nature genetics

SUPPLEMENTARY INFORMATION

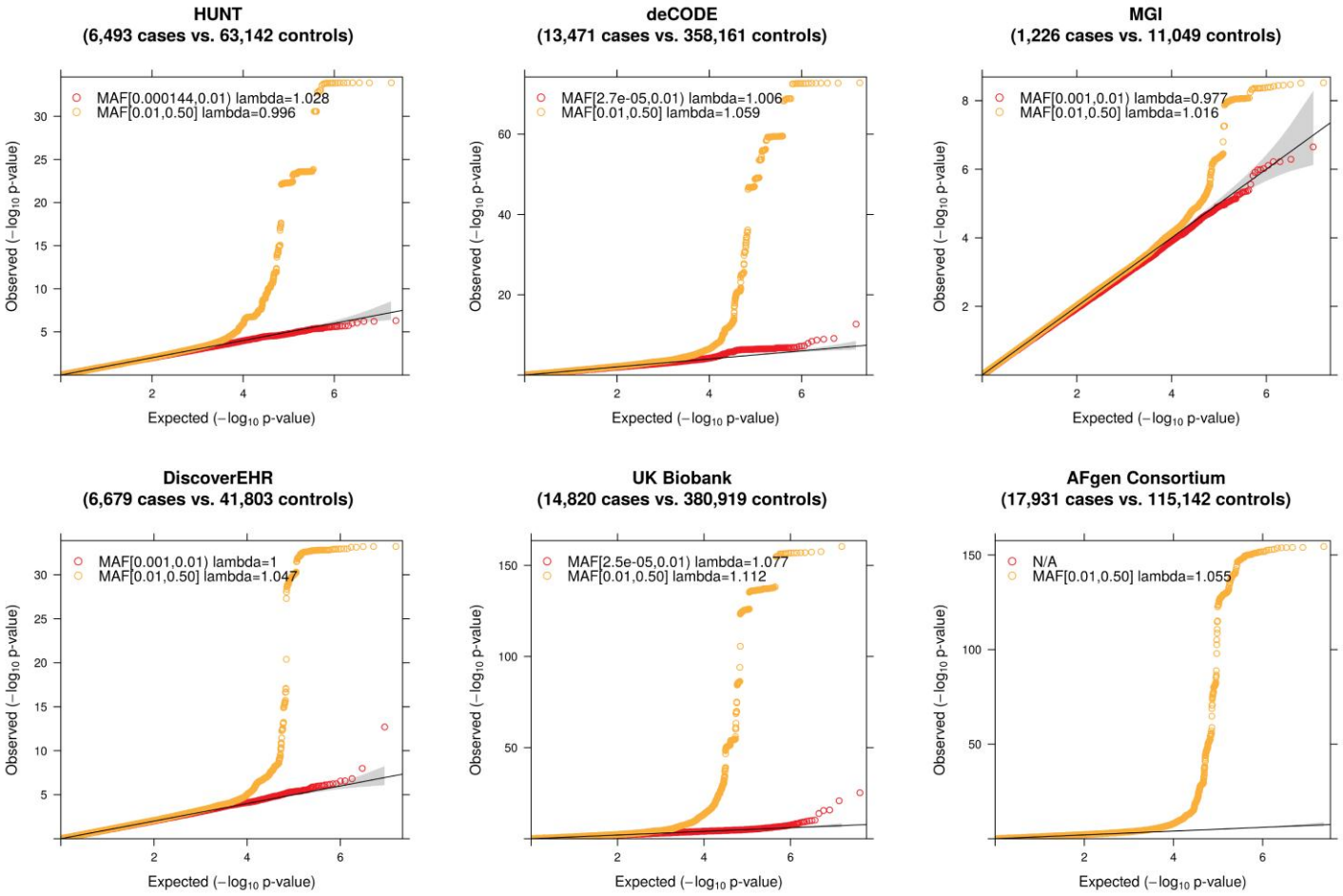
https://doi.org/10.1038/s41588-018-0171-3

In the format provided by the authors and unedited.

