The genetic makeup of the electrocardiogram

Niek Verweij[a,b], Jan-Walter Benjamins[a], Michael P. Morley[c], Yordi J. van de Vegte[a], Alexander Teumer[d,e], Teresa Trenkwalder[f,g], Wibke Reinhard[f], Thomas P. Cappola[h], Pim van der Harst[a,i]

Author affiliations:
a. University of Groningen, University Medical Center Groningen, Department of Cardiology, Groningen, the Netherlands.
b. Genomics plc, Oxford, United Kingdom.
c. The Cardiovascular Institute at the Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA.
d. Institute for Community Medicine, University Medicine Greifswald, Germany.
e. DZHK (German Center for Cardiovascular Research), partner site Greifswald, Greifswald, Germany.
f. Klinik für Herz- und Kreislauferkrankungen, Deutsches Herzzentrum München, Technical University Munich, Munich, Germany.
g. DZHK (German Center for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany.
h. Division of Cardiovascular Medicine at the Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA.
i. Department of Cardiology, Heart and Lung Division, University Medical Center Utrecht, Utrecht, The Netherlands

Address for correspondence:

Dr. Niek Verweij

University of Groningen, UMCG

Department of Cardiology

PO Box 30.001, 9700 RB Groningen, The Netherlands

Email: mail@niekverweij.com

Phone: +31 (0)50 3612355

Abstract

The electrocardiogram (ECG) is one of the most useful non-invasive diagnostic tests for a wide array of cardiac disorders. Traditional approaches to analyzing ECGs focus on individual segments. Here, we performed comprehensive deep phenotyping of 77,190 ECGs in the UK Biobank across the complete cycle of cardiac conduction, resulting in 500 spatial-temporal datapoints, across 10 Million genetic variants. In addition to characterizing polygenic risk scores for the traditional ECG segments, we identified over 300 genetic loci that are statistically associated with the high-dimensional representation of the ECG. We established the genetic ECG signature for dilated cardiomyopathy, associated the *BAG3*, *HSPB7/CLCNKA*, *PRKCA*, *TMEM43*, and *OBSCN* loci with disease risk, and confirmed this association in an independent cohort. In total, our work demonstrates that a high-dimensional analysis of the entire ECG provides unique opportunities for studying cardiac biology, disease and furthering drug development.

A record of this paper's Transparent Peer Review process is included in the Supplemental Information.

Keywords: electrocardiogram, genetics, dilated cardiomyopathy, cardiovascular risk

URL: www.ecgenetics.org

Abbreviation list

ECG = Electrocardiogram

GWAS = Genome Wide Association Study

HES = Health Episode Statistics

NHS = UK National Health Service

PCA = Prinicpal component (analysis)

Introduction

Since its original description in 1893 by Willem van Einthoven, the electrocardiogram (ECG) has been instrumental in the recognition of a wide array of cardiac disorders(AlGhatrif and Lindsay, 2012; Das and Zipes, 2012). The ECG is one of the most useful non-invasive diagnostic tests for many cardiac disorders as it informs about the presence of cardiac structure abnormalities such as chamber enlargement, conduction disturbances and arrythmias. It also provides indication of electrolyte imbalance and drug toxicity. So far, genetic studies of the ECG identified more than hundred genetic loci(Arking et al., 2014; van den Berg et al., 2017; Bihlmeyer et al., 2018; Eppinga et al., 2016; van der Harst et al., 2016; Nielsen et al., 2018; Prins et al., 2018; Roselli et al., 2018; van Setten et al., 2018; Verweij et al., 2016), but have focused around specific ECG fragments representing peak characteristics such as the PR interval, QRS duration and QT interval.

Current genome-wide association studies (GWASes) of cardiac conduction provided insights into cardiac (conduction) biology by pointing out new genetic loci and genes. However, one of the major limitations of this approach is that it ignores the high dimensional nature of the ECG. It is also challenging to compare findings across different ECG GWASes because the traits themselves are fragmented across the ECG, measured in different populations, using different methods and disseminated across different publications.

We aimed to overcome these limitations by following the effects of individual genetic variants through the complete cardiac cycle the ECG represents, in order to identify additional insights into cardiac biology and disease. To this end, we downloaded, processed over 70 thousand 3-lead ECG recordings of the UK Biobank. For each individual, we averaged the ECG recording of multiple heart beats down to 500 datapoints that represented one normal cardiac conduction cycle in rest and we refer to it as the 'ECG morphology phenotype'. The 500 datapoints, can be seen as 500 traits or timepoints, that each represent separate segments of the cardiac cycle. Aggerating these data across all individuals, on a population level, we then studied how the ECG morphology phenotype was influenced by genetics. We show that genetic variants are not bound by classical interpretations of the ECG, but have unique association signatures across the ECG, or 'genetic ECG signatures'. By exploiting aberrant genetic ECG signatures, we identified genetic loci and endo-phenotypes for cardiac disorders. Our results demonstrate how an integrated approach to analyze high-dimensional data can further our understanding of the ECG, adding to the earlier undertaken snapshot analyses of individual ECG components.

Results

An enhanced understanding of the influence of genetic variants on the complete cardiac cycle represented by the ECG could generate new hypotheses on cardiac physiology, disease and effects of drugs. To better characterize the impact of genetic variants on the ECG, we obtained all 77,190 3-lead ECGs of the UK Biobank that contained raw signal data necessary for the analysis. After individual level and population level quality control to remove abnormal beats and ECGs(Simson, 1981) (**Methods, Figure 1**), 63,706 individuals remained for the primary analyses. The primary ECG morphology phenotype was constructed by averaging the cardiac cycle that the ECG represents. We segmented the cardiac cycle into 500 data points representing ECG voltages along a 1000ms window, dictated by the 500hz ECG sampling rate. To study possible effects of heart rate, we performed additional, secondary, analyses in which we normalized that representation for the individual beat-to-beat variation (the R-R interval) and refer to it as the 'R-R adjusted' ECG morphology phenotype.

To demonstrate that this approach also captures the classical ECG traits, we used previously described genetic variants in aggregate and isolation to visualize their morphological effect on the ECG(Arking et al., 2014; van den Berg et al., 2017; Bihlmeyer et al., 2018; Eppinga et al., 2016; van der Harst et al., 2016; Nielsen et al., 2018; Prins et al., 2018; Roselli et al., 2018; van Setten et al., 2018; Verweij et al., 2016). By plotting 500 association signals of each datapoint as $-\log_{10}$ P-values along the time axis of one beat (**Figure 2**), we found that the polygenic risk score of PR-interval associated with a shift of the P-wave; the polygenetic risk score of QRS-duration associated with Q and S-wave durations; the polygenic risk score of 12-lead-sum area matched the area under the curve of the QRS-complex; and the polygenetic risk score of QT-time associated with T-wave prolongation. To obtain insights into the potential effects of heart rate on cardiac conduction and ECG amplitudes, we included the R-R adjusted ECG morphology phenotype as a secondary analysis to show the potential effect of heart rate on the genetic ECG signatures. The classical ECG traits are largely unaffected by heart rate (**Online Data Figure 1**), but the polygenic score of heart rate shows a very different genetic ECG signature for the R-R adjusted phenotype, as expected (**Figure 2**).

When individual variants of these risk scores were plotted, a plethora of different ECG morphologies was observed, suggesting they are not bound by classic ECG patterns and indicate differences in the underlying biology (**Figure 3a-b**, **Online Data Figure 2**). For atrial fibrillation some interesting examples include the major atrial fibrillation locus *PITX2* (paired-like homeodomain transcription factor) and *SYNPO2L (*synaptopodin 2-like*),* both showing a genetic ECG signature consistent with the genetic ECG signature of atrial fibrillation. The loci encompassing *TTN* and *PLN* however show very different genetic ECG signatures not captured by any of the polygenic risk scores in **Figure 2**. *TTN* encodes titin, the largest known human protein and major component of the muscle's sarcomere. *PLN* encodes phospholamban, which inhibits calcium uptake by the sarco–/endoplasmic reticulum Ca^{2+}-ATPase-2 (SERCA-2) pump, involved in atrial and ventricular relaxation.

To better understand the contribution of common genetic variation to the ECG morphology phenotype in general, we studied the SNP heritability (h^2_g). Here, we observed that electrically active (non-isoelectric) ECG points are more heritable than isoelectric points, consistent with what would be expected based on the molecular mechanisms (see **Figure 3c**). This analysis also suggested that amplitudes between the S and T wave (the ST segment) are highly heritable, suggesting this part of the ECG may be particularly interesting.

Next, we performed genome wide association studies (GWAS) on the amplitude of each of the 500 temporal data points of the ECG generating 500 association values per genetic variant (**Methods**). In total 414 independent (r^2 <0.001) genetic variants were identified in 331 2MB regions at the traditional genome wide significance threshold ($P<5x10^{-8}$) or 203 independent variants in 166 2MB regions at the stringent Bonferroni corrected significance level ($P<6.25x10^{-10}$). These 414 independent genetic variants were assigned to 584 candidate genes (**Figure 4a., Table S1-3**). Of all 331 ECG associated regions, 127 were shared (1MB region-based) with loci found in prior GWASes of classical ECG traits and 64 regions were shared with the atrial fibrillation phenotype (**Table S1 and 4**), at the time of this study. We identified 179 genetic association signals that had not been reported before(Arking et al., 2014; van den Berg et al., 2017; Bihlmeyer et al., 2018; Christophersen et al., 2017; Eppinga et al., 2016; van der Harst et al., 2016; Nielsen et al., 2018; Prins et al., 2018; Roselli et al., 2018; van Setten et al., 2018; Verweij et al., 2014, 2016). The impact of all genome wide significant genetic variants on ECG-morphology is plotted in **Online Data Figure 3,** and all genetic variants studied can be viewed online in our ECG-browser (http://www.ecgenetics.org). The three strongest observed association signals were for rs7132327 ($P=4x10^{-140}$) near *TBX3*, rs6801957 ($P=3x10^{-133}$) near SCN5A and rs2074238 ($P=2x10^{-129}$) in *KCNQ1*. The most significant Pathways by DEPICT were "heart development" and "decreased cardiac muscle contractility". The most significant tissues identified by DEPICT were "heart", "heart ventricles" and "heart atria" (**Table S5-6**).

To group genetic variants with similar main effects on the ECG we performed unbiased clustering of the normalized ECG morphology association profiles. This analysis suggested 5 subsets of genetic effects on the ECG (**Figure 4b**). Four clusters A, B, D and E contained variants that associated primarily with differences in Q-wave, P-wave, R-S Wave and T-wave morphology, respectively (**Figure 4c**). The 5th cluster (cluster C) contained a diversity of ECG morphologies affecting the edges of the ECG and influence a component of heart rate. Although genetic variants grouped into a cluster sharing morphological features, many expressed unique features within each cluster, pointing towards differences in underlying biology.

