Deep learning-derived cardiovascular age shares a genetic basis with other cardiac phenotypes

Julian Libiseller-Egger
London School of Hygiene and Tropical Medicine
Jody Phelan
London School of Hygiene and Tropical Medicine https://orcid.org/0000-0001-8323-7019
Zachi Attia
Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, USA. https://orcid.org/0000-0002-9706-7900
Ernest Diez Benavente
University Medical Center Utrecht
Susana Campino
London School of Hygiene and Tropical Medicine https://orcid.org/0000-0003-1403-6138
Paul Friedman
Mayo Clinic https://orcid.org/0000-0001-5052-2948
Francisco Lopez-Jimenez
Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, USA.
David Leon
London School of Hygiene and Tropical Medicine https://orcid.org/0000-0001-9747-1762
Taane Clark (taane.clark@lshtm.ac.uk)
London School of Hygiene & Tropical Medicine https://orcid.org/0000-0001-8985-9265

Article

Keywords:

Posted