped (see Online Methods) at least 1 variant in the 111 loci. We prioritized biological candidate genes that: (1) harbored a protein-altering index variant itself or in high LD ($r^2 > 0.80$; Supplementary Table 8); (2) had expression levels that were associated and colocalized with atrial fibrillation-associated variants ($P < 1.14 \times 10^{-9}$ in Genotype-Tissue Expression (GTEx) consortium data)[14]; (3) were highlighted by DEPICT (FDR <0.05; Supplementary Table 10); or (4) were nearest to the index variant in a locus. Using these criteria, we prioritized 151 candidate genes (Supplementary Tables 2 and 11).

To identify tissues in which the 151 prioritized candidate genes showed enhanced expression, we used 'Tissue Specific Expression Analysis' (TSEA)[15] and found enrichment in heart ($P = 1 \times 10^{-16}$), blood vessel (9×10^{-13}) muscle tissues ($P = 7 \times 10^{-11}$). To assess the empirical significance of these results, we performed 1,000 permutations of the same number of genes selected: (1) randomly from the genome, and (2) subsets of the 3,048 genes within the 111 atrial fibrillation loci. We determined that the observed TSEA *P* values were substantially more significant than expected by chance (Fig. 3). The finding of increased expression of these genes in heart support that the genes we prioritized are strong candidates for being involved in atrial fibrillation.

Interestingly, we identified as functional candidates at least 18 genes likely to be involved in cardiac and skeletal muscle function and integrity (*AKAP6*, *CFL2*, *MYH6*, *MYH7*, *MYO18B*, *MYO1C*, *MYOCD*, *MYOT*, *MYOZ1*, *MYPN*, *PKP2*, *RBM20*, *SGCA*, *SSPN*, *SYNPO2L*, *TTN*, *TTN-AS*, *WIPF1*); these included *SGCG*, which has been associated with muscular dystrophy[16], *RBM20*, which has been associated with dilated cardiomyopathy[17], and *PKP2*, which has been associated with arrhythmogenic right ventricular cardiomyopathy[18]. We identified at least 13 genes likely to be involved in mediation of developmental events (*ARNT2*, *EPHA3*, *FGF5*, *GATA4*, *GTF2I*, *HAND2*, *LRRC10*, *NAV2*, *NKX2-5*, *PITX2*, *SLIT3*,
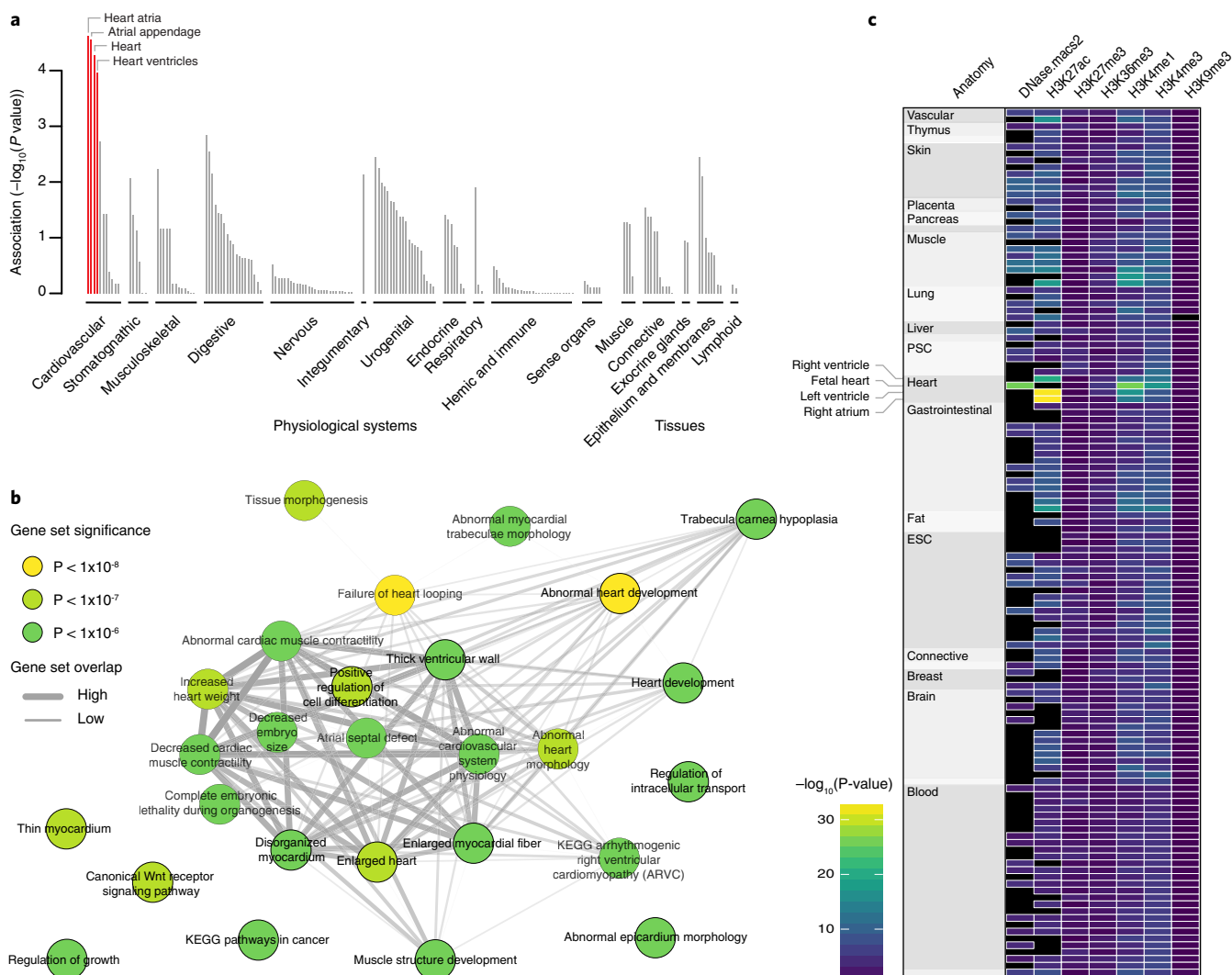
Fig. 2 | Tissues, reconstituted gene sets, and regulatory elements implicated in atrial fibrillation. a, Based on expression patterns across 37,427 human mRNA microarrays, DEPICT predicted genes within atrial fibrillation-associated loci to be highly expressed across various cardiac tissues. Tissues are grouped by type and significance. Red columns represent statistically significant tissues following Bonferroni correction ($P < 0.05/209 = 0.0002$). **b**, Top ($P < 1 \times 10^{-6}$) reconstituted gene sets (out of 826 with FDR < 0.05 and after exclusion of 'gene subnetworks') found by DEPICT to be significantly enriched for genes in atrial fibrillation-associated loci. Each node, colored according to the permutation *P* value, represents a gene set and the gray connecting lines represent pairwise overlap of genes within the gene sets. **c**, Heatmap indicating the overlap between atrial fibrillation-associated risk variants and regulatory elements across 127 Roadmap Epigenomics tissues (each represented by a row) using GREGOR. Black indicates no data. PSC, pluripotent stem cell; ESC, embryonic stem cell.