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*]

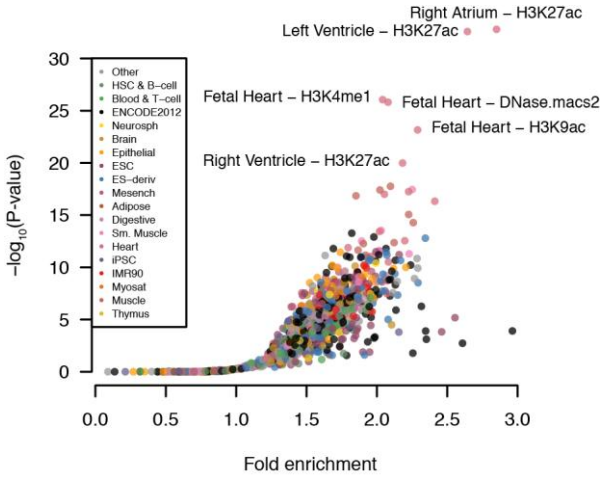
[1]Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor, MI, USA. [2]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA. [3]deCODE genetics/Amgen, Inc., Reykjavik, Iceland. [4]Faculty of Medicine, University of Iceland, Reykjavik, 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



Supplementary Figure 1

Quantile–quantile plots for genome-wide single-variant association analyses for the six contributing study cohorts.

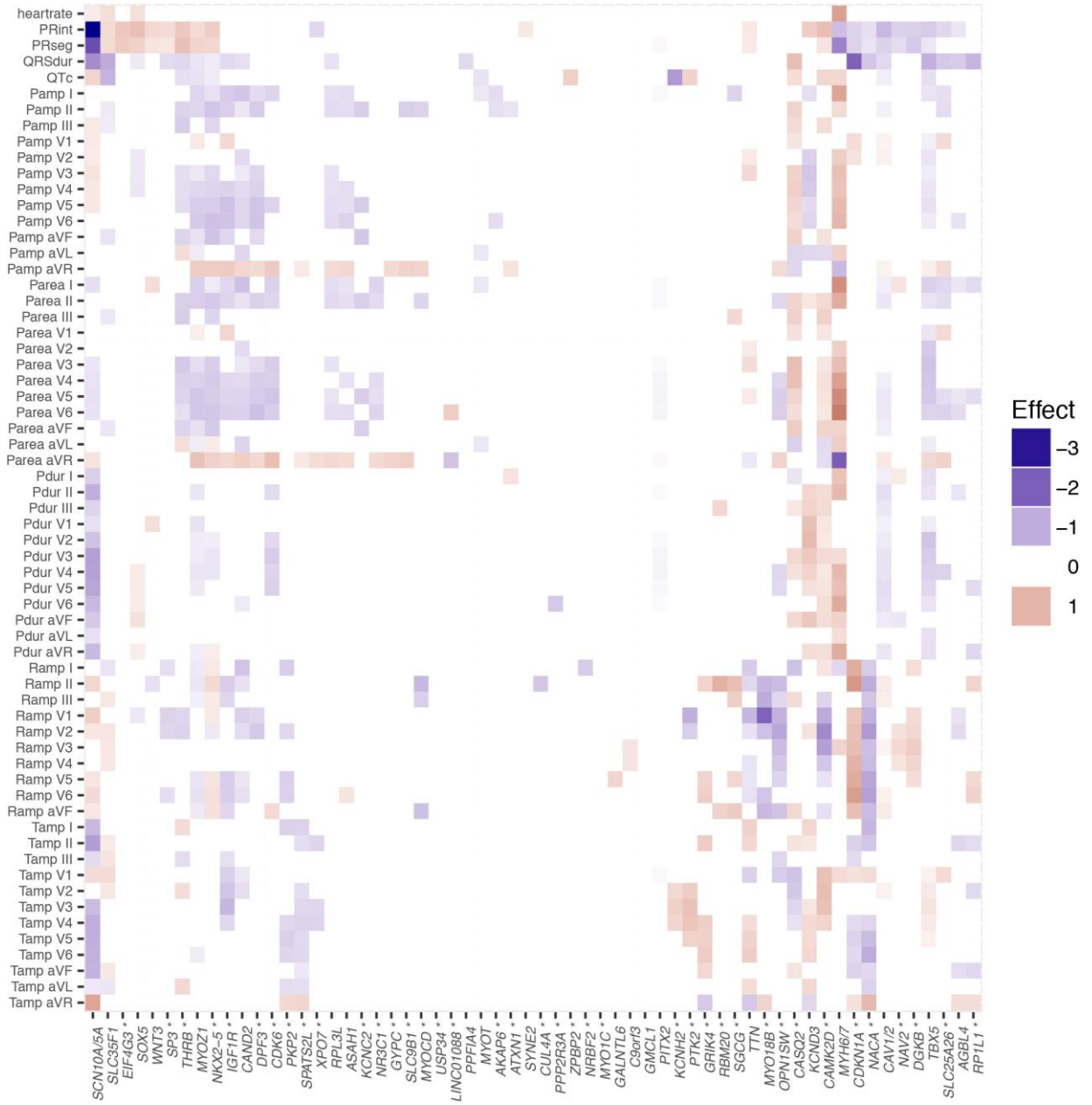
Markers are stratified by minor allele frequency below versus above 0.01. A genomic control factor of 1.38 was applied to the deCODE association results. Dots indicate observed *P* values (–$\log_{10}$ (*P* value)) compared with those expected by chance under the null hypothesis (no association). The black line indicates the identity (no association) with corresponding 95% confidence intervals.