To exploit the potential for obtaining clinical insights of this resource, we focused on a particular point of the Q-wave cluster that showed strong overlap with dilated cardiopathy risk. rs2234962 in *BAG3* was among the most significant loci ($P<5x10^{-54}$) in our analysis that have not been reported previously in GWASes of the ECG. The genetic ECG signature for this variant concentrated around the Q-R upslope of the ECG at -18ms from the R-wave (**Figure 5**, http://www.ecgenetics.org). This variant was previously identified in a GWAS of dilated cardiopathy (Meder et al., 2014; Stark et al., 2010; Villard et al., 2011). Another previously unreported ECG locus was rs1763604 near

HSPB7/CLCNKA ($P<5x10^{-20}$), this variant has also been identified by the same GWAS of dilated cardiopathy and exhibiting the same ECG pattern as *BAG3* (**Figure 5**), suggesting that this particular genetic ECG signature of the Q-R wave may have something in common with dilated cardiomyopathy. *BAG3* is part of the conserved family of cyto-protective co-chaperone proteins. The gene is mainly expressed in striated muscle and the protein colocalizes with the Z-disc of the sarcomere. The locus encompassing *HSPB7 / CLCNKA* includes two potential candidate causal genes. CLCNKA encodes a renal ClC-Ka chloride channel, however *HSPB7* may be a more likely candidate gene as it is expressed in cardiomyocytes. *HSPB7*, is a member of the small heat-shock protein family. The protein is located in the intercalated disc adjacent to the Z-line of the adult cardiac muscle and is essential for maintaining integrity of myofiber structure and cardiac conduction(Liao et al., 2017). The genetic associations for rs2234962 and rs1763604 persisted in individuals not diagnosed with dilated cardiopathy or other cardiac conditions **(Table S7)**, indicating that dilated cardiomyopathy patients are not driving the associations between these variants and the ECG at -18ms.

The observation that the ECG captures information relevant to dilated cardiopathy is important, because it shows that genetic-disease associations can be found in healthy individuals using an inexpensive ECG test. Dilated cardiopathy is a severe disease with poor prognosis: the 5 year survival rate after diagnosis is only 50%, yet the disease can be asymptyomatic for a long time. One hypothesis is that the Q-R wave marks sub-clinical dilated cardiopathy and reflects the structural integrity of the heart relevant for dilated cardiopathy risk. It is notable that here is no clear single diagnosis criterium of the ECG for dilated cardiopathy (Bozkurt et al., 2016; Japp et al., 2016).

We hypothesized that this Q-R upslope feature represents a biomarker for dilated cardiopathy. To test this, we carried out an ECG-wide Mendelian randomization (fixed effects inverse variance weighted method) in UK Biobank to reveal the genetic ECG signature for dilated cardiopathy. This analysis confirmed that -18ms was most associated with dilated cardiopathy, but also revealed a secondary effect just before the Q-R wave (**Figure 5a**).

To further substantiate the hypotheses that the Q-R upslope at -18ms is associated with dilated cardiopathy, we executed a more detailed Mendelian randomization analyses using up to 47 genetic variants that are associated with amplitudes of -18ms. Here, we observed that there is a consistent relationship between the genetic ECG signature and increased risk of dilated cardiopathy (**Figure 5c.**, **Table S8-9**). These findings remained consistent in several sensitivity analyses, across the UK Biobank and the independent MAGNet study cohort. Also, exclusion of the 2 previously known variants did not change our findings (**Figure 5e.**).

In addition to the known *BAG3* and *CLCNKA* loci, we identified 3 other variants for dilated cardiopathy passing the Bonferroni significant threshold ($P<0.001$) by using the Q-R slope trait as "bait": rs9909004 ($P=3.5x10^{-4}$) in *PRKCA*, rs4685090 ($P=4.2x10^{-6}$) near *TMEM43* and rs60836537 ($P=1.4x10^{-4}$) in *OBSCN* (**Figure 5c**, **Table S8-9**). The *PRKCA* locus encodes the enzyme Protein kinase C alpha that has previously been identified tob e associated with dilated cardiomyopathy via it's role in adverse left ventricular modeling (Hu et al., 2018). *PRKCA* affects dephosphorylation of cardiac the phospholamban (*PLN*) involved in inhibiting the SERCA2 pump. Mutations in *TMEM43 (*transmembrane protein

43*)* are commonly identified as a cause for right ventricular dysplasia subtype 5 (H et al., 2017). However, the exact molecular function of *TMEM43* and role in cardiac biology is unknown. *OBSCN* encodes Obscurin, a large 720kDa multidomain protein that plays a primary role in the formation and stabilization of the sarcomere, it harbors mutations causing hypertrophic and dilated cardiopathy (Marston, 2017). In total 13 of 47 variants were nominally associated (P<0.05) with dilated cardiopathy, which is considerably more than expected by chance (P_{binomial} = 3.4 x 10^{-7}). These analyses support this new hypothesis that the Q-R slope differences are linked to dilated cardiopathy risk and that it is a powerful tool to study the genetics of this disease. Further functional studies are important to understand the nature of these associations and how this genetic ECG signature can be used in precision medicine approaches (Rosenbaum et al., 2019), used as a biomarker, used for stratification, or used to monitor progression of dilated cardiomyopathy.

Other examples of cardiac diseases in which the ECG is critical for clinical diagnosis include early repolarization. The diagnoses of this condition is based on strict criteria of passing a certain threshold on the ECG. We hypothesized that the biological underpinnings of early repolarization does not adhere to a strict binary phenomenon, and that our understanding can be enhanced by studying continuous traits as well. Recently, a GWAS of the strict binary definition of early repolarization identified one locus, rs1545300 in *KCND3* (Teumer et al., 2019). When analyzing the full ECG-morphology of rs1545300, we observed an association peak at 44ms after the R wave that coincided with the early repolarization criterion (**Figure 5b**, http://www.ecgenetics.org). We used this 44ms data-point as "bait" to identify 8 additional loci possibly relevant to early repolarization at $P < 5 \times 10^{-8}$ (**Table S10**). We tested these for association with early repolarization in the independent LifeLines cohort (**Figure 5f-g.**, **Table S8-9**). Two loci (1 known, 1 novel) were significant at the Bonferroni level: rs776293589 (*KCND3*, P= 1.05×10^{-7}) and rs1733724 (*DKK1*, $P = 1.8 \times 10^{-3}$) and 5 out of 9 variants showed a suggestive association (P<0.05), more than expected by chance ($P_{\text{binomial}} = 3.3 \times 10^{-5}$). *KCDN3* encodes Kv4.3 that is responsible for the transient outward potassium current, I(to), and important for the early phase of repolarization. *DKK1* encodes the Dickkopf WNT Signaling Pathway Inhibitor 1, and may be involved in regulating Wnt signaling that is responsible for cell fate during formation of the heart.

One of the most promintent genetic loci influencing elecotrcardiographic traits and diseases is the *SCN5A* locus. *SCN5A* encodes the alpha subunit of the main cardiac sodium channel Nav1.5 and has a critical role in regulation of cardiac electrophysiological function. SCN5A has been found causative for a variety of cardiomyopathies and cardiac arrythmias including atrial fibrillation, long QT syndrome and the Brugada syndrome. Common variants in or near SCN5A have been associated with almost all ECG traits (Arking et al., 2014; van den Berg et al., 2017; Bihlmeyer et al., 2018; Christophersen et al., 2017; Eppinga et al., 2016; van der Harst et al., 2016; Nielsen et al., 2018; Prins et al., 2018; Roselli et al., 2018; van Setten et al., 2018; Verweij et al., 2014, 2016). Because this is such a prominent locus for cardiac electrophysiology, we aimed to dissect the complex genetic architecture of *SCN5A* by leveraging the power multiple ECG traits. We performed a comprehensive conditional analysis across the entire ECG morphology phenotype by taking into account all 1,000 ECG traits simultaneously (see **Methods**). Using this multivariate approach, we identified 10 independent variants in the *SCN5A*

locus with an effect on one or more of the 500 data-points of the ECG-morphology (at $P<5x10^{-8}$). The identified variants all coincide with cardiac enhancer mark, which was more than expected by chance, $P_{hypergeometric}=1x10^{-7}$. This finding suggests that the causal common variants at *SCN5A* are likely to be of regulatory nature and provides support for the hypothesis that the identified variants may be causal. We found that these 10 independent *SCN5A* variants have different morphological ECG signatures, possibly pointing towards different biological mechanisms and consequences for the cardiac conduction system (**Online Data Figure 4, Table S11**).

Discussion

Although genetic influences of the ECG have long been established, these analyses have mainly been based on isolated segments of the ECG (Arking et al., 2014; van den Berg et al., 2017; Bihlmeyer et al., 2018; Eppinga et al., 2016; van der Harst et al., 2016; Nielsen et al., 2018; Prins et al., 2018; Roselli et al., 2018; van Setten et al., 2018; Verweij et al., 2016). In the current study, we provide the first high-dimensional analyses of the ECG that is not limited by a definition of classically defined ECG features. Studying the complete ECG in an integrative manner provided a unique opportunity to obtain more meaningful insights into biology compared to studying individual ECG fragments one at a time.

We indentified over 300 genome wide genetic loci, of which the majority have not been reported before in relation to the ECG. These loci showed unique association signatures on the ECG. For example, variants near genes encoding pottasium and calcium channels exchibited different signatures, suggesting differences in ion-channel activity during the cardiac cycle. These data provide unique insights into the biological mechanisms of cardiac physiology, and could inform *in-silico* models of cardiac conduction (Sadrieh et al., 2014).

We expect that these types of integrative analyses of genome and phenome data, leveraging many correlated traits at the same time, can be applied to other high dimensional datasets as well, such as lipidomic and metabolomics data. This will be essential to move the GWAS field forward and to provide more useful information compared to studying individual traits in isolation.

Study limitations

First, we were only able to interrogate common genetic variants in the UK Biobank, future efforts should integrate rare coding variants once the full exome sequencing data becomes available. This will be particularly relevant considering their prominent role in dilated cardiomyopathy(Rosenbaum et al., 2019) and other cardiac disorders. Secondly, the ECG device to measure cardiac conduction consisted of only 3 electrodes that were attached to the limbs, these surface ECG parameters are not specific enough to identify the exact anatomical region of the heart. However, there are currently no single large genetic datasets available of more precise measurements, such as could be derived from more sophisticated surface ECG equipment or intra-cardiac ECG measurements. Third, the individuals that our study is based on are predominantly of European ancestry, it is unclear if conclusions can be extrapolated to other ethnic groups. Fourth, this work is based on statistical analyses and does not provide functional experimental validation; the genetic variants and genes that are prioritized require further experimental validation.

Key Changes Prompted by Reviewer Comments

In response to reviewer comments we lightly modified our main text and figures to better highlight the main findings contained in them. Although the reviewers did not ask for

more work, we added the genetic ECG signature of dilated cardiomyopathy as part of Figure 5. We also added additional information about the genes in the results section. For context, the complete Transparent Peer Review Record is included within the Supplemental Information.

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 12010. We thank Ruben N. Eppinga, Tom Hendriks, M. Yldau van der Ende, Yanick Hagemeijer, Hilde Groot, Jan-Walter Benjamins and Ming W. Yeung University of Groningen, University Medical Center Groningen, Department of Cardiology, for their contributions to the extraction and processing of data in the UK Biobank. None of the mentioned contributors received compensation, except for their employment at the University Medical Center Groningen.