SOX15, *TBX5*) along with genes likely to be involved in intracellular calcium handling in the heart (*CALU*, *CAMK2D*, *CASQ2*, *PLN*), angiogenesis (*TNFSF12*, *TNFSF12-TNFSF13*), hormone signaling (*CGA*, *ESR2*, *IGF1R*, *NR3C1*, *THRB*), and function of cardiac ion channels (*HCN4*, *KCND3*, *KCNH2*, *KCNJ5*, *KCNN2*, *KCNN3*, *SCN10A*, *SCN5A*, *SLC9B1*).

We tested the 111 atrial fibrillation index variants for association with 123 electrocardiogram (ECG) parameters in 62,974 Icelanders in sinus rhythm, after exclusion of atrial fibrillation cases (Supplementary Fig. 3 and Supplementary Table 12). Sixty variants were associated with at least 1 ECG parameter when we controlled for an FDR of 0.05 at the variant level, 39 of which were novel atrial fibrillation variants, including many with substantial ECG effects, such as the variants near *NACA*, *THRB*, *CAMK2D*, *NKX2-5*, and *CDKN1A*. Many of the associated ECG parameters, including heart rate, P-wave duration, PR interval, and heart rate-corrected QT interval, are well-established intermediate phenotypes for atrial

fibrillation[19–22]. Accordingly, several of the highlighted associations can be seen as indirect replications of the atrial fibrillation risk variants identified through GWASs. The results also indicate that many of the 111 atrial fibrillation risk variants act in the heart before atrial fibrillation. The type and direction of the ECG parameter associations might help inform the biology underlying the specific loci.

For the locus around index variant rs422068 on chromosome 14, our approach prioritized *MYH6* and *MYH7* as the most likely functional genes (Supplementary Table 2). A rare missense mutation in *MYH6* has in prior GWASs been associated with sick sinus syndrome[23], atrial fibrillation[7], and coarctation of the aorta[24], and several protein-altering variants in *MYH7* have been linked to hypertrophic cardiomyopathy[25]. *MYH6* and *MYH7* encode the molecular motors of cardiac muscle that transduce chemical energy from ATP hydrolysis into mechanical energy of each heartbeat. *MYH6* encodes α-myosin heavy chain (α-MyHC), which is the faster molecular motor of the thick filaments of the contractile
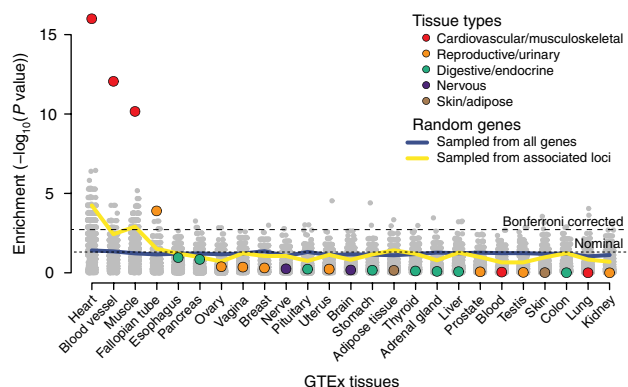
Fig. 3 | Significance of the expression enrichment for the atrial fibrillation candidate genes. This figure compares the tissue-specific gene expression enrichment for the 151 biological candidate genes (colored dots) to a null distribution derived by randomly selecting the same number of genes from the whole genome or from the associated loci. Tissue-specific gene expression data ($n = 25$ tissues) were obtained from Genotype-Tissue Expression (GTEx) consortium data. The gray dots represent the P values for each of the permutations for the randomized tests (1,000 for both sampling scenarios for each tissue), and the blue and yellow lines represent the per-tissue P value thresholds comparable to a false positive rate of 0.05.

apparatus in healthy adult atrial muscle[26]. On the other hand, *MYH7* encodes β-MyHC, a slower molecular motor[27], which is expressed only in the atria during cardiac development and not in the normal adult atria. It has been established that *MYH6* and *MYH7* are regulated in an inverse manner in the ventricles of the heart, and in heart failure and other cardiac disorders in humans, β-MyHC is upregulated, whereas α-MyHC is downregulated, resulting in diminution of cardiac performance[28]. Importantly, recent experiments have demonstrated that *MYH7* expression is elevated in atrial myocytes of patients with chronic atrial fibrillation as well as in an ovine model of chronic atrial fibrillation[29].

To explore potential mechanisms of *MYH6* and *MYH7* in atrial fibrillation, we developed an ischemic heart failure model for atrial fibrillation in rabbits. Ischemia was produced by chronic ligation of the left circumflex artery during thoracotomy with subsequent development of ischemic heart failure (>4 weeks postoperatively), profound left atrial dilation, and development of long-lasting atrial fibrillation following burst pacing (Fig. 4 and Supplementary Fig. 4). We found that MYH7 expression was only detectable in the heart failure remodeled left atrium (Fig. 4 and Supplementary Fig. 5). The control left atrium did not express detectable levels of MYH7 and exclusively expressed MYH6. More importantly, in the dilated left atrium, MYH7 expression was heterogeneously distributed (Fig. 4 and Supplementary Fig. 6), which thus resulted in contractile and metabolic heterogeneity, both of which are probably arrhythmogenic. Although the association between the rs422068 locus and atrial fibrillation could potentially be mediated through structural heart defects such as coarctation of the aorta or hypertrophic cardiomyopathy via genetic variation not captured in this study, it is likely that the heterogeneously distributed switch from the adult to the fetal isoform of myosin heavy chain that we observed in the dilated left atrium may predispose rabbit (and possibly human) hearts to developing long-lasting atrial fibrillation.