Supplementary Figure 2

Enrichment of atrial fibrillation–associated risk variants in regulatory elements across 127 Roadmap Epigenomics tissue groups.

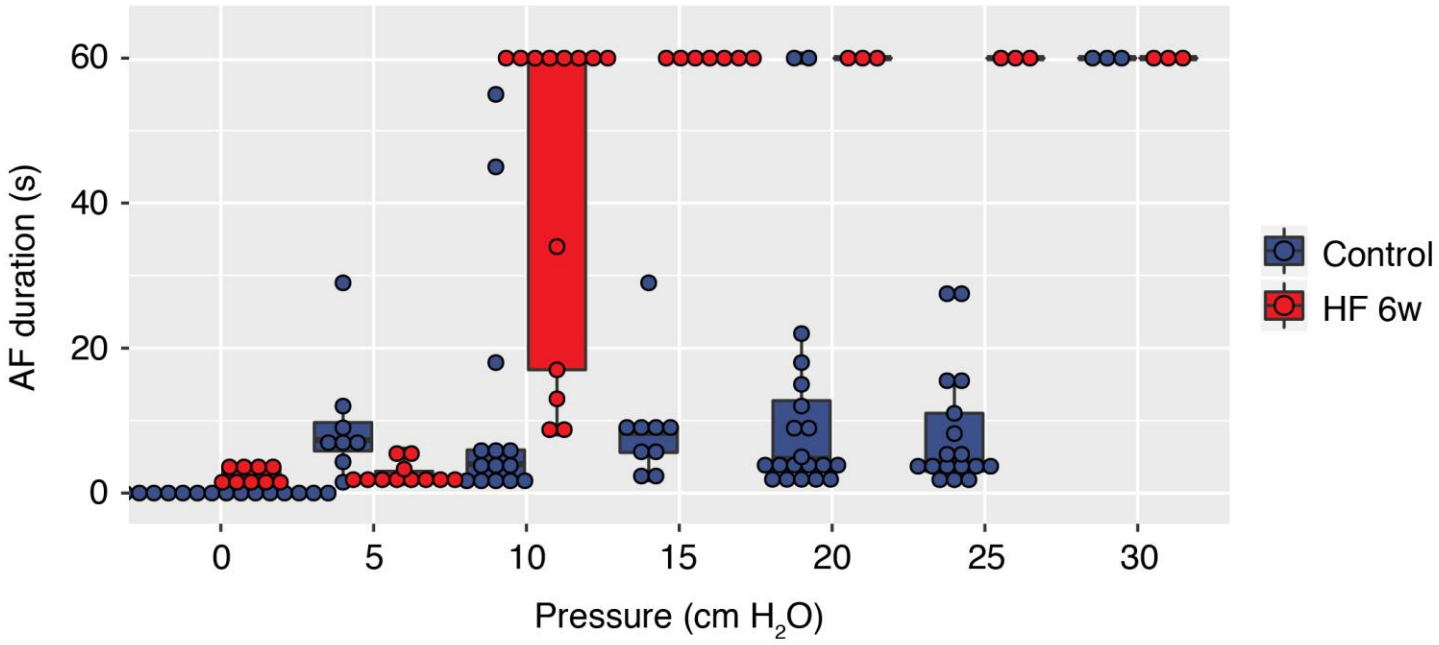
A total of 785 combinations of regulatory features and tissues were examined. *P* values and fold enrichment were estimated using GREGOR. The most statistically significant findings comprised an overlap with H3K27 in right atrium and left ventricle along with H3K4me1 and DNase sites in fetal heart.