We thank the Centre for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high-performance computing cluster. In addition, we thank the "Medische en Informatie Technologie Systeembeheer" of the University Medical Center of Groningen for their support and maintenance of our own computing cluster.

The work of N.V was supported by NWO VENI grant 016.186.125. The work of M.P.M and T.P.C was supported by NIH grants HL105993 and HL088577. The work of AT, TT and WR was supported by the German Center for Cardiovascular Research (DZHK Shared Expertise SE081 – STATS). The funders had no role in the design and conduct of the study; collection, management, analysis and interpretation of the data; preparation, review or approval of the manuscript; or decision to submit the manuscript for publication.

Declaration of Interests

N.V. was employed by Genomics plc and is a paid consultant for Regeneron Pharmaceuticals. The other authors declare no competing financial interests.

(STAR) Methods

RESOURCE AVAILABILITY

Lead Contact

Further information and data requests should be directed to Niek Verweij (mail@niekverweij.com).

Data and Code Availability

Individual level data are available to researchers on successful application to the UK Biobank. Genome-wide summary statistics can be queried and downloaded from www.ecgenetics.org. Code to parse the data is available here at https://github.com/niekverw/ecgenetics. Further datasets generated and/or analyzed during this study are available upon reasonable request.

Materials Availability

This study did not generate new unique reagents.

METHOD DETAILS

UK Biobank individuals

Participants were recruited with an age range of 40-69 that registered with a general practitioner of the UK National Health Service (NHS). Between 2006–2010, a total of 503,325 individuals were recruited to participate. All study participants provided informed consent and the study was approved by the North West Multi-center Research Ethics Committee. Detailed methods used by UK Biobank have been described elsewhere(Bycroft et al., 2018). The prevalence and incidence of cardiac conditions and events were captured by data collected at the Assessment Centre in-patient Health Episode Statistics (HES) download on September 10, 2017.

Genotyping and imputation

The Wellcome Trust Centre for Human Genetics performed quality control before imputation, and imputed the genotypes to the HRC v1.1 reference panel that was released on March 7th 2018. For the current analysis, individual sample outliers excluded based on heterozygosity and missingness were excluded, as well as those with gender discrepancies between the reported and inferred gender using X-chromosome heterozygosity test. A detailed description of the quality control and imputation, is available elsewhere(Bycroft et al., 2018).

ECG morphology phenotypes and quality control

Three-lead exercise ECG data were provided by the UK Biobank as bulk in separate xml-files. To isolate R-waves, we employed the gQRS algorithm by George Moody, the construe algorithm was used to further refine the localization of R-waves (Teijeiro et al., 2018). Individual ECG beats were processed and averaged (Simson, 1981), which we refer to as the 'ECG-morphology' phenotype.

Two entities of ECG morphology were constructed: The primary morphology trait was defined as the classical signal averaged electrocardiographic beat (Simson, 1981) that consisted of an averaged 1000ms window surrounding the R wave at a resolution of 500hz resulting in 500 averaged data points or ECG traits for each individual. As such, these beats were unadjusted for individual R-R intervals. The secondary 'ECG morphology' trait consisted of R-R intervals made of equal length (500 data points) so that the resulting averaged ECG beat was adjusted for each individual R-R interval.

Only information of the rest phase was used, as defined by the first 15 seconds of ECG assessment. From all 99,539 3-lead ECGs recorded in 96,567 participants, 77,190 ECGs contained full disclosure data necessary to detect R wave. The R-wave is traditionally used as a reference point to detect all other points on the ECG, hence this should be sufficient to identify major changes ECG-beat, while also easy to understand and visualize.

Before signal averaging, ECG beats were quality controlled on the individual level. First, the 3-leads were averaged to create a single ECG signal vector. Then, individual beats containing excess noise were removed as described previously(van de Vegte et al., 2018; Verweij et al., 2018), this was based on a moving standard deviation using a window-size of 3 data points under the assumption that ECG signals without noise have a moving standard deviation close to 0. Thirdly, individual beats were matched on a template and

discarded if they were dissimilar based on a Pearson correlation function between each beat; beats with mean negative correlations and those that fell outside the standard 1.5 interquartile range rule were removed. This procedure was repeated until no outlier beats were left to exclude. At least 6 beats were required at any stage of the averaging process; otherwise the entire electrocardiogram was excluded from the analysis. We explored whether the averaging process was improved when accounting for the lag at which the cross-correlation function between each ECG beats shows its maximum, but this did not make a difference; suggesting that the R-wave detection was already sufficient. In total 67,440 ECGs of 66,240 individuals passed the individual level quality control.

Finally, to further detect and exclude abnormal ECGs on the population level, we calculated the standard deviation of the difference between each averaged ECG beat of an individual and the population-mean averaged ECG beat; outliers were discarded according to the standard 1.5 interquartile range rule per ECG-phenotype. Observations of the second follow-up visits were used when no baseline observation was available. This resulted in the inclusion of 63,706 individuals in the primary analysis (unadjusted R-R intervals). The 2,534 excluded individuals based on population level QC were much more likely to be diagnosed with a history of bundle branch block (OR=16.8, se= 3.1, z=15.41), cardiomyopathy (OR=8.8 se=2.3, z=8.58), myocardial infarction (OR=2.9, se=0.27, z=12.2), atrial fibrillation (OR=2.0, se=0.25, z=5.6), and heart failure (OR=7.3, se=1.2,z=12.55). For the secondary morphology phenotype (where we adjusted for R-R intervals), 65,183 individuals remained.

Another 3,750 individuals with a history of myocardial infarction, atrial fibrillation, heart failure, cardiomyopathy, bundle-branch-block or pacemaker implantation were excluded in sensitivity analyses and Mendelian randomization analyses of dilated cardiomyopathy.

Diagnoses in UK Biobank were defined as follows: myocardial infarction (ICD10: I21, I22, I23, I252; ICD9: 410, 412 or self-reported myocardial infarction), atrial fibrillation (ICD10:I48; ICD9: 4273, OPCS-4: K621, K622, K623 or self-reported atrial fibrillation or flutter), heart failure (ICD10: I50, I110, I130, I132; ICD9: 428 or self-reported heart failure), cardiomyopathy (ICD10: I42, I43, O903, I255, O944; ICD9: 425 or self-reported cardiomyopathy or hypertrophic cardiomyopathy), dilated cardiomyopathy (ICD10:I42.0 or ICD9: 4254), bundle-branch-block (ICD10: I447, I451 or ICD9: 4263, 4264) or pacemaker implantation (OPCS-4: K59, K60, K61 or self-reported pacemaker or defibrillator insertion). Diagnoses in UK Biobank were extracted using https://github.com/niekverw/ukpheno.

Polygenic scores of ECG-traits and known genetic variants

Polygenic risk scores were created following an additive model for atrial fibrillation, QT-interval, QRS duration, PR-interval, QRS-voltage traits (12-lead sum area, cornel area and Sokolow-Lyon area) and heart rate separately (**Table S13**), as previously described(Verweij et al., 2017). In short, the number of alleles for each individual (0, 1 or 2) was summed after multiplying the alleles with the previously reported effect size of the variant-trait association. Effect sizes estimated in UK Biobank data were avoided to reduce potential overestimation. If multiple effect sizes were available, those estimated in the largest sample size were used (e.g., the combined replication and discovery phase). Single-nucleotide polymorphisms were excluded if they were missing in UK Biobank data. In instances where multiple correlated variants in the same locus were reported for

the same trait, we used only independent variants that were selected by the linkage disequilibrium clumping procedure (at $r^2 < 0.01$) implemented in PLINK version 1.9.

Pathway analyses and candidate genes

DEPICT was employed to discover pathways, tissues and genes underlying GWAS-loci of the electrocardiographic morphology phenotype. Please see Pers *et al.* for a detailed description of the method(Pers et al., 2015). All independently associated variants ($r^2>0.005$) passing a traditional genome wide significance threshold $P < 5 \times 10^{-8}$ were used as input to the DEPICT framework. DEPICT was run using default settings. Because DEPICT uses 1000 Genomes as a reference panel which does not include certain UK Biobank specific variants, we also included all variants in LD ($r^2>0.8$) with the input-variants to ensure 100% coverage of the signal. Pairwise correlations across all significantly enriched pathways (false discovery rate > 0.5%) were computed and visualized using the Gephi software (www.gephi.org) after filtering out edges with correlations lower than 0.5.

For the differential pathway analysis, we repeated the DEPICT pathway analysis but excluded variants that were of interest (variants-of-interest). We calculated the reduction in pathway enrichment by subtracting the $-\log_{10}(P)$ for enrichment before and after excluding variants-of-interest. A null distribution was created by repeating the DEPICT run 100 times on the same set of variants but each time excluding a matched number of random variants not belonging to the variants-of-interest. If all 100 null runs indicated that the reduction in significance for a given pathway was more using the variants-of-interest than for the null runs (indicating $P_{differential}<0.01$), the pathway was considered to be differentially enriched for the variants-of-interest. Particularly, we were interested whether different biological pathways were active at different ECG intervals. To test this, we used the clusters of variants that were identified in the unbiased clustering approach by t-sne and k-means as 'variants-of-interest' (see *Genetics and association testing*); which grouped variants by their primary effect on the different ECG wave intervals in an unbiased manner.

Nearest gene or any gene within 10kb of the lead variants were used to annotate candidate causal genes. We also searched for coding variants in LD with the identified variants ($r^2>0.8$) to further prioritize candidate causal genes. Variants were also annotated by proximity (250kb) to known rare genetic cardiac disease genes in Orphanet, code ORPHA:97929. An additional line of evidence for candidate causal genes was performed by DEPICT, taking into account gene-gene similarities across loci(Pers et al., 2015). We specifically did not incorporate annotations from datasets such as Hi-C and eQTL, as we strongly believe these may not be informative for cardiac conduction biology and need targeted experiments on a per locus basis to interpret them (Stacey et al., 2019; Wang and Goldstein, 2020).

Dilated cardiomyopathy

The association between genetic variants and risk of dilated cardiomyopathy was tested in case-control GWASes of UK Biobank and MAGNet (Myocardial Applied Genomics Network) and combined via inverse-variance weighted meta-analysis to increase power totaling 1,375 cases and 241,325 controls.

In the UK Biobank cohort individuals with and without dilated cardiomyopathy were selected to be independent of those taking part in the ECG measurements. 744 dilated cardiomyopathy cases were identified according to ICD10 and ICD9 codes (I42.0 and 4254), 239,729 control individuals were not diagnosed with dilated cardiomyopathy and had no family history of heart disease. Association testing was performed as described under *'Genetic and association testing'*, coefficients were re-scaled to log odds.

For the MAGNet study, 631 subjects with dilated cardiomyopathy were recruited; patients were included in the study if they were diagnosed with heart failure with reduced ejection fraction (< 40%) in the absence of hypertension, primary valvular disease, or coronary artery disease from the University of Pennsylvania Health System; 1,596 controls were recruited from the same center who had no history of heart disease. All subjects provided written informed consent. Genotyping was performed on the Illumina OmniExpressExome Array and imputed using the 1000 Genomes Project with Minimac. Genotypes between cases and controls were compared using an additive genetic model adjusting for gender and two principal components of race using SNPtest; there was no genomic inflation (1.0335).