Next, we investigated whether any of the 151 biological candidate genes that we identified could potentially represent a novel drug target for already developed drugs or drugs undergoing development by querying the Drug Gene Interaction Database[30]. We found 1 or more potential drug or substance interactions for 32 of the 151 prioritized genes, totaling 475 drugs. Of these, 78 drugs targeting

14 genes are already known to be able to control or trigger atrial fibrillation or other cardiac arrhythmias (Supplementary Table 13). In addition to a number of drugs that could potentially impact atrial fibrillation via an effect on cardiac ion channels (for example, rottlerin, bepridil), including drugs already used for treating various neuropsychiatric disorders (for example, fosphenytoin, flunarizine), we also identified a number of anti-inflammatory drugs, including several glucocorticoids, and a cardiac-specific myosin activator (omecamtiv mecarbil), which is currently being tested for treatment of heart failure[31]. Whether the highlighted drugs can be used to treat or prevent atrial fibrillation requires further evaluation, but the findings can be used as a foundation for directing future functional experiments and clinical trials.

We constructed a polygenic risk score based on the 111 locus index variants and the 31 additional risk variants, identified through stepwise conditional analyses, weighted by effect estimates obtained from meta-analyses excluding the UK Biobank (Supplementary Table 14). We found that the risk score predicted prevalent atrial fibrillation in the UK Biobank with an unadjusted area under the receiver operator curve of 65%. We then used the polygenic risk score to test for association with 1,494 International Classification of Diseases (ICD) code-defined disease groups in UK Biobank participants of white British ancestry[32]. In addition to a strong association with atrial fibrillation ($P = 2 \times 10^{-920}$), we found association to additional mainly cardiovascular conditions ($P < 0.05/1,494 = 3.3 \times 10^{-5}$), including palpitations, heart valve disorders, heart failure, ischemic heart disease, and stroke (Supplementary Table 15 and Supplementary Fig. 7). However, when participants diagnosed with any type of cardiac arrhythmia ($n = 24,681$) were excluded from the analyses to avoid assessment bias (termed an exclusion phenome-wide association study)[33], the atrial fibrillation risk score was no longer associated with any ICD disease group (all $P > 3.3 \times 10^{-5}$). This suggests that the atrial fibrillation polygenic risk score is specific for atrial fibrillation and that the additional associations identified were mediated through atrial fibrillation, either as a result of a more thorough clinical examination (for example, heart valve disorders) or because atrial fibrillation is an intermediate step towards the disease (for example, stroke).

To examine the genetic impact on age of onset of atrial fibrillation, we generated a polygenic risk score ($n = 142$ markers) in which weights were based on information from all contributing studies (Supplementary Table 14) and tested for association with atrial fibrillation age of onset in the HUNT Study. In agreement with our previous report[11], we found that younger atrial fibrillation age of onset was associated with a higher genetic burden of atrial fibrillation (Supplementary Fig. 8). This finding supports previous epidemiological studies indicating that the risk of atrial fibrillation increases with decreasing atrial fibrillation age of onset in close relatives[34,35].

After acceptance of this manuscript, a GWAS of AF in 65,446 cases identifying 97 loci was published online[36]. We meta-analyzed the index variants from this report with samples from HUNT, deCODE, MGI, and DiscovEHR, comprising the largest possible independent dataset (up to 93,315 cases), and identified 24 genome-wide significant loci that are independent of the 111 we report here (Supplementary Table 16). We suggest that combining the initial 111 locus index variants, the 31 additional risk identified through approximate stepwise conditional analyses, and the 24 new locus index variants ($n = 166$ variants) would comprise the current optimal variant list for a polygenic risk score for atrial fibrillation (Supplementary Table 16).

In summary, we substantially increased the number of genome-wide significant risk variants for atrial fibrillation through a large GWAS meta-analysis. We highlighted genes important for function of cardiac ion channels and calcium signaling, along with cardiac transcription factors, which in turn could also affect the electrical properties of the myocardium[37,38], and in addition prioritized multiple atrial fibrillation functional candidate genes likely to be
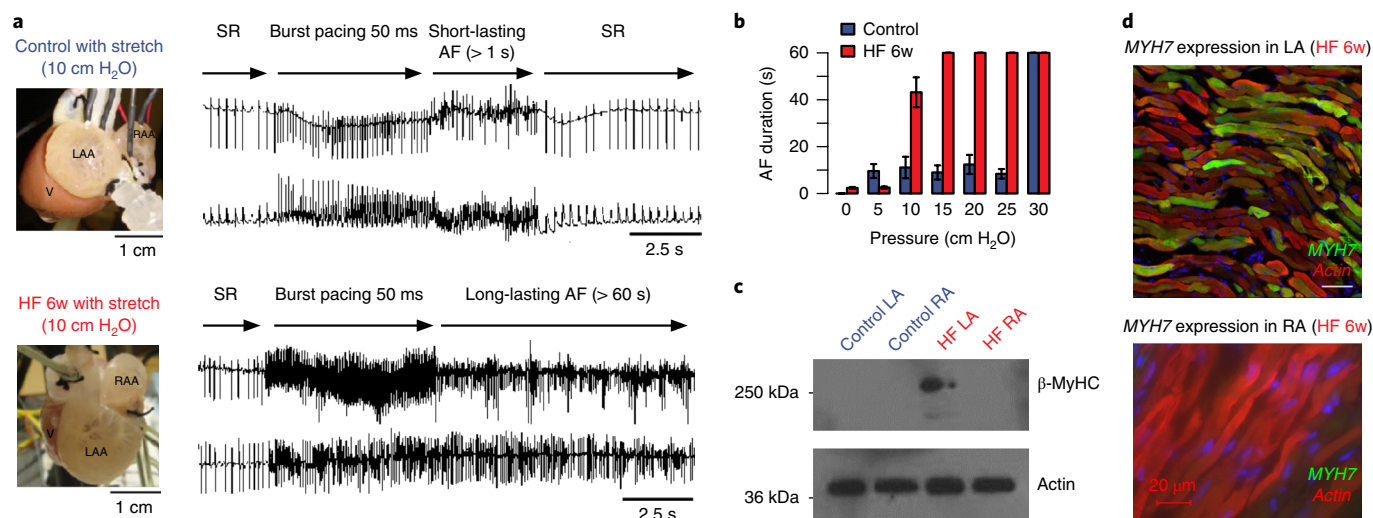
Fig. 4 | Atrial fibrillation is associated with heterogeneous changes in left atrial myosin isoform expression. a, Langendorff-perfused rabbit hearts from control (blue, top) or heart failure rabbits (red, bottom panel) were tested for atrial fibrillation inducibility and duration following burst pacing at 50 ms cycle length. Heart failure was induced by chronic left circumflex artery ligation and was allowed to develop over 6 weeks. During heart failure progression, severe left atrial hypertrophy occurred. **b**, Heart failure hearts ($n = 4$) developed long-lasting atrial fibrillation (>60 s) when intra-atrial pressure was increased to 10 cm H_2O. Control hearts ($n = 4$) did not develop long-lasting atrial fibrillation until intra-atrial pressure was increased to 30 cm H_2O. The colored bars represent mean atrial fibrillation duration and the black error bars represent the corresponding standard errors of the mean. All individual data points are shown in the more detailed Supplementary Fig. 4. **c**, Western blotting for MYH7 expression (β-MyHC protein) indicates MYH7 expression exclusively in the remodeled heart failure left atrium. The experiment was repeated for two independent heart failure animals and two control animals with similar results. An uncropped version of the western blot is shown as Supplementary Fig. 5. **d**, Immunostaining and confocal microscopy revealed heterogeneous MYH7 expression (green) in the heart failure left atrium. Consistent with western blotting data, the heart failure right atrium did not express MYH7. The experiment was repeated for two independent heart failure animals with similar results. Supplementary Figure 6 shows an additional image. LAA, left atrial appendage; RAA, right atrial appendage; V, ventricle; AF, atrial fibrillation; HF, heart failure; SR, sinus rhythm; w, week; LA, left atrium; RA, right atrium.