Supplementary Figure 3

Heat map showing the effects of atrial fibrillation (AF) variants on electrocardiogram (ECG) traits in sinus rhythm ECGs, excluding AF cases.

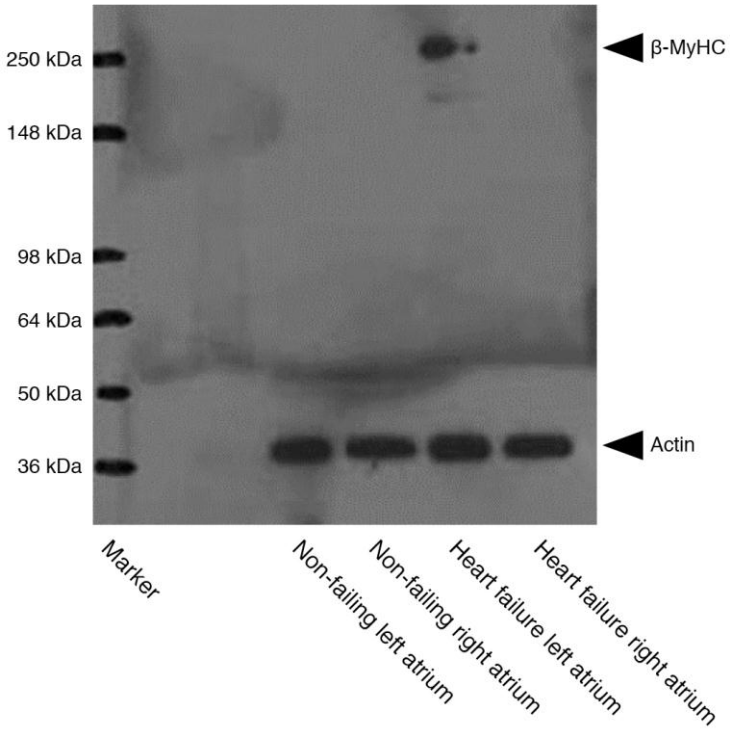
Sinus rhythm ECG measurements were available for 62,974 Icelandic individuals without diagnosis of AF. Each column shows the estimated effect of the AF risk allele on various ECG traits. The effect of each variant, annotated with the locus gene name, is scaled with the log AF odds ratio. Novel variants are marked with an asterisk. Red represents a positive effect of the AF risk allele on the ECG variable, and blue represents negative effect. The effect is shown only for significant associations after adjusting for multiple testing with a false discovery rate procedure for each variant. Non-significant associations are white in the heat map. Sixty of 111 variants with at least one association are shown. *P* values and effect estimates were obtained using BOLT-LMM. For readability, selected highly correlated lead-specific time duration ECG variables (P interval, $r^2 > 0.51$; PR segment, $r^2 > 0.46$; QRS duration, $r^2 > 0.47$; and T duration, $r^2 > 0.16$) have been omitted from the plot. A complete set of association results is provided in Supplementary Table 12. Print, PR interval; PRseg, PR segment; QRSdur, QRS interval duration; Pamp, P-wave amplitude; Parea, P-wave area; Pdur, P-wave duration; Ramp, R-wave amplitude; Tamp, T-wave amplitude.