Early repolarization

To assess the clinical relevance of the ECG morphology phenotypes in early repolarization, we diagnosed individuals 1,253 with and 11,463 without early repolarization of the Lifelines cohort. For this, 12-lead ECGs were manually curated by trained cardiologists for the presence of early repolarization according to the established definition by Haisaguerre and Macfarlane(Haïssaguerre et al., 2008; Macfarlane et al., 2015). The criteria for early repolarization were defined as an elevation of the J-point (≥0.1 mV) in at least two corresponding leads above the level of QRS onset. Leads V1 to V3 were excluded from scoring to avoid confusion or overlap with Brugada syndrome or arrhythmogenic right ventricular dysplasia. The morphology of early repolarization was assessed as either notching, slurring or both. The ST segment was defined according to Tikkanen *et al.* as either rapidly ascending (concave) or horizontal descending(Tikkanen et al., 2011). ECGs with QRS duration > 120 ms or non-sinus rhythm (for example atrial fibrillation or pacemaker rhythm) were excluded from the analysis. Inter-observer correlations have been described previously(Bastiaenen et al., 2018).

Exhaustive joint conditional analysis of the SCN5A locus

The *SCN5A* locus has been reported in almost all cardiac conduction GWAS with various independent lead SNPs, it harbors many mutations causing different types of arrhythmias, and it encodes an essential sodium channel $Na_v1.5$ that is a common target of anti-arrhythmic drugs. To further increase our understanding of complex genetics at *SCN5A*, we performed an exhaustive joint conditional analysis across all ECG datapoints. Finemapping this locus and understanding the consequences of the multiple genetic variants at play may improve our understanding of this gene and cardiac disease pathophysiology.

We performed an in-depth conditional analysis of this locus using information of the entire ECG using a deterministic approach. For this, a stepwise conditional analysis was conducted followed by the full model containing all of the identified SNPs $P < x\ 10^{-8}$; after this the stepwise conditional analysis was redone with as starting point the SNPs that

were also the lead SNP in the full model. This was repeated until all SNPs in the model were the lead associated SNP in the full model.

The first stepwise model including all 1309 variants between 38500000 and 38900000bp on chromosome 3 identified 11 genetic variants. After 3 iterations, a final model was constructed of 10 genetic variants that were independently associated and also the top-associated variants across all ECG association analyses. Please see **Table S11** for more information on the exact configurations that were tested and **Supplementary Data Figure 4** for the conditional regional association plots and impact on ECG morphology for both the univariate and joint (including all 10 variants) model.

Notably, all of the 10 genetic variants were located in a cardiac enhancer (https://pubs.broadinstitute.org/mammals/haploreg). When summarizing the overlap of all variants in the *SCN5A* region, only 21% overlapped with a cardiac enhancer. The fact that we find all (10 out of 10) genetic variants to overlap a cardiac enhancer represents a considerable enrichment (P=1x10$^{-7,}$ hypergeometric test) on the single locus level.

Previously identified genetic variants for Brugada syndrome at rs10428132 and rs11708996 were also found back as independent signals in this analysis. Rs10428132 was the genetic variant with the lowest *P* values overall and was fine-mapped to rs6801957; rs11708996 was in perfect LD with rs41312411, highlighting that this approach is able to capture disease relevant genetic variants. Both variants had a strong effect on the P-wave interval, whereas Brugada is characterized by ST abnormalities, the syndrome is also known to present a concealed abnormal atrial, P-wave, phenotype (Conte et al., 2018). Rs41312411 also has a specific effect on the beginning of the QRS-complex, which seems to affect Q-R duration and/or Q amplitude.

QUANTIFICATION AND STATISTICAL ANALYSIS

Genetics and regression analyses

All of the genetic analyses in UK Biobank that are reported in this manuscript have been adjusted for age, gender, BMI, height and the first 30 principal components (PCAs) to account for population stratification, and genotyping array (Affymetrix UK Biobank Axiom® array or Affymetrix UK BiLEVE Axiom array). The ECG variables were inverse rank normalized prior to association analyses

Genome wide association analyses in UK Biobank were performed using BOLT-LMM v2.3beta2. BOLT-LMM fits a mixed linear model that accounts for the population structure and cryptic relatedness.(Loh et al., 2015) For this, directly genotyped variants were used that passed quality control, and which were extracted from the imputed dataset to ensure 100% call rate, and pruned on linkage disequilibrium (first $r^2 < 0.05$ and a second round of $r^2<0.045$) to obtain roughly 400k variants across the genome. The genomic control lambda's, heritability by BOLT-LMM, LD-Score intercepts and the attenuation ratio statistics are listed in **Table S12**, these suggested no inflation due to non-polygenic signals (Loh et al., 2018). Regression analyses of genetic risk scores and individual genetic variant associations across the ECG-morphologies were performed with sandwich robust standard errors that were clustered by family to account for relatedness in STATA-MP v15. Families were inferred from the kinship matrix, clustering all 3rd degree relatives or higher together (kinship coefficient > 0.0442).

For the GWAS we focused on the ECG morphology unadjusted for R-R interval as it was more powerful compared to the R-R adjusted morphology phenotype; higher heritability, captured the large majority of previously identified genetic variants of the ECG and was easiest to interpret. We performed 500 GWASes, one for each time-point. Using an eigenvalue-based measure(Galwey, 2009), we estimated that testing the 500 ECG data points resulted in effectively 79.9 independent tests and therefore used an alpha of $5x10^{-8}/80 = 6.25x10^{-10}$ to indicate Bonferroni corrected genome wide significance.

Normalized association profiles of variants were plotted as heatmaps where red colors indicate the most strongly associated effect at a time point and the blue color the most strongly associated effect in the opposite direction. To visualize the similarities further on a 2-d plot, we used dimensionality reduction by t-sne on these normalized association profiles for all genome-wide variants, and k-means for clustering. This method only provides a global overview of the primary effects among variants.

Mendelian randomization

Lead variants associated at $P < 5 \times 10^{-8}$ were used as instrumental variables in the MR. Enrichment of low P values ($P < 0.05$) among variants was calculated using a binomial distribution. Mendelian randomization analyses included the inverse-variance-weighted fixed-effects, random-effects meta analyses and weighted median. Heterogeneity was assessed through I^2 index and Cochran's Q. Pleiotropy analyses included the MR-Egger intercept, MR-RAPS and MR-LASSO estimates(Slob and Burgess, 2019). MR pleiotropy residual sum and outlier (MR-PRESSO)(Verbanck et al., 2018) was used as MR analysis to detect and remove pleiotropic variants. MR-PRESSO accounts for pleiotropic effects of the genetic variants comparing the observed distance of all the variants to the regression line (residual sum of squares) to the expected distance under the null hypothesis of no

horizontal pleiotropy. MR-PRESSO is based on the assumption that at least 50% of the variants are valid instruments.

For the Mendelian randomization of dilated cardiomyopathy, variant effects on the ECG morphology at -18 ms were estimated using only individuals without a history of cardiac disease. We performed an additional analysis to explore the impact of pleiotropic signals by conducting a 2-stage Mendelian randomization using a discovery and replication stage. First, we assessed the causal relationship in UK Biobank and identified pleiotropic outliers by MR-PRESSO. Secondly, we replicated this analysis by repeating the MR in an independent cohort of 631 dilated cardiomyopathy patients from MAGNet. In this second stage, we excluded the variants identified by MR-PRESSO in the first stage, hence correcting for pleiotropic effects in an unbiased manner.

References

AlGhatrif, M., and Lindsay, J. (2012). A brief review: history to understand fundamentals of electrocardiography. J. Community Hosp. Intern. Med. Perspect. *2*.

Arking, D.E., Pulit, S.L., Crotti, L., Van Der Harst, P., Munroe, P.B., Koopmann, T.T., Sotoodehnia, N., Rossin, E.J., Morley, M., Wang, X., et al. (2014). Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. Nat. Genet. *46*, 826–836.

Bastiaenen, R., Nolte, I.M., Munroe, P.B., Riese, H., Nelson, C., O'Connor, H., Gang, Y., Warren, H.R., Cabrera, C., Reinhard, W., et al. (2018). The narrow-sense and common single nucleotide polymorphism heritability of early repolarization. Int. J. Cardiol.

van den Berg, M.E., Warren, H.R., Cabrera, C.P., Verweij, N., Mifsud, B., Haessler, J., Bihlmeyer, N.A., Fu, Y.P., Weiss, S., Lin, H.J., et al. (2017). Discovery of novel heart rate-associated loci using the Exome Chip. Hum. Mol. Genet. *26*, 2346–2363.

Bihlmeyer, N.A., Brody, J.A., Smith, A.V., Warren, H.R., Lin, H., Isaacs, A., Liu, C.-T., Marten, J., Radmanesh, F., Hall, L.M., et al. (2018). ExomeChip-Wide Analysis of 95 626 Individuals Identifies 10 Novel Loci Associated With QT and JT Intervals. Circ Genom Precis Med *11*, e001758.

Bozkurt, B., Colvin, M., Cook, J., Cooper, L.T., Deswal, A., Fonarow, G.C., Francis, G.S., Lenihan, D., Lewis, E.F., McNamara, D.M., et al. (2016). Current Diagnostic and Treatment Strategies for Specific Dilated Cardiomyopathies: A Scientific Statement from the American Heart Association. Circulation *134*, e579–e646.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

Christophersen, I.E., Magnani, J.W., Yin, X., Barnard, J., Weng, L.-C., Arking, D.E., Niemeijer, M.N., Lubitz, S.A., Avery, C.L., Duan, Q., et al. (2017). Fifteen Genetic Loci Associated With the Electrocardiographic P Wave. Circ. Cardiovasc. Genet. *10*.

Conte, G., Caputo, M.L., Volders, P.G.A., Luca, A., Mainardi, L., Schotten, U., Corino, V.D.A., Regoli, F., Zeemering, S., Zink, M., et al. (2018). Concealed abnormal atrial phenotype in patients with Brugada syndrome and no history of atrial fibrillation. Int. J. Cardiol. *253*, 66–70.

Das, M.K., and Zipes, D.P. (2012). Electrocardiography of arrhythmias : a comprehensive review : a companion to Cardiac electrophysiology : from cell to bedside (Elsevier/Saunders).

Eppinga, R.N., Hagemeijer, Y., Burgess, S., Hinds, D.A., Stefansson, K., Gudbjartsson, D.F., Van Veldhuisen, D.J., Munroe, P.B., Verweij, N., and Van Der Harst, P. (2016). Identification of genomic loci associated with resting heart rate and shared genetic predictors with all-cause mortality. Nat. Genet. *48*, 1557–1563.

Galwey, N.W. (2009). A new measure of the effective number of tests, a practical tool

for comparing families of non-independent significance tests. Genet. Epidemiol. *33*, 559–568.

H, C., D, C., and F, M. (2017). Risk Stratification in Arrhythmogenic Right Ventricular Cardiomyopathy. Circulation *136*.