involved in structural integrity and function of heart and skeletal muscle. We performed pathway and functional enrichment analyses that highlighted fetal heart tissue and pathways related to cardiac development as important for developing atrial fibrillation. This might reflect that atrial fibrillation risk variants are acting in the developing heart or that the variants are important for reactivating fetal genes or pathways as a response to stress in the adult heart. We demonstrated an example of the latter; experiments in rabbits with heart failure and left atrial dilation identified a heterogeneous distributed molecular switch from the adult to the fetal isoform of myosin heavy chain, which resulted in contractile and functional heterogeneity that may predispose to initiation and maintenance of atrial fibrillation. These findings need confirmation but provide a foundation for directing future functional experiments to better understand the biology underlying atrial fibrillation.

URLs. GotCloud, https://genome.sph.umich.edu/wiki/GotCloud; Michigan Imputation Server, https://imputationserver.sph.umich.edu/index.html; METAL, http://genome.sph.umich.edu/wiki/METAL_Documentation; PLINK1.9, https://www.cog-genomics.org/plink/1.9; DEPICT, https://data.broadinstitute.org/mpg/depict/; COJO-GCTA software, http://cnsgenomics.com/software/gcta/; Roadmap Epigenomics project, http://www.roadmapepigenomics.org/; GTEx database, http://gtexportal.org; GREGOR, http://csg.sph.umich.edu/GREGOR/; Unified Medical Language System, https://www.nlm.nih.gov/research/umls/.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0171-3.

References

1. Jin, S. C. et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017).
2. Goette, A. et al. EHRA/HRS/APHRS/SOLAECE expert consensus on atrial cardiomyopathies: definition, characterization, and clinical implication. *EP Eur.* **18**, 1455–1490 (2016).
3. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
4. Wu, Y., Zheng, Z., Visscher, P. M. & Yang, J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biol.* **18**, (2017).
5. Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R. & Smoller, J. W. Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet.* **13**, e1006711 (2017).
6. Christophersen, I. E. et al. Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nat. Genet.* **49**, 946–952 (2017).
7. Thorolfsdottir, R. B. et al. A missense variant in *PLEC* increases risk of atrial fibrillation. *J. Am. Coll. Cardiol.* **70**, 2157–2168 (2017).
8. Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
9. Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
10. Orr, N. et al. A mutation in the atrial-specific myosin light chain gene (*MYL4*) causes familial atrial fibrillation. *Nat. Commun.* **7**, 11303 (2016).
11. Nielsen, J. B. et al. Genome-wide study of atrial fibrillation identifies seven risk loci and highlights biological pathways and regulatory elements involved in cardiac development. *Am. J. Hum. Genet.* **102**, 103–115 (2018).
12. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
13. Schmidt, E. M. et al. GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* **31**, 2601–2606 (2015).

14. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
15. Wells, A. et al. The anatomical distribution of genetic associations. *Nucleic Acids Res.* **43**, 10804–10820 (2015).
16. Noguchi, S. et al. Mutations in the dystrophin-associated protein γ-sarcoglycan in chromosome 13 muscular dystrophy. *Science* **270**, 819–822 (1995).
17. Brauch, K. M. et al. Mutations in ribonucleic acid binding protein gene cause familial dilated cardiomyopathy. *J. Am. Coll. Cardiol.* **54**, 930–941 (2009).
18. Gerull, B. et al. Mutations in the desmosomal protein plakophilin-2 are common in arrhythmogenic right ventricular cardiomyopathy. *Nat. Genet.* **36**, 1162–1164 (2004).
19. Skov, M. W. et al. Association between heart rate at rest and incident atrial fibrillation (from the Copenhagen Electrocardiographic Study). *Am. J. Cardiol.* **118**, 708–713 (2016).
20. Nielsen, J. B. et al. P-wave duration and the risk of atrial fibrillation: results from the Copenhagen ECG Study. *Heart Rhythm* **12**, 1887–1895 (2015).
21. Nielsen, J. B. et al. Risk of atrial fibrillation as a function of the electrocardiographic PR interval: Results from the Copenhagen ECG Study. *Heart Rhythm* **10**, 1249–1256 (2013).
22. Nielsen, J. B. et al. J-shaped association between QTc interval duration and the risk of atrial fibrillation: results from the Copenhagen ECG study. *J. Am. Coll. Cardiol.* **61**, 2557–2564 (2013).
23. Holm, H. et al. A rare variant in *MYH6* is associated with high risk of sick sinus syndrome. *Nat. Genet.* **43**, 316–320 (2011).
24. Bjornsson, T. et al. A rare missense mutation in *MYH6* confers high risk of coarctation of the aorta. Preprint at *bioRxiv* https://doi.org/10.1101/180794 (2017).
25. Maron, B. J. & Maron, M. S. Hypertrophic cardiomyopathy. *Lancet* **381**, 242–255 (2013).
26. England, J. & Loughna, S. Heavy and light roles: myosin in the morphogenesis of the heart. *Cell. Mol. Life Sci.* **70**, 1221–1239 (2013).
27. Herron, T. J., Korte, F. S. & McDonald, K. S. Loaded shortening and power output in cardiac myocytes are dependent on myosin heavy chain isoform expression. *Am. J. Physiol. Heart Circ. Physiol.* **281**, H1217–H1222 (2001).
28. Miyata, S., Minobe, W., Bristow, M. R. & Leinwand, L. A. Myosin heavy chain isoform expression in the failing and nonfailing human heart. *Circ. Res.* **86**, 386–390 (2000).
29. Cañón, S. et al. miR-208b upregulation interferes with calcium handling in HL-1 atrial myocytes: Implications in human chronic atrial fibrillation. *J. Mol. Cell. Cardiol.* **99**, 162–173 (2016).
30. Wagner, A. H. et al. DGIdb 2.0: mining clinically relevant drug–gene interactions. *Nucleic Acids Res.* **44**, D1036–D1044 (2016).
31. Teerlink, J. R. et al. Dose-dependent augmentation of cardiac systolic function with the selective cardiac myosin activator, omecamtiv mecarbil: a first-in-man study. *Lancet* **378**, 667–675 (2011).
32. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
33. Fritsche, L. G. et al. Association of polygenic risk scores for multiple cancers in a phenome-wide study: results from The Michigan Genomics Initiative. *Am. J. Hum. Genet.* **102**, 1048–1061 (2018).
34. Lubitz, S. A. et al. Association between familial atrial fibrillation and risk of new-onset atrial fibrillation. *JAMA* **304**, 2263–2269 (2010).
35. Oyen, N. et al. Familial aggregation of lone atrial fibrillation in young persons. *J. Am. Coll. Cardiol.* **60**, 917–921 (2012).
36. Roselli, C. et al. Multi-ethnic genome-wide association study for atrial fibrillation. *Nat. Genet.* https://doi.org/10.1038/s41588-018-0133-9 (2018).
37. Costantini, D. L. et al. The homeodomain transcription factor Irx5 establishes the mouse cardiac ventricular repolarization gradient. *Cell* **123**, 347–358 (2005).
38. Veerman, C. C. et al. The Brugada Syndrome susceptibility gene *HEY2* modulates cardiac transmural ion channel patterning and electrical heterogeneity. *Circ. Res.* **121**, 537–548 (2017).