Supplementary Figure 4

Relationship between left atrium pressure and duration of atrial fibrillation (AF) following burst pacing of rabbit hearts.

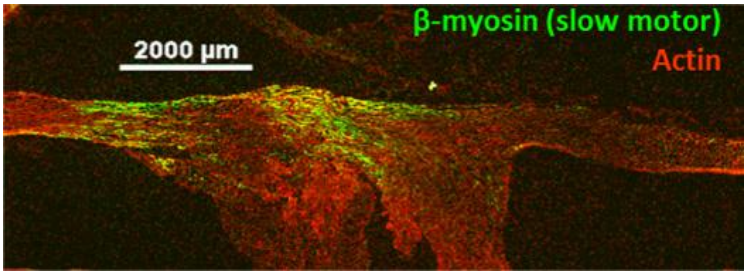
This is an extended version of Fig. 4b showing all individual data points. Heart failure (HF) hearts (*n* = 4) developed long-lasting AF (>60 s) when intra-atrial pressure was increased to 10 cm H_2O. Control hearts (*n* = 4) did not develop long-lasting AF until intra-atrial pressure was increased to 30 cm H_2O. Each individual measurement (represented by a dot) is superimposed on box plots showing the median (horizontal black lines), interquartile range (upper and lower box boarders), and interquartile range × 1.5 (vertical black lines) of AF duration.



Supplementary Figure 5

Western blotting for MYH7 expression (β-MyHC protein) indicates MYH7 expression exclusively in the remodeled heart failure left atrium.