Haïssaguerre, M., Derval, N., Sacher, F., Jesel, L., Deisenhofer, I., de Roy, L., Pasquié, J.-L., Nogami, A., Babuty, D., Yli-Mayry, S., et al. (2008). Sudden cardiac arrest associated with early repolarization. N. Engl. J. Med. *358*, 2016–2023.

van der Harst, P., van Setten, J., Verweij, N., Vogler, G., Franke, L., Maurano, M.T., Wang, X., Mateo Leach, I., Eijgelsheim, M., Sotoodehnia, N., et al. (2016). 52 Genetic Loci Influencing Myocardial Mass. J. Am. Coll. Cardiol. *68*, 1435–1448.

Hu, R., Morley, M.P., Brandimarto, J., Tucker, N.R., Parsons, V.A., Zhao, S.D., Meder, B., Katus, H.A., Rühle, F., Stoll, M., et al. (2018). Genetic Reduction in Left Ventricular Protein Kinase C-α and Adverse Ventricular Remodeling in Human Subjects. Circ. Genomic Precis. Med. *11*, e001901.

Japp, A.G., Gulati, A., Cook, S.A., Cowie, M.R., and Prasad, S.K. (2016). The Diagnosis and Evaluation of Dilated Cardiomyopathy. J. Am. Coll. Cardiol. *67*, 2996–3010.

Liao, W.-C., Juo, L.-Y., Shih, Y.-L., Chen, Y.-H., and Yan, Y.-T. (2017). HSPB7 prevents cardiac conduction system defect through maintaining intercalated disc integrity. PLOS Genet. *13*, e1006984.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290.

Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A.P., and Price, A.L. (2018). Mixed-model association for biobank-scale datasets. Nat. Genet. *50*, 906–908.

Macfarlane, P.W., Antzelevitch, C., Haissaguerre, M., Huikuri, H. V, Potse, M., Rosso, R., Sacher, F., Tikkanen, J.T., Wellens, H., and Yan, G.-X. (2015). The Early Repolarization Pattern: A Consensus Paper. J. Am. Coll. Cardiol. *66*, 470–477.

Marston, S. (2017). Obscurin variants and inherited cardiomyopathies. Biophys. Rev. *9*, 239–243.

Meder, B., Rühle, F., Weis, T., Homuth, G., Keller, A., Franke, J., Peil, B., Lorenzo Bermejo, J., Frese, K., Huge, A., et al. (2014). A genome-wide association study identifies 6p21 as novel risk locus for dilated cardiomyopathy. Eur. Heart J. *35*, 1069–1077.

Nielsen, J.B., Thorolfsdottir, R.B., Fritsche, L.G., Zhou, W., Skov, M.W., Graham, S.E., Herron, T.J., McCarthy, S., Schmidt, E.M., Sveinbjornsson, G., et al. (2018). Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. Nat. Genet. *50*, 1234–1239.

Pers, T.H., Karjalainen, J.M., Chan, Y., Westra, H.-J., Wood, A.R., Yang, J., Lui, J.C., Vedantam, S., Gustafsson, S., Esko, T., et al. (2015). Biological interpretation of genome-wide association studies using predicted gene functions. Nat. Commun. *6*, 5890.

Prins, B.P., Mead, T.J., Brody, J.A., Sveinbjornsson, G., Ntalla, I., Bihlmeyer, N.A., van

den Berg, M., Bork-Jensen, J., Cappellani, S., Van Duijvenboden, S., et al. (2018). Exome-chip meta-analysis identifies novel loci associated with cardiac conduction, including ADAMTS6. Genome Biol. *19*, 87.

Roselli, C., Chaffin, M.D., Weng, L.-C., Aeschbacher, S., Ahlberg, G., Albert, C.M., Almgren, P., Alonso, A., Anderson, C.D., Aragam, K.G., et al. (2018). Multi-ethnic genome-wide association study for atrial fibrillation. Nat. Genet.

Rosenbaum, A.N., Agre, K.E., and Pereira, N.L. (2019). Genetics of dilated cardiomyopathy: practical implications for heart failure management. Nat. Rev. Cardiol.

Sadrieh, A., Domanski, L., Pitt-Francis, J., Mann, S.A., Hodkinson, E.C., Ng, C.-A., Perry, M.D., Taylor, J.A., Gavaghan, D., Subbiah, R.N., et al. (2014). Multiscale cardiac modelling reveals the origins of notched T waves in long QT syndrome type 2. Nat. Commun. *5*, 5069.

van Setten, J., Brody, J.A., Jamshidi, Y., Swenson, B.R., Butler, A.M., Campbell, H., Del Greco, F.M., Evans, D.S., Gibson, Q., Gudbjartsson, D.F., et al. (2018). PR interval genome-wide association meta-analysis identifies 50 loci associated with atrial and atrioventricular electrical activity. Nat. Commun. *9*, 2904.

Simson, M.B. (1981). Use of signals in the terminal QRS complex to identify patients with ventricular tachycardia after myocardial infarction. Circulation *64*, 235–242.

Slob, E.A.W., and Burgess, S. (2019). A Comparison Of Robust Mendelian Randomization Methods Using Summary Data. BioRxiv 577940.

Stacey, D., Fauman, E.B., Ziemek, D., Sun, B.B., Harshfield, E.L., Wood, A.M., Butterworth, A.S., Suhre, K., and Paul, D.S. (2019). ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. Nucleic Acids Res. *47*, e3–e3.

Stark, K., Esslinger, U.B., Reinhard, W., Petrov, G., Winkler, T., Komajda, M., Isnard, R., Charron, P., Villard, E., Cambien, F., et al. (2010). Genetic Association Study Identifies HSPB7 as a Risk Gene for Idiopathic Dilated Cardiomyopathy. PLoS Genet. *6*, e1001167.

Teijeiro, T., Felix, P., Presedo, J., and Castro, D. (2018). Heartbeat Classification Using Abstract Features From the Abductive Interpretation of the ECG. IEEE J. Biomed. Heal. Informatics *22*, 409–420.

Teumer, A., Trenkwalder, T., Kessler, T., Jamshidi, Y., van den Berg, M.E., Kaess, B., Nelson, C.P., Bastiaenen, R., de Bortoli, M., Rossini, A., et al. (2019). KCND3 potassium channel gene variant confers susceptibility to electrocardiographic early repolarization pattern. JCI Insight *4*.

Tikkanen, J.T., Junttila, M.J., Anttonen, O., Aro, A.L., Luttinen, S., Kerola, T., Sager, S.J., Rissanen, H.A., Myerburg, R.J., Reunanen, A., et al. (2011). Early repolarization: electrocardiographic phenotypes associated with favorable long-term outcome. Circulation *123*, 2666–2673.

van de Vegte, Y.J., van der Harst, P., and Verweij, N. (2018). Heart rate recovery 10 seconds after cessation of exercise predicts death. J. Am. Heart Assoc. *7*.

Verbanck, M., Chen, C.-Y., Neale, B., and Do, R. (2018). Detection of widespread

horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat. Genet. *50*, 693–698.

Verweij, N., Leach, I.M., Van Den Boogaard, M., Van Veldhuisen, D.J., Christoffels, V.M., Hillege, H.L., Van Gilst, W.H., Barnett, P., De Boer, R.A., and Van Der Harst, P. (2014). Genetic Determinants of P Wave Duration and PR Segment. Circ. Cardiovasc. Genet. *7*, 475–481.

Verweij, N., Leach, I.M., Isaacs, A., Arking, D.E., Bis, J.C., Pers, T.H., Van Den Berg, M.E., Lyytikäinen, L.P., Barnett, P., Wang, X., et al. (2016). Twenty-eight genetic loci associated with ST-T-wave amplitudes of the electrocardiogram. Hum. Mol. Genet. *25*, 2093–2103.

Verweij, N., Eppinga, R.N., Hagemeijer, Y., and Van Der Harst, P. (2017). Identification of 15 novel risk loci for coronary artery disease and genetic risk of recurrent events, atrial fibrillation and heart failure. Sci. Rep. *7*.

Verweij, N., Van De Vegte, Y.J., and Van Der Harst, P. (2018). Genetic study links components of the autonomous nervous system to heart-rate profile during exercise. Nat. Commun. *9*.

Villard, E., Perret, C., Gary, F., Proust, C., Dilanian, G., Hengstenberg, C., Ruppert, V., Arbustini, E., Wichter, T., Germain, M., et al. (2011). A genome-wide association study identifies two loci associated with heart failure due to dilated cardiomyopathy. Eur. Heart J. *32*, 1065–1076.

Wang, X., and Goldstein, D.B. (2020). Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. Am. J. Hum. Genet. *106*, 215–233.

Figure legends

Figure 1. Overview of the study design. A schematic overview of the phenotyping approach and downstream *in-silico* annotations and analyses.

Figure 2 Polygenic risk scores of ECG indices are associated with the trait expected segments of the ECG. For each of the classical ECG traits we calculated polygenic scores and associated them with each of the datapoint that makeup the ECG morphology phenotype. The left y-axis depicts the micro voltage scale, the y-axis on the right indicates the signed -log10(*P*-values). The x-axis is time in milliseconds (ms), or percentage (%) from the RR in the case of the RR adjusted ECG morphology phenotype. The blue lines are the average ECG amplitude of the full cohort and the red lines the P-value for association with each datapoint of the ECG morphology phenotype (n=500 timepoints) on a $\log_{10}$ scale, signed to show direction of association. The dashed vertical black lines mark point of strongest negative and positive association. Additional plots for the R-R adjusted ECG phenotype and sensitivity analyses can be found in the appendix.

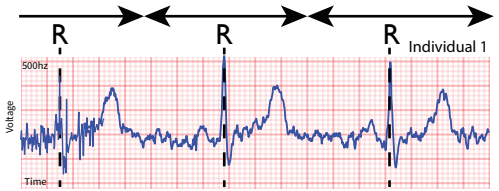
Figure 3. Genetic variants display unique morphological signatures and the heritability is highest at ECG segments showing greatest electrical activity. a, A heatmap of the genetic ECG signatures that were previously found to be associated with AF. We normalized the genetic ECG signatures of genetic variants in order to compare the effects across loci. Effects were orientated to the most positively associated allele across all time points and colored in red on the heatmap; a blue color indicates a negative effect while yellow indicates no effect. On the left side is the unadjusted ECG phenotype, on the right side the R-R adjusted ECG phenotype. **b.** excerpts of genetic ECG signatures for previously reported genetic variants, plotted in the same way as the polygenic risk scores in Figure 2. The red line indicates the signed -$\log_{10}$(*P*-value) for association across the heartbeat. And the blue line the average ECG amplitudes in the population **c.** The SNP-heritability's (h^2_g) are plotted in red for the unadjusted and the R-R interval adjusted ECG morphology phenotype. The SNP heritability was highest for the ECG morphology phenotype unadjusted for the R-R interval (left plots). The maximum observed heritability was 0.29 (se=0.01) at the ST-wave segment. Heritability estimates were high for ECG segments that have high electrical activity, consistent with molecular mechanism of cardiac conduction. The plots on the bottom shows the pairwise Pearson correlation matrix on the phenotype level between each ECG datapoint across the ECG morphology phenotypes. Red indicates a high positive correlation, blue indicates a high negative correlation, yellow indicates no correlation; both the x and y axis depict the ECG in time and match the average ECG amplitudes plotted on the top.