Author contributions

J.B.N., R.B.T., L.G.F., W.Z., M.W.S., S.E.G., S.M., E.M.S., G.S., I.S., M.L., B.N.W., R.D., P.S., U.T., and X.W. performed the computational analyses. M.R.M., M.E.G., A.H.S., O.L.H., H.D., J.H.C., J.D.B., D.O.A., U.T., A.B., C.O., A.G.H., W.H., S.K., C.M.B., and T.M.T. conducted data acquisition. T.J.H., M.Y., R.D.C., J.K., J.A.S., and J.J. performed wet lab experiments. O.L.H., F.E.D., M.B., S.L., H.M.K., H.H., D.J.C., D.F.G., K.S., B.M., G.R.A., K.H., and C.J.W. designed and supervised the study. All authors contributed to manuscript preparation and read, commented on, and approved the manuscript.

Competing interests

R.B.T., G.S., D.O.A., P.S., U.T., D.F.G., H.H., and K.S. are employed by deCODE genetics/Amgen, Inc., Reykjavik, Iceland. A.G.H. is employed by Novo Nordisk A/S, Bagsværd, Denmark. S.M., J.H.C., J.D.B., A.B., C.O., F.E.D., G.R.A., and T.M.T. are employed by Regeneron Pharmaceuticals, Inc., Tarrytown, New York, USA. D.J.C. is employed by Geisinger Health System, Danville, Pennsylvania, USA.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0171-3.

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

Correspondence and requests for materials should be addressed to G.R.A. or K.H. or C.J.W.

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

Methods

Discovery cohorts. Additional details on selected cohorts are provided in the Supplementary Note.

HUNT. The Nord-Trøndelag Health Study (HUNT) is a population-based health survey conducted in the county of Nord-Trøndelag, Norway, since 1984[39]. We used a combination of hospital, out-patient, and emergency room discharge diagnoses (ICD-9 and ICD-10) to identify 6,493 atrial fibrillation cases and 63,142 atrial fibrillation-free controls with genotype data. Participation in the HUNT Study is based on informed consent, and the study has been approved by the Data Inspectorate and the Regional Ethics Committee for Medical Research in Norway.

deCODE. The Icelandic atrial fibrillation population consisted of all patients diagnosed with atrial fibrillation (ICD-10 code I48 and ICD-9 code 427.3) at Landspitali, The National University Hospital, in Reykjavik, and Akureyri Hospital (the two largest hospitals in Iceland) from 1987 to 2015. All atrial fibrillation cases, a total of 13,471, were included. Controls were 358,161 Icelanders recruited through different genetic research projects at deCODE genetics, excluding those in the atrial fibrillation cohort. The study was approved by the Icelandic Data Protection Authority and the National Bioethics Committee of Iceland (no. VSNb2015030021).

MGI. MGI is a hospital-based cohort collected at Michigan Medicine, USA. Atrial fibrillation cases ($n = 1,226$) were defined as patients with ICD-9 billing code 427.31, and controls were individuals without atrial fibrillation, atrial flutter, or related phenotypes (ICD-9 426-427.99). MGI was reviewed and approved by the Institutional Review Board of the University of Michigan Medical School.

DiscovEHR. The DiscovEHR collaboration cohort is a hospital-based cohort including 58,124 genotyped individuals of European ancestry from the ongoing MyCode Community Health Initiative of the Geisinger Health System, USA[40]. Atrial fibrillation cases ($n = 6,679$) were defined as DiscovEHR participants with at least one electronic health record problem list entry or at least two diagnosis code entries for two separate clinical encounters on separate calendar days for ICD-10 I48: atrial fibrillation and flutter. Corresponding controls ($n = 41,803$) were defined as individuals with no electronic health record diagnosis code entries (problem list or encounter codes) for ICD-10 I48. The Study was approved by the Geisinger Institutional Review Board.

UK Biobank. The UK Biobank is a population-based cohort collected from multiple sites across the United Kingdom[32]. Cases of atrial fibrillation were selected using ICD-9 and ICD-10 codes for atrial fibrillation or atrial flutter (ICD-9 427.3 and ICD-10 I48). Controls were participants without any ICD-9 or ICD-10 codes specific for atrial fibrillation, atrial flutter, other cardiac arrhythmias, or conduction disorders.

AFGen Consortium. Published atrial fibrillation association summary statistics from 31 cohorts representing 17,931 atrial fibrillation cases and 115,142 controls were obtained from the authors[6].