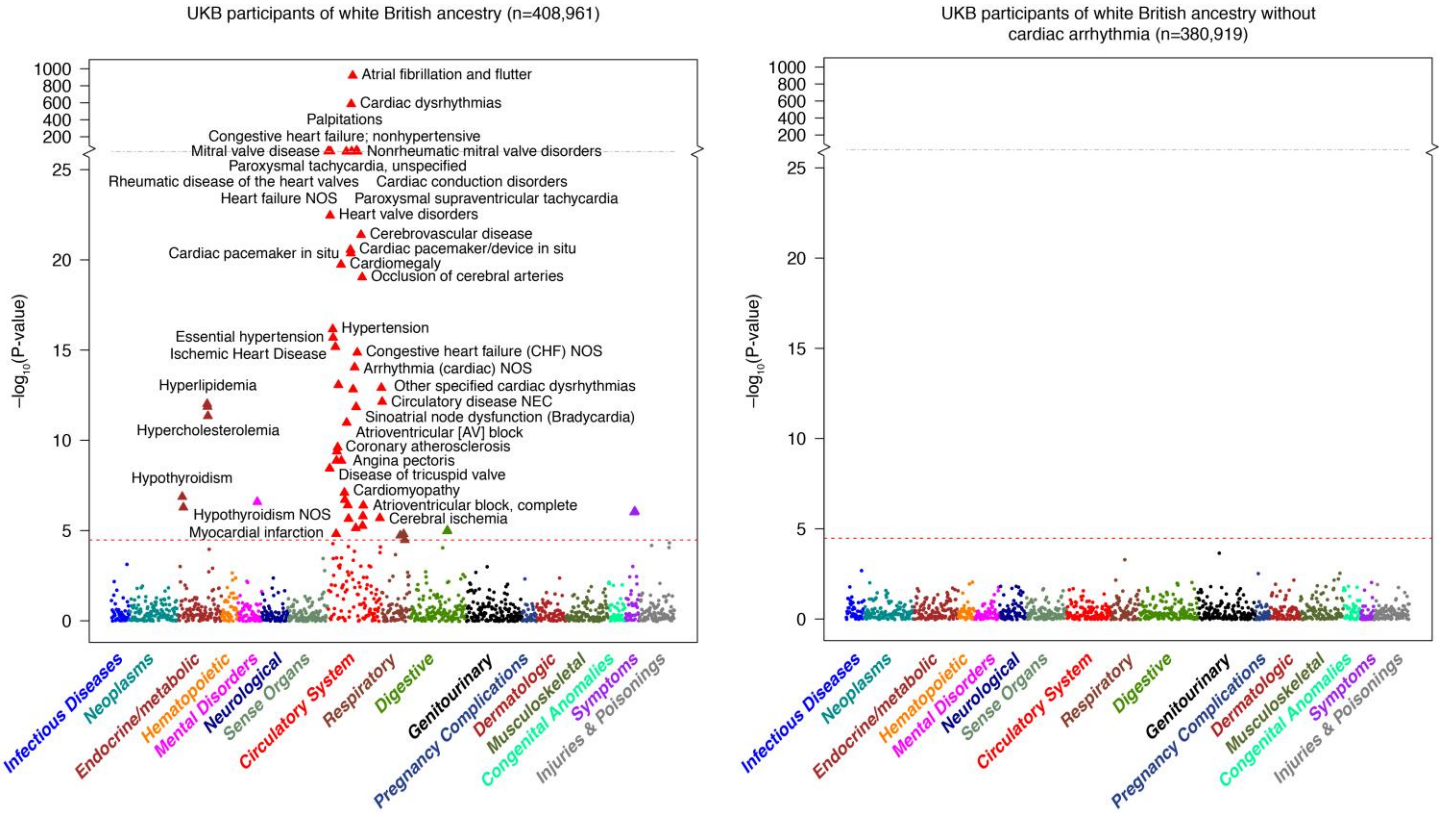
Uncropped version of Fig. 4c.



Supplementary Figure 6

Immunostaining and confocal microscopy reveal heterogeneous MYH7 expression in the heart failure left atrium.

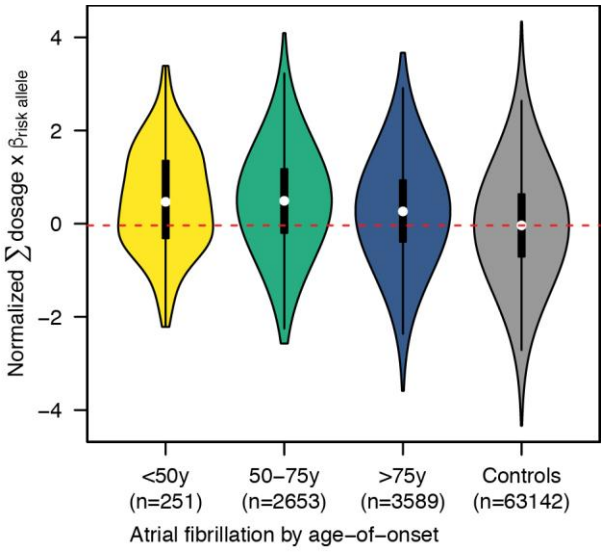
Green represents MYH7 expression (β-myosin), and red represents actin filaments.



Supplementary Figure 7

Association between atrial fibrillation polygenic risk score (*n* = 142 markers) and 1,494 ICD-based traits in UK Biobank participants of white British ancestry.