Figure 4. Unbiased clustering maps main effects of genetic variants on the electrocardiogram. a, Manhattan plot of the ECG morphology phenotype (smallest *P* value across all traits is shown), variants in red indicate those passing $P<5x10^{-8}$, the top

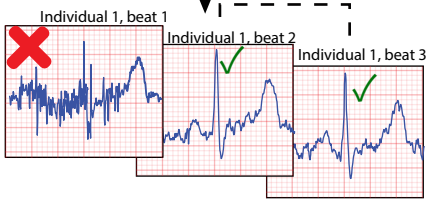
loci have been annotated with their nearby genes. **b**, The t-SNE plot was derived from a large matrix that contained all genetic ECG signatures (signed -log(P) values) for all lead genetic variants that were identified in the GWAS. An interactive t-SNE plot is available on www.ecgenetics.org. **c**, Exemplar ECG signatures were annotated with the most likely candidate causal gene, illustrating how each cluster can be linked back to their primary ECG effect.

Figure 5. Mendelian randomization of dilated cardiomyopathy and early repolarization. **a**, the left plot illustrates the ECG signature of two loci previously identified for dilated cardiomyopathy, but never before for the ECG. The right plot illustrates an ECG-wide genetic signature for dilated cardiomyopathy: for this we carried out a fixed effects Mendelian randomization (MR) for each time point, including all lead variants $P<5x10^{-8}$ per time point of the ECG as instrumental variables and dilated cardiomyopathy as outcome variable. The red lines indicate the P-value for association with each datapoint of the ECG morphology phenotype (n=500 timepoints) on a $\log_{10}$ scale, signed to show direction of association. **b**, The genetic ECG signature of *KCND3* shows an association peak that is consistent with the early repolarization criterion at 44 ms after the R peak. **c**, Forest plot of the Mendelian randomization estimates of the ECG at -18ms on dilated cardiomyopathy for each variant individually. Variants in red are Bonferroni significant ($P<0.001$), variants in bold are $P < 0.05$ for dilated cardiomyopathy. **d**, Mendelian randomization estimates of the ECG at -18ms on dilated cardiomyopathy **e.** Mendelian randomization estimates of the ECG at -18ms on dilated cardiomyopathy in the discovery (UK Biobank) and replication (MAGNet) cohort controlling for pleiotropic variants identified by MR-PRESSO. **f**. Forest plot of the Mendelian randomization estimates of the ECG at 44ms on early repolarization for each variant individually. Variants in red are Bonferroni significant ($P<0.001$), variants in bold are $P < 0.05$ for early repolarization.

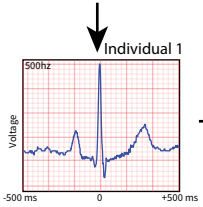
Phenotyping of 77,190 ECG's of UK Biobank



Detect R waves and overlap beats by slicing it up in windows

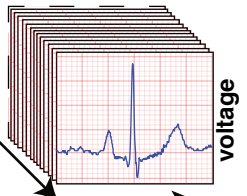


Individual level QC, remove noise and abnormal beats.



Averaged individual beats

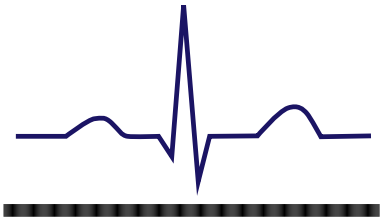
Population level QC, remove abnormal electrocardiograms and standardize



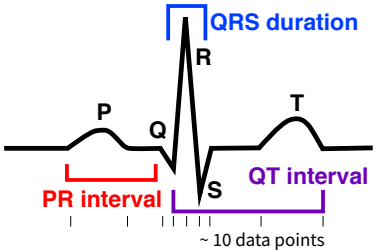
data on >60,000 individuals

500 traits representing the 'ECG morphology phenotype'

Current approach: study the complete ECG morphology phenotype



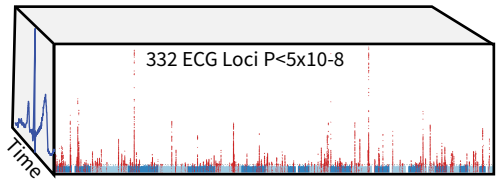
Classically defined ECG sections



In-silico analyses

Perform association analysis between the ECG morphology phenotype represented by 500 traits and polygenic risk scores of classically defined ECG traits.

Perform ECG-wide heritability analysis and genome wide association analysis across the 'ECG morphology phenotype' (500 GWASes)



Visualizing genetic ECG signatures by plotting the signed P-values along the ECG's x-axis

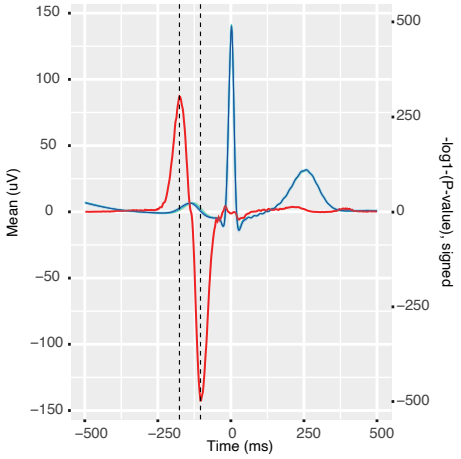
t-SNE and clustering analysis of 332 genetic ECG signatures

Biological insights: candidate causal gene identification and differential pathway analysis across the ECG

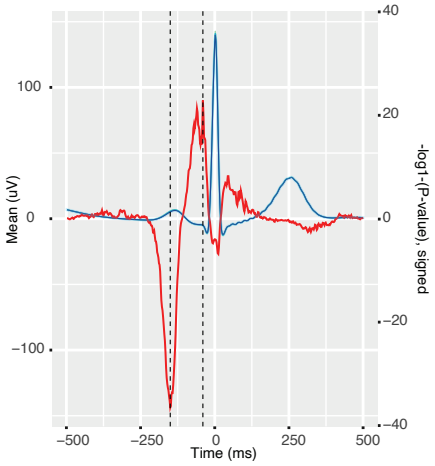
Disease insights: Dilated cardiomyopathy & Early Repolarization
- **Identification of ECG endophentypes of disease**
- **Identification of new genetic variants associated with disease**

ECG-wide conditional analysis of the SCN5A locus

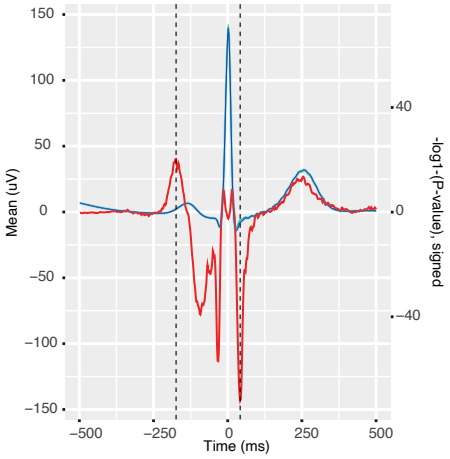
PR interval



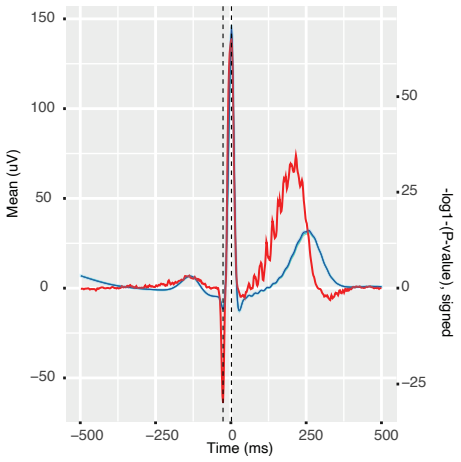
Atrial Fibrillation



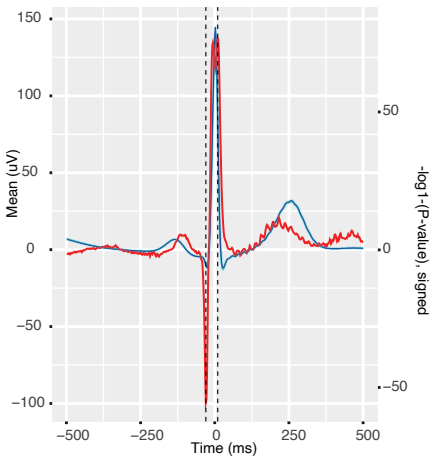
QRS duration



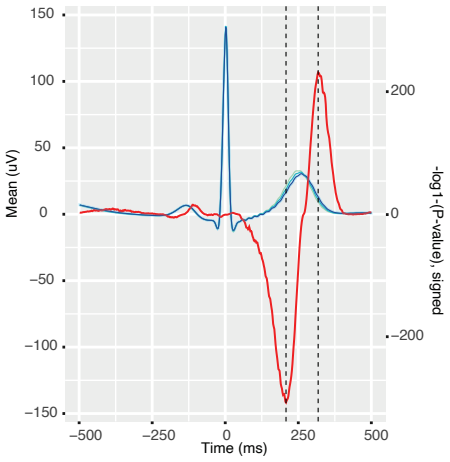
12 Lead sum area



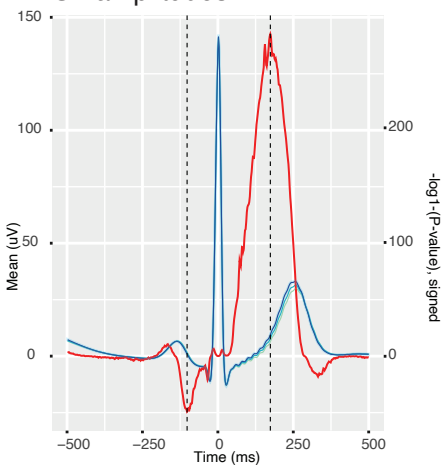
Sokolow-Lyon area



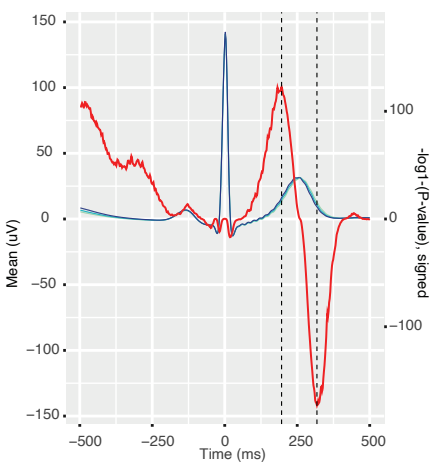
QT interval



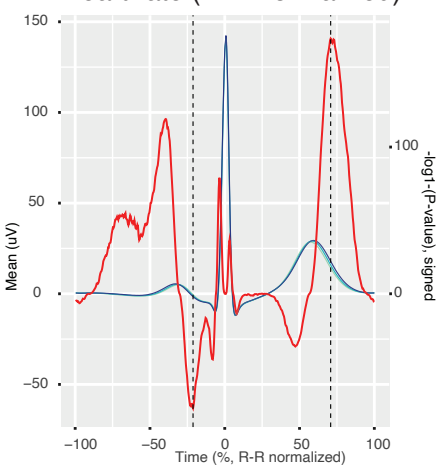
ST amplitudes

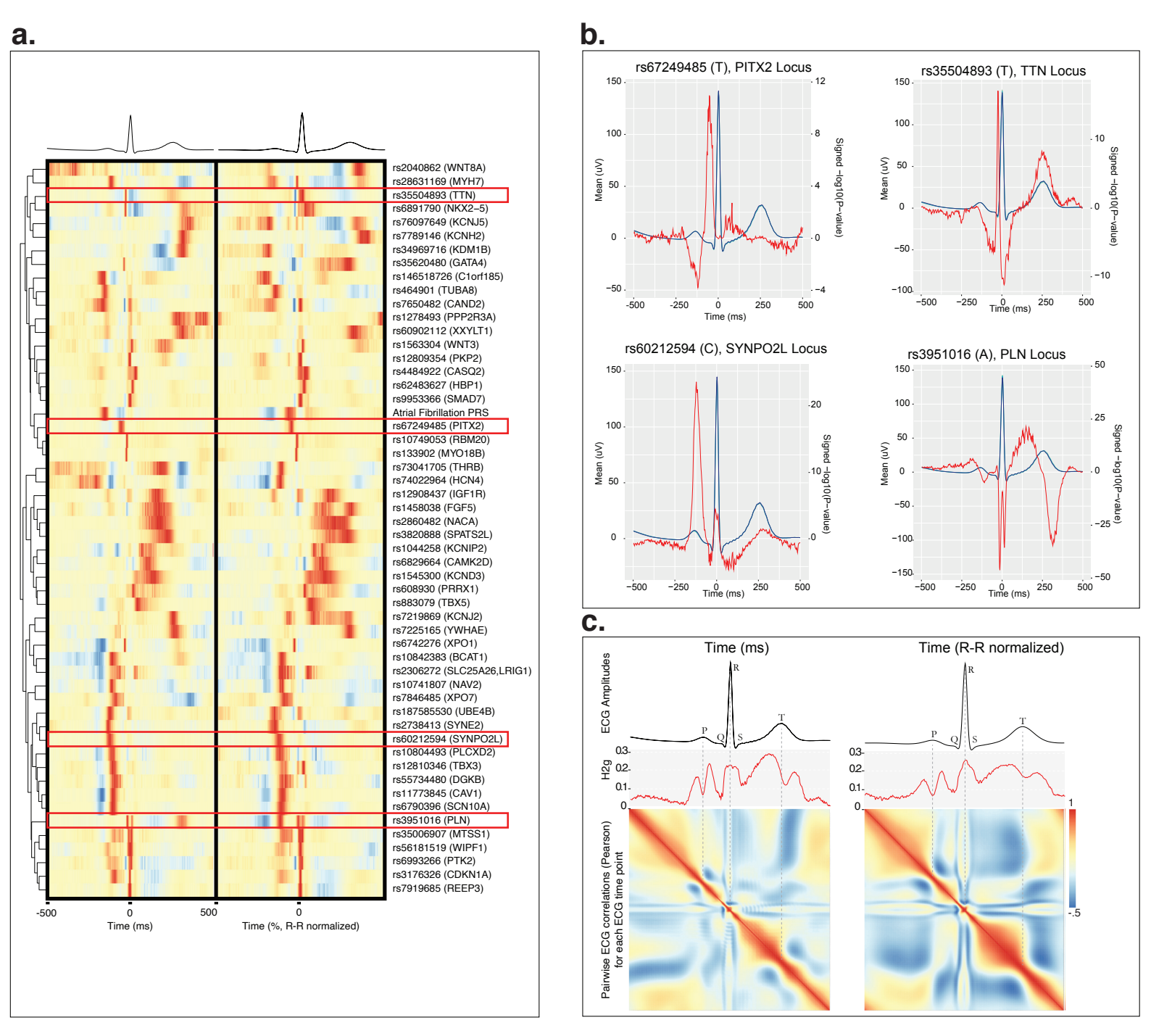


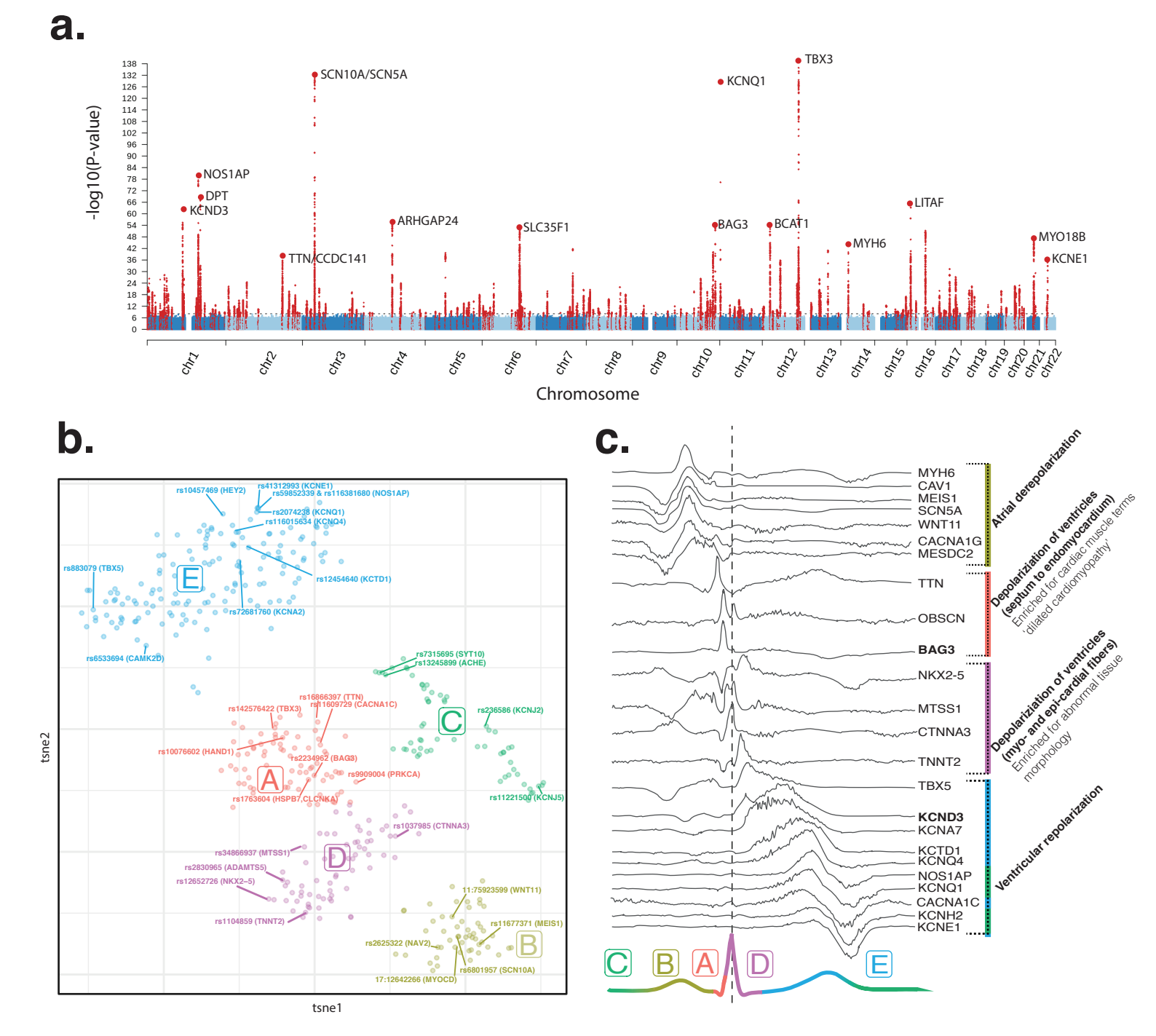
Heart rate

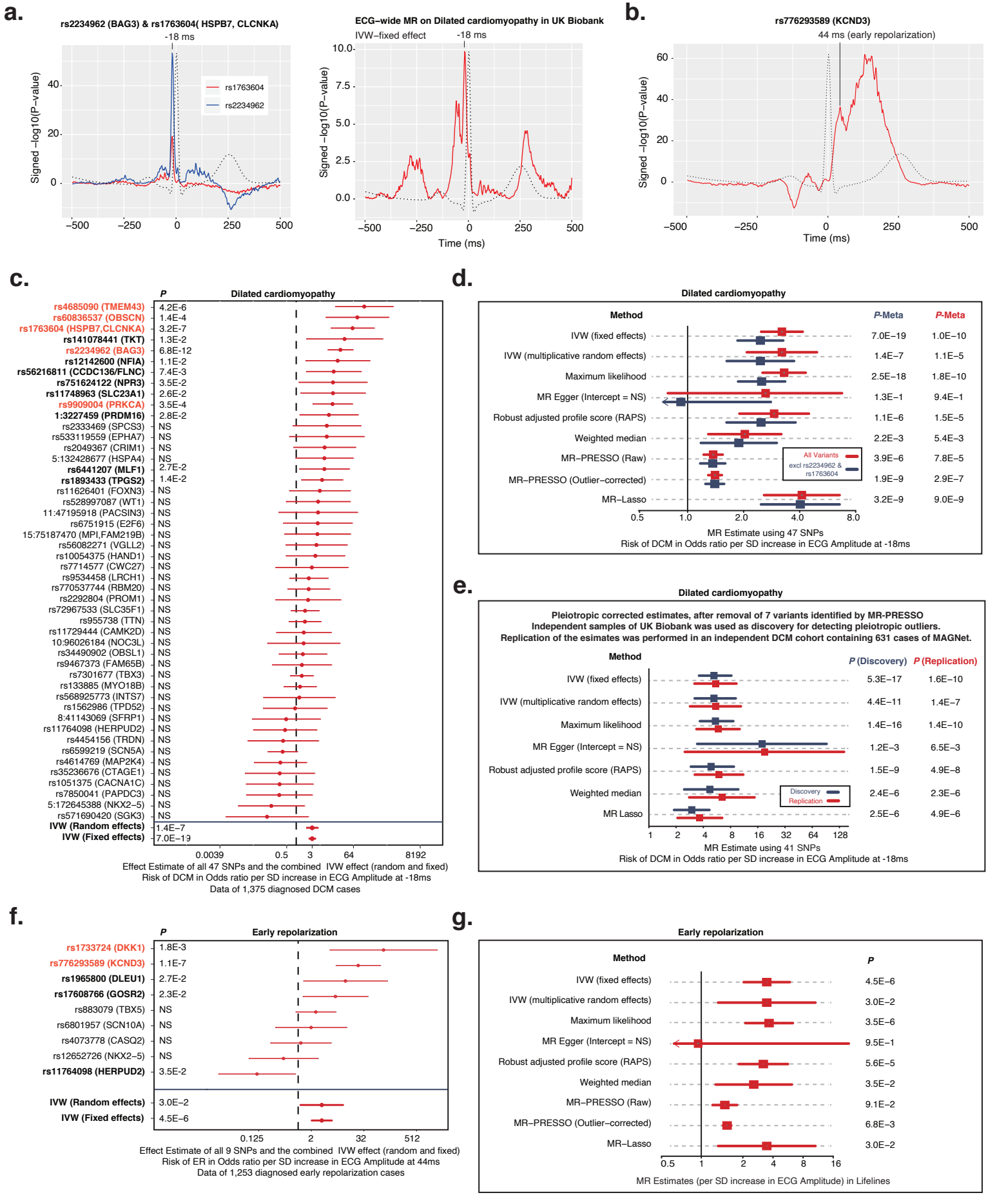


Heart rate (R-R normalized)




a.
rs2040862 (WNT8A)
rs28631169 (MYH7)
rs35504893 (TTN)
rs6891790 (NKX2–5)
rs76097649 (KCNJ5)
rs7789146 (KCNH2)
rs34969716 (KDM1B)
rs35620480 (GATA4)
rs146518726 (C1orf185)
rs464901 (TUBA8)
rs7650482 (CAND2)
rs1278493 (PPP2R3A)
rs60902112 (XXYLT1)
rs1563304 (WNT3)
rs12809354 (PKP2)
rs4484922 (CASQ2)
rs62483627 (HBP1)
rs9953366 (SMAD7)
Atrial Fibrillation PRS
rs67249485 (PITX2)
rs10749053 (RBM20)
rs133902 (MYO18B)
rs73041705 (THRB)
rs74022964 (HCN4)
rs12908437 (IGF1R)
rs1458038 (FGF5)
rs2860482 (NACA)
rs3820888 (SPATS2L)
rs1044258 (KCNIP2)
rs6829664 (CAMK2D)
rs1545300 (KCND3)
rs608930 (PRRX1)
rs883079 (TBX5)
rs7219869 (KCNJ2)
rs7225165 (YWHAE)
rs6742276 (XPO1)
rs10842383 (BCAT1)
rs2306272 (SLC25A26,LRIG1)
rs10741807 (NAV2)
rs7846485 (XPO7)
rs187585530 (UBE4B)
rs2738413 (SYNE2)
rs60212594 (SYNPO2L)
rs10804493 (PLCXD2)
rs12810346 (TBX3)
rs55734480 (DGKB)
rs11773845 (CAV1)
rs6790396 (SCN10A)
rs3951016 (PLN)
rs35006907 (MTSS1)
rs56181519 (WIPF1)
rs6993266 (PTK2)
rs3176326 (CDKN1A)
rs7919685 (REEP3)
-500
0
500
Time (ms)
0
Time (%, R-R normalized)
b.
rs67249485 (T), PITX2 Locus
rs35504893 (T), TTN Locus
rs60212594 (C), SYNPO2L Locus
rs3951016 (A), PLN Locus
Mean (uV)
Signed −log10(P−value)
Time (ms)
c.
Time (ms)
Time (R-R normalized)
ECG Amplitudes
H2g
Pairwise ECG correlations (Pearson) for each ECG time point
R
P
Q
S
T


a.
-log10(P-value)
SCN10A/SCN5A
KCNQ1
TBX3
NOS1AP
DPT
KCND3
ARHGAP24
SLC35F1
BAG3
BCAT1
LITAF
MYH6
MYO18B
KCNE1
TTN/CCDC141
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
Chromosome
b.
rs10457469 (HEY2)
rs41312993 (KCNE1)
rs59852339 & rs116381680 (NOS1AP)
rs2074238 (KCNQ1)
rs116015634 (KCNQ4)
rs883079 (TBX5)
rs12454640 (KCTD1)
rs72681760 (KCNA2)
rs6533694 (CAMK2D)
rs7315695 (SYT10)
rs13245899 (ACHE)
rs16866397 (TTN)
rs11609729 (CACNA1C)
rs142576422 (TBX3)
rs236586 (KCNJ2)
rs10076602 (HAND1)
rs2234962 (BAG3)
rs9909004 (PRKCA)
rs1763604 (HSPB7,CLCNKA)
rs11221500 (KCNJ5)
rs1037985 (CTNNA3)
rs34866937 (MTSS1)
rs2830965 (ADAMTS5)
rs12652726 (NKX2-5)
11:75923599 (WNT11)
rs1104859 (TNNT2)
rs11677371 (MEIS1)
rs2625322 (NAV2)
rs6801957 (SCN10A)
17:12642266 (MYOCD)
A
B
C
D
E
tsne1
tsne2
c.
MYH6
CAV1
MEIS1
SCN5A
WNT11
CACNA1G
MESDC2
TTN
OBSCN
BAG3
NKX2-5
MTSS1
CTNNA3
TNNT2
TBX5
KCND3
KCNA7
KCTD1
KCNQ4
NOS1AP
KCNQ1
CACNA1C
KCNH2
KCNE1
Atrial derepolarization
Depolarization of ventricles (septum to endomyocardium)
Enriched for cardiac muscle terms
dilated cardiomyopathy
Depolarization of ventricles (myo- and epi-cardial fibers)
Enriched for abnormal tissue morphology
Ventricular repolarization
C
B
A
D
E


a.
rs2234962 (BAG3) & rs1763604(HSPB7, CLCNKA)
-18 ms
Signed −log10(P−value)
rs1763604
rs2234962
ECG−wide MR on Dilated cardiomyopathy in UK Biobank
IVW−fixed effect
-18 ms
Time (ms)
b.
rs776293589 (KCND3)
44 ms (early repolarization)
c.
P
Dilated cardiomyopathy
rs4685090 (TMEM43) 4.2E-6
rs60836537 (OBSCN) 1.4E-4
rs1763604 (HSPB7,CLCNKA) 3.2E-7
rs141078441 (TKT) 1.3E-2
rs2234962 (BAG3) 6.8E-12
rs12142600 (NFIA) 1.1E-2
rs56216811 (CCDC136/FLNC) 7.4E-3
rs751624122 (NPR3) 3.5E-2
rs11748963 (SLC23A1) 2.6E-2
rs9909004 (PRKCA) 3.5E-4
1:3227459 (PRDM16) 2.8E-2
rs2333469 (SPCS3) NS
rs533119559 (EPHA7) NS
rs2049367 (CRIM1) NS
5:132428677 (HSPA4) NS
rs6441207 (MLF1) 2.7E-2
rs1893433 (TPGS2) 1.4E-2
rs11626401 (FOXN3) NS
rs528997087 (WT1) NS
11:47195918 (PACSIN3) NS
rs6751915 (E2F6) NS
15:75187470 (MPI,FAM219B) NS
rs56082271 (VGLL2) NS
rs10054375 (HAND1) NS
rs7714577 (CWC27) NS
rs9534458 (LRCH1) NS
rs770537744 (RBM20) NS
rs2292804 (PROM1) NS
rs72967533 (SLC35F1) NS
rs955738 (TTN) NS
rs11729444 (CAMK2D) NS
10:96026184 (NOC3L) NS
rs34490902 (OBSL1) NS
rs9467373 (FAM65B) NS
rs7301677 (TBX3) NS
rs133885 (MYO18B) NS
rs568925773 (INTS7) NS
rs1562986 (TPD52) NS
8:41143069 (SFRP1) NS
rs11764098 (HERPUD2) NS
rs4454156 (TRDN) NS
rs6599219 (SCN5A) NS
rs4614769 (MAP2K4) NS
rs35236676 (CTAGE1) NS
rs1051375 (CACNA1C) NS
rs7850041 (PAPDC3) NS
5:172645388 (NKX2−5) NS
rs571690420 (SGK3) NS
IVW (Random effects) 1.4E−7
IVW (Fixed effects) 7.0E−19
0.0039
0.5
3
64
8192
Effect Estimate of all 47 SNPs and the combined IVW effect (random and fixed)
Risk of DCM in Odds ratio per SD increase in ECG Amplitude at -18ms
Data of 1,375 diagnosed DCM cases
d.
Dilated cardiomyopathy
Method
P-Meta
P-Meta
IVW (fixed effects) 7.0E−19 1.0E−10
IVW (multiplicative random effects) 1.4E−7 1.1E−5
Maximum likelihood 2.5E−18 1.8E−10
MR Egger (Intercept = NS) 1.3E−1 9.4E−1
Robust adjusted profile score (RAPS) 1.1E−6 1.5E−5
Weighted median 2.2E−3 5.4E−3
MR−PRESSO (Raw) 3.9E−6 7.8E−5
MR−PRESSO (Outlier−corrected) 1.9E−9 2.9E−7
MR−Lasso 3.2E−9 9.0E−9
All Variants
excl rs2234962 & rs1763604
0.5
1.0
2.0
4.0
8.0
MR Estimate using 47 SNPs
Risk of DCM in Odds ratio per SD increase in ECG Amplitude at -18ms
e.
Dilated cardiomyopathy
Pleiotropic corrected estimates, after removal of 7 variants identified by MR-PRESSO
Independent samples of UK Biobank was used as discovery for detecting pleiotropic outliers.
Replication of the esimates was performed in an independent DCM cohort containing 631 cases of MAGNet.
Method
P (Discovery)
P (Replication)
IVW (fixed effects) 5.3E−17 1.6E−10
IVW (multiplicative random effects) 4.4E−11 1.4E−7
Maximum likelihood 1.4E−16 1.4E−10
MR Egger (Intercept = NS) 1.2E−3 6.5E−3
Robust adjusted profile score (RAPS) 1.5E−9 4.9E−8
Weighted median 2.4E−6 2.3E−6
MR Lasso 2.5E−6 4.9E−6
Discovery
Replication
1
2
4
8
16
32
64
128
MR Estimate using 41 SNPs
Risk of DCM in Odds ratio per SD increase in ECG Amplitude at -18ms
f.
P
Early repolarization
rs1733724 (DKK1) 1.8E-3
rs776293589 (KCND3) 1.1E-7
rs1965800 (DLEU1) 2.7E-2
rs17608766 (GOSR2) 2.3E-2
rs883079 (TBX5) NS
rs6801957 (SCN10A) NS
rs4073778 (CASQ2) NS
rs12652726 (NKX2−5) NS
rs11764098 (HERPUD2) 3.5E-2
IVW (Random effects) 3.0E−2
IVW (Fixed effects) 4.5E−6
0.125
2
32
512
Effect Estimate of all 9 SNPs and the combined IVW effect (random and fixed)
Risk of ER in Odds ratio per SD increase in ECG Amplitude at 44ms
Data of 1,253 diagnosed early repolarization cases
g.
Early repolarization
Method
P
IVW (fixed effects) 4.5E−6
IVW (multiplicative random effects) 3.0E−2
Maximum likelihood 3.5E−6
MR Egger (Intercept = NS) 9.5E−1
Robust adjusted profile score (RAPS) 5.6E−5
Weighted median 3.5E−2
MR−PRESSO (Raw) 9.1E−2
MR−PRESSO (Outlier−corrected) 6.8E−3
MR−Lasso 3.0E−2
0.5
1
2
4
8
16
MR Estimates (per SD increase in ECG Amplitude) in Lifelines