Genotyping array, imputation, and association analysis. *HUNT.* Genotyping was performed at the Norwegian University of Science and Technology (NTNU) using the Illumina HumanCore Exome v1.0 and v1.1. Quality control was performed at the marker and sample level. A total of 2,201 individuals were whole-genome sequenced at low pass and genotype calls were generated using GotCloud pipeline (see URLs). Variants from the HUNT low-pass genomes were imputed into The Haplotype Reference Consortium (HRC) samples and vice versa to generate a single imputation reference panel of ~34,000 individuals including 2,201 HUNT study-specific samples. Imputation was performed using Minimac3, and variants with imputation $r^2 > 0.3$ were taken forward. We performed testing for association with atrial fibrillation using a generalized mixed model, including covariates birth year, sex, genotype batch, and principal components 1–4 as implemented in SAIGE[41].

deCODE. The study is based on whole-genome sequence data from 15,220 Icelanders participating in various disease projects at deCODE genetics. The sequencing was done using Illumina standard TruSeq methodology to a mean depth of 35× (s.d. 8)[9]. Autosomal SNPs and indels were identified using the Genome Analysis Toolkit version 3.4.0[42]. Variants that did not pass quality control were excluded from the analysis according to Genome Analysis Toolkit best practices. Variants identified through sequencing (SNPs and indels) were then imputed into 151,677 Icelanders genotyped using Illumina SNP chips and their close relatives (familial imputation)[43]. Variants for the meta-analysis were selected based on matching with either the 1000 Genomes Project reference panel (Phase 3) or the Haplotype Consortium reference panel[44] (based on allele, frequency, and correlation matching). Logistic regression was used to test for association between SNPs and atrial fibrillation, treating disease status as the response and allele counts from direct genotyping or expected genotype counts from imputation as covariates. Other available individual characteristics that correlate with phenotype status

were also included in the model as nuisance variables. These characteristics were: sex, county of birth, current age or age at death (first- and second-order terms included), blood sample availability for the individual, and an indicator function for the overlap of the lifetime of the individual with the time span of phenotype collection. To account for inflation in test statistics due to cryptic relatedness and stratification, we applied the method of LD score regression[45]. The estimated correction factor for atrial fibrillation based on LD score regression was 1.38 for the additive model.

MGI. Genotyping was performed at the University of Michigan using the Illumina Human Core Exome v1.0 and v1.1. Quality control was performed at the marker and sample level. Imputation of variants from the HRC reference panel was performed using the Michigan Imputation Server (see URLs), and variants with imputation $r^2 > 0.3$ were included. Association with atrial fibrillation was determined using the Firth bias-corrected logistic likelihood ratio test[46] with adjustment for age, sex, and principal components 1–4.

DiscovEHR. Aliquots of DNA were sent to Illumina for genotyping on the Human OmniExpress Exome Beadchip. All individuals of European ancestry, as determined using principal component analysis, were imputed to the HRC reference panel using the Michigan Imputation Server. Markers with imputation $r^2 > 0.3$ and MAF > 0.001 were carried forward for analysis. BOLT-LMM[47] was used to analyze BGEN dosage files, and variants were tested for association with atrial fibrillation under an additive genetic model, adjusting for sex, age, age^2, and the first four principal components of ancestry; additionally, a genetic relatedness matrix (calculated using variants with MAF > 0.001, per-genotype missing data rate $< 1\%$, and Hardy–Weinberg equilibrium $P < 10^{-15}$) was included as a random-effects variable in the model.

UK Biobank. Details on quality control, genotyping, and imputation can be found elsewhere[48]. In brief, study participants were genotyped using two very similar genotyping arrays (Applied Biosystems UK BiLEVE Axiom Array and UK BioBank Axiom Array) designed specifically for the UK Biobank. Phasing and imputation were done by the UK Biobank analysis team based on the HRC reference panel and the UK10K haplotype resource[48]. We restricted our analyses to HRC-imputed markers only as there have been reports of incorrect estimates for non-HRC markers in the first 500,000 people-release from UK Biobank. We performed testing for association with atrial fibrillation in people of white British ancestry using a generalized mixed model including covariates birth year, sex, genotype batch, and principal components 1–4 as implemented in SAIGE[41].

Meta-analysis. We included all markers that were available for analyses in any of the six contributing studies. For the DiscovEHR that applied the BOLT-LMM mixed model, we obtained an approximation of the allelic log-odds ratio and corresponding variance from the linear model as described previously[49]. Following this, we performed meta-analyses using the inverse variance method implemented in the software package METAL (see URLs)[50]. When estimating the cross-cohort allele frequencies, we only included participating studies where individuals were sampled independent of atrial fibrillation status (HUNT, deCODE, MGI, DiscovEHR, UK Biobank). This was done to avoid sampling bias. Heterogeneity tests were performed as implemented in METAL[50].

Definition of independent loci. Independent loci were defined as genetic markers > 1 Mb and > 0.25 cM apart in physical and genomic distance, respectively, with at least one genetic variant associated with atrial fibrillation at a genome-wide significance threshold of $P < 5 \times 10^{-8}$. Loci borders were defined as the highest and lowest genomic positions within the locus reaching genome-wide significance plus an additional 1 Mb on either side.

LD estimation. We used 5,000 unrelated individuals that were randomly sampled among the HUNT Study participants for calculating LD r^2 by using the software PLINK1.9 (see URLs). We additionally used the 1000 Genomes Project phase 3 European (EUR) sample for LD estimation.

Approximate, stepwise conditional analyses. To identify independent risk variants within the identified atrial fibrillation-associated loci, we used the COJO-GCTA software (see URLs) to perform approximate, stepwise conditional analyses based on summary statistics from the meta-analyses and an LD matrix obtained from 5,000 unrelated individuals randomly sampled from the HUNT Study[3]. Only variants with MAF > 0.01 were included in the analyses and variants were only considered truly independent if they were not in LD ($r^2 < 0.10$) with the locus index variant and any of the other independent risk variants.

Estimation of heritability. The genome-wide heritability explained by all markers was estimated based on GWAS summary statistics, LD-score regression, and European-ancestry LD information from the 1000 Genomes Project[45]. The heritability explained by atrial fibrillation-associated index variants and additional independent risk variants was calculated on the basis of odds ratios and risk allele frequencies as described previously[51].

Identifying candidate functional genes, gene sets, and tissues using DEPICT.
We employed DEPICT (see URLs) to identify: (1) the most likely causal gene at associated loci, (2) reconstituted gene sets enriched for atrial fibrillation loci, and (3) tissues and cell types in which genes at associated loci that are preferentially expressed[8]. DEPICT uses gene expression data derived from a panel of 77,840 messenger RNA expression arrays[52] together with 14,461 existing gene sets defined based on molecular pathways derived from experimentally verified protein–protein interactions[53], genotype–phenotype relationships from the Mouse Genetics Initiative[54], Reactome pathways[55], KEGG pathways[56], and gene ontology terms[57]. Based on similarities across the microarray expression data, DEPICT reconstitutes the 14,461 existing gene sets by assigning each gene in the genome a likelihood of membership in each gene set. Using these precomputed gene sets and a set of trait-associated loci, DEPICT quantifies whether any of the 14,461 reconstituted gene sets are significantly enriched for genes in the associated loci and prioritizes genes that share predicted functions with genes from the other associated loci more often than expected by chance. Additionally, DEPICT uses a set of 37,427 human mRNA microarrays to identify tissues and cell types in which genes from associated loci are highly expressed (all genes residing within an LD of $r^2 > 0.5$ from index variant).

We ran DEPICT using all atrial fibrillation-associated index variants and all variants identified through stepwise conditional analyses, regardless of LD structure. For the gene sets significantly enriched for atrial fibrillation-associated loci ($P < 1 \times 10^{-6}$, FDR <0.05), we computed a weighted pairwise similarity based on the number of overlapping genes for genes with a Z score <4.75 (corresponding to $P < 1 \times 10^{-6}$) for being part of the gene set. For gene sets with no genes with a Z score <4.75, we included the three most significant genes as suggested previously[58].

Identification of regulatory elements using GREGOR. We tested for enrichment of index variants with functional domains using the software GREGOR (see URLs)[13]. This method tests for an increase in the number of atrial fibrillation-associated index variants, or their LD proxies, overlapping with the regulatory feature more often than expected by chance by comparing to permuted control sets where the variants are matched for frequency, number of LD proxies, and distance to the nearest gene. We use a saddle-point approximation to estimate the P value by comparing to the distribution of permuted statistics[13]. We ran GREGOR using all atrial fibrillation-associated index variants and all variants identified through stepwise conditional analyses, regardless of LD structure, and narrow peak BED-files from the Roadmap Epigenomics project (see URLs)[12].

Identification of expression quantitative trait loci (eQTLs) using GTEx data.
We performed eQTL look-up using the GTEx database (see URLs)[14] version 6p, which holds *cis*-eQTL expression data of up to 190,000,000 single nucleotide variants across 44 tissues, by searching for all atrial fibrillation-associated loci index variants, all independent risk variants identified from the stepwise conditional analyses, and any variants in strong LD ($r^2 > 0.80$) with these variants using an eQTL significance threshold of $P < 1.14 \times 10^{-9}$ ($5 \times 10^{-8}/44$ tissues). For all statistically significant genes, we queried all markers in the GTEx database that affected the expression of the affected genes and tested whether the eQTL markers colocalized with the GWAS signal as described previously[59].

Ischemic heart failure model of atrial fibrillation susceptibility. Ischemic heart failure was modeled using a previously described rabbit model of left circumflex artery ligation[60]. In this model, the left atrium progressively dilates following the ischemic insult as heart failure develops. Figure 4a shows images of Langendorff perfused hearts of control (sham operated) and heart failure animals highlighting the overt dilation of the left atrium in heart failure. With equivalent left atrial pressure, atrial fibrillation was induced in each condition with high frequency burst pacing as shown in the ECG traces and as described previously[61]. Protein expression analysis was performed using western blot. Use of animals was reviewed and approved by the Care and Use Committee of the University of Michigan. All pre-, intra-, and postoperative surgical procedures were developed in collaboration with veterinarians on staff in the Unit for Laboratory Animal Medicine of the University of Michigan.

Tissue-specific expression analysis (TSEA). The TSEA analyses were performed using the R software pSI package (see URLs)[15]. For the calculations, predefined pSI values provided by the pSI package creators were used. To get null distributions for the P values for the prioritized genes, we performed two sets of permutations: randomly selected from the entire human genome and randomly selected from the associated loci (also matching the number of genes picked in each of the loci), as done previously[62]. In both scenarios, 1,000 permutations were performed.

ECG-wide association analyses. ECG data were collected from Landspitali University Hospital in Reykjavik and included all ECGs obtained and digitally stored from 1998 to 2015, including a total of 434,000 ECGs from 88,217 individuals. A total of 289,297 ECGs of 62,974 individuals were sinus rhythm (heart rate 50–100 beats per minute) ECGs of individuals without the diagnosis of atrial fibrillation. The ECGs were digitally recorded with the Philips PageWriter Trim III, PageWriter 200, Philips Page Writer 50, and Phillips Page Writer 70

cardiographs and stored in the Philips TraceMasterVue ECG Management System. These were ECGs obtained in all hospital departments, from both inpatients and outpatients. Digitally measured ECG waveforms and parameters were extracted from the database for analysis. The Philips PageWriter Trim III QT interval measurement algorithm has been previously described and shown to fulfill industrial ECG measurement accuracy standards[63]. The Philips PR interval and QRS complex measurements have been shown to fulfill industrial accuracy standards[64].

We tested the 111 genome-wide significant atrial fibrillation index variants for association with 123 ECG parameters using a linear mixed effects model implemented in the Bolt software package[47], treating the ECG measurement as the response and the genotype as the covariate. All measures except heart rate and QT interval are presented for all 12 ECG leads. For this analysis, we used 289,297 sinus rhythm ECGs (heart rate 50–100 beats per minute) from 62,974 individuals who have not been diagnosed with atrial fibrillation according to our databases. This was done to assess the effect of the atrial fibrillation variants on ECG measures and cardiac electrical function in the absence of atrial fibrillation. Individuals with pacemakers were also excluded. The ECG measurements were adjusted for sex, year of birth, and age at measurement and were subsequently quantile standardized to have a normal distribution. For individuals with multiple ECG measurements, the mean standardized value was used. We assume that the quantitative measurements follow a normal distribution with a mean that depends linearly on the expected allele at the variant and a variance–covariance matrix proportional to the kinship matrix[65]. Since 123 traits were tested, the Benjamini–Hochberg FDR procedure controlling the FDR at 0.05 at each marker was used to account for multiple testing.

Polygenic risk scores. For each study participant in the UK Biobank and in the HUNT Study, we constructed an inverse normal-transformed polygenic risk score for atrial fibrillation using summarized dosage-weighted risk estimates from the list of 142 independent risk variants. For the UK Biobank risk score, risk estimates (beta coefficients) were obtained by meta-analyzing the risk variants across all contributing studies excluding the UK Biobank. To explore the association between the genetic burden of atrial fibrillation and the age of onset of atrial fibrillation, which we assumed was independent of the case status used for obtaining the risk estimates, we obtained risk estimates from meta-analyses of the full sample size.

Phenome-wide association analyses in the UK Biobank. We used a previously published scheme to define disease-specific binary phenotypes by combining hospital ICD-9 codes into hierarchical PheCodes, each representing a particular disease group[66]. ICD-10 codes were mapped to PheCodes using a combination of available maps through the Unified Medical Language System (see URLs), string matching, and manual review. UK Biobank study participants were labeled with a PheCode if they had one or more of the PheCode-specific ICD codes. Cases were all UK Biobank study participants with the PheCode of interest and controls were all UK Biobank study participants without the PheCode of interest or any related PheCodes. Sex checks were performed, so PheCodes specific for one sex could not mistakenly be assigned to the other sex. The associations between the polygenic risk score and each of the defined phenotypes were tested using a logistic regression adjusted for sex and birth year.

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

Data availability. The meta-analysis summary association statistics that support the findings of this study are available for download at http://csg.sph.umich.edu/willer/public/afib2018.

References

39. Krokstad, S. et al. Cohort Profile: The HUNT Study, Norway. *Int. J. Epidemiol.* **42**, 968–977 (2013).
40. Carey, D. J. et al. The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. *Genet. Med.* **18**, 906–913 (2016).
41. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Preprint at *bioRxiv* https://doi.org/10.1101/212357 (2017).
42. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
43. Kong, A. et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–1075 (2008).
44. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
45. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

46. Ma, C., Blackwell, T., Boehnke, M. & Scott, L. J., GoT2D investigators. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet. Epidemiol.* **37**, 539–550 (2013).

47. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

48. Bycroft, C. et al. Genome-wide genetic data on ~500,000 UK Biobank participants. Preprint at *bioRxiv* https://doi.org/10.1101/166298 (2017).

49. Cook, J. P., Mahajan, A. & Morris, A. P. Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. *Eur. J. Hum. Genet.* **25**, 240–245 (2017).

50. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

51. So, H.-C., Gui, A. H. S., Cherny, S. S. & Sham, P. C. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genet. Epidemiol.* **35**, 310–317 (2011).

52. Fehrmann, R. S. N. et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47**, 115–25 (2015).

53. Lage, K. et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309–316 (2007).

54. Bult, C. J. et al. Mouse genome informatics in a new age of biological inquiry. in *Proc. IEEE International Symposium on Bio-Informatics and Biomedical Engineering* 29–32 (IEEE, Piscataway, New Jersey, USA, 2000).

55. Croft, D. et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697 (2011).

56. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).

57. Raychaudhuri, S. et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5**, e1000534 (2009).

58. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

59. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. *PLoS Genet.* **13**, e1006646 (2017).

60. Herron, T. J. et al. Ca2+-independent positive molecular inotropy for failing rabbit and human cardiac muscle by alpha-myosin motor gene transfer. *FASEB J.* **24**, 415–424 (2010).

61. Yamazaki, M., Filgueiras-Rama, D., Berenfeld, O. & Kalifa, J. Ectopic and reentrant activation patterns in the posterior left atrium during stretch-related atrial fibrillation. *Prog. Biophys. Mol. Biol.* **110**, 269–277 (2012).

62. Ferreira, M. A. et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat. Genet.* **49**, 1752–1757 (2017).

63. Zhou, S. H., Helfenbein, E. D., Lindauer, J. M., Gregg, R. E. & Feild, D. Q. Philips QT interval measurement algorithms for diagnostic, ambulatory, and patient monitoring ECG applications. *Ann. Noninvasive Electrocardiol* **14**(Suppl.), S3–S8 (2009).

64. Lindauer, J., Gregg, R., Helfenbein, E., Shao, M. & Zhou, S. Global QT measurements in the Philips 12-lead algorithm. *J. Electrocardiol.* **38**, 90 (2005).

65. Benonisdottir, S. et al. Epigenetic and genetic components of height regulation. *Nat. Commun.* **7**, 13490 (2016).

66. Denny, J. C. et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).

nature research

Corresponding author(s): Cristen J. Willer; Kristian Hveem; Gonçalo R. Abecasis

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

▶ Experimental design

1. Sample size

   Describe how sample size was determined.

   > The sample size of this study was not predetermined. We included as many study participants as possible to improve power for discovering novel genetic associations.

2. Data exclusions

   Describe any data exclusions.

   > Some of the participating studies excluded from the control population individuals with cardiac arrhythmias/conduction disorders other than atrial fibrillation. This was predetermined.

3. Replication

   Describe the measures taken to verify the reproducibility of the experimental findings.

   > We replicated 89% of the previously reported atrial fibrillation risk loci and we found that the effect sizes of the index variants that we report were homogeneous across all included studies. This support a high external validity of our findings. Further, for many of the reported index variants, we found an association with one or more ECG traits, reflecting cardiac physiology and anatomy, and hence support that the variants we report are important for cardiac function and might be involved in atrial fibrillation. Finally, we performed a number of down-stream pathway and enrichment analyses analyses, all pointing to relevant biology and hence further support our findings.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   > Randomization was not relevant to the human data due to the observational study design. For the rabbit experiment, rabbits were randomly allocated into a control group (n=4) and an intervention group (n=4).

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   > Blinding was not relevant to this observational study. Participants were analyzed according to observed disease status (atrial fibrillation yes vs. no).

   Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

    For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a | Confirmed

☐ ☒ The <u>exact sample size</u> (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)

☐ ☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ ☒ A statement indicating how many times each experiment was replicated

☐ ☒ 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 any assumptions or corrections, such as an adjustment for multiple comparisons

☐ ☒ Test values indicating whether an effect is present
Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.

☐ ☒ A clear description of statistics including <u>central tendency</u> (e.g. median, mean) and <u>variation</u> (e.g. standard deviation, interquartile range)

☐ ☒ Clearly defined error bars in <u>all</u> relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

▶ Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

> The softwere most essential to this study were: METAL 2011-03-25, GCTA 1.91.1, Plink 1.9, SAIGE 0.19, BOLT-LMM 2.3.1, EPACTS 3.3, Genome Analysis Toolkit version 3.4.0, DEPICT 2014-07-21, GREGOR 1.4.0, R software pSI package 1.1, Philips TraceMasterVue ECG Management System, R 3.4.3, Python 3.4.5, BAFRegress 0.9.3, Illumina GenomeStudio (module 1.9.4, algorithm GenTrain 2.0), LASER 2.0, Minimac3, SQLite3.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

▶ Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

> No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

> No antibodies used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

> No eukaryotic cell lines were used.

b. Describe the method of cell line authentication used.

> No eukaryotic cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination.

> No eukaryotic cell lines were used.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

> No commonly misidentified cell lines were used.

▶ Animals and human research participants

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

> We used New Zealand white male rabbits, 1-2kg in size and about 1 year old. A total of 8 rabbits were used, including 4 controls (sham operated) and 4 in which heart failure was induced.

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

> More that one million participants were recruited via various genetic studies. All participants were older than 18 years of age. Both sexes were included. All patients gave written informed consent. The study was conducted in accordance with the Declaration of Helsinki.