Association tests were performed using a logistic regression adjusted for sex and birth year. The horizontal dotted red line represents a *P*-value threshold of significance based on Bonferroni correction ($P < 0.05/1{,}494 = 3.3 \times 10^{-5}$). Some labels have been omitted on the left plot (see Supplementary Table 15 for details on association results).



Supplementary Figure 8

Polygenic risk score distributions for atrial fibrillation–associated variants stratified by age of onset of disease.

Results are based on the HUNT Study only. White dots represent the median, black boxes represent interquartile ranges, black whiskers are the interquartile range times 1.5, and the colored areas show the probability density of the data. The horizontal red dotted line represents the median score for controls.

Supplementary Note for the paper *"Biobank-driven genomic discovery yields new insight into atrial fibrillation biology"* by Nielsen J.B., Thorolfsdottir R.B. et. al.

Extended description of selected study cohorts

The HUNT Study

Sample ascertainment and phenotype definition

The Nord-Trøndelag Health Study (HUNT) is a population-based health survey conducted in the county of Nord-Trøndelag, Norway. Individuals were included at three different time points during approximately 20 years (HUNT1 [1984-1986], HUNT2 [1995-1997] and HUNT3 [2006-2008]).[1] At each time point, the entire adult population (≥ 20 years) was invited to participate by completing questionnaires, attending clinical examinations and interviews. Participation rates have generally been high: 89.4% (n = 77,212), 69.5% (n = 65 237) and 54.1% (n = 50 807) in HUNT1, HUNT2 and HUNT3, respectively.[1] Taken together, the health studies included information from over 120,000 different individuals from Nord-Trøndelag. Biological samples including DNA have been collected for approximately 70,000 participants. Atrial fibrillation was defined based on ICD-10 codes collected from local hospitals and out-patient clinics between 1999-2016. Cases were defined as individual with one or more ICD-9 or ICD-10 codes specific for atrial fibrillation ("I48" or "427.3") whereas controls were all individuals without a code specific for atrial fibrillation.

Genotyping, quality control, and imputation

In total, DNA from 71,860 HUNT samples was genotyped using one of three different Illumina HumanCoreExome arrays (HumanCoreExome12 v1.0, HumanCoreExome12 v1.1 and UM HUNT Biobank v1.0). We excluded samples that failed to reach a 99% call rate, had contamination > 2.5% as estimated with BAF Regress,[2] large chromosomal copy number variants, lower call rate of a technical duplicate pair and twins, gonosomal constellations other than XX and XY, or whose inferred sex contradicted the reported gender. Samples that passed quality control were analyzed in a second round of genotype calling following the Genome Studio quality control protocol described elsewhere.[3] Genomic position, strand orientation and the reference allele of genotyped variants were determined by aligning their probe sequences against the human genome (Genome Reference Consortium Human genome build 37 and revised Cambridge Reference Sequence of the human mitochondrial DNA; http://genome.ucsc.edu) using BLAT.[4] PLINK v1.90[5] was then used to exclude variants if their probe sequences could not be perfectly mapped, cluster separation was < 0.3, Gentrain score < 0.15, showed deviations from Hardy Weinberg equilibrium in unrelated samples of European ancestry with p-value <

0.0001), had a call rate < 99%, or another assay with higher call rate genotyped the same variant. Ancestry of all samples was inferred by projecting all genotyped samples into the space of the principal components of the Human Genome Diversity Project (HGDP) reference panel (938 unrelated individuals; downloaded from http://csg.sph.umich.edu/chaolong/LASER/),[6,7] using PLINK. Recent European ancestry was defined as samples that fell into an ellipsoid spanning exclusively European populations of the HGDP panel. The different arrays were harmonized by reducing to a set of overlapping variants and excluding variants that showed frequency differences > 15% between data sets, or that were monomorphic in one and had MAF > 1% in another data set. The resulting genotype data were phased using Eagle2 v2.3.[8] Imputation was performed on samples of recent European ancestry using Minimac3 (v2.0.1, http://genome.sph.umich.edu/wiki/Minimac3)[9] and a merged reference panel that was constructed by combining the Haplotype Reference Consortium panel (release version 1.1)[10] and a local reference panel based on 2,202 whole-genome sequenced HUNT study participants. A maximal set of relatively unrelated individuals (kinship coefficient < 0.0884) was chosen using KING[11] and FastIndep[12].

Association analysis

We performed testing for association with AF using a generalized mixed model including covariates birth year, sex, genotype batch, and principal components (PC) 1-4 as implemented in SAIGE.[13] PCs were computed using PLINK. Additional filters applied to the analysis included minor allele count ≥ 10 and imputation $r^2 \geq 0.3$.

Michigan Genomics Initiative (MGI)

DNA from blood samples of surgical patients at the University of Michigan Health System was genotyped on a customized Illumina HumanCore Exome array. Genotypes of the Haplotype Reference Consortium were imputed into the phased MGI genotypes, resulting in dense mapping at over 20 million variants. Atrial fibrillation cases were derived from electronic health records for patients with at least 1 encounter with ICD-9 billing code 427.31 ('Atrial fibrillation'). We performed a genome-wide association analysis for Atrial fibrillation in 1,226 cases and 11,049 controls of European ancestry (excluding 1st and 2nd degree relatives). Exclusion criteria for controls included individuals with ICD-9 codes 426-427.99. We used the Firth bias-corrected logistic likelihood ratio test with adjustment for age, sex, and principal components 1-4.

DiscovEHR Collaboration Cohort

The DiscovEHR human genetics cohort analyzed here included 58,124 consented enrollees of European ancestry from the ongoing MyCode Community Health Initiative of the Geisinger Health System ("DiscovEHR study"). Participants were recruited from outpatient primary care and specialty clinics, the cardiac catheterization laboratory, and from patient

populations referred for bariatric and abdominal vascular surgery between 2007 and 2016. Clinical laboratory measurements, International Classification of Diseases, Ninth Revision (ICD-9) disease diagnosis codes, medications, and procedural codes were extracted from the electronic health records (EHR) recording a median of 15 years of clinical care.

Cases 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 were defined as individuals with no EHR diagnosis code entries (problem list or encounter codes) for ICD-10 I48.*.

Aliquots of DNA were sent to Illumina for genotyping on the Human OmniExpress Exome Beadchip. All individuals of European ancestry, as determined using principal components (PC) analysis, were imputed to the HRC Reference Panel using the Michigan Imputation Server. Markers with imputation $r^2 > 0.3$ were carried forward for analysis with no filtering for MAF. BOLT-LMM 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 PCs of ancestry; additionally, a genetic relatedness matrix (calculated using variants with MAF > 1%, per-genotype missing data rate < 1%, and HWE P-value < 1×10^{-15}) was included as a random-effects variable in the model.

Supplementary references

1. Krokstad, S. *et al.* Cohort profile: The HUNT study, Norway. *Int. J. Epidemiol.* **42,** 968–977 (2013).
2. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91,** 839–848 (2012).
3. Guo, Y. *et al.* Illumina human exome genotyping array clustering and quality control. *Nat. Protoc.* **9,** 2643–2662 (2014).
4. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
5. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4,** 7 (2015).
6. Wang, C. *et al.* Ancestry estimation and control of population stratification for sequence-based association studies. *Nat. Genet.* **46,** 409–415 (2014).

7. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319,** 1100–1104 (2008).

8. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48,** 1443–1448 (2016).

9. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48,** 1284–1287 (2016).

10. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48,** 1279–1283 (2016).

11. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinforma. Oxf. Engl.* **26,** 2867–2873 (2010).

12. 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).

13. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *bioRxiv* 212357 (2017). doi:10.1101/212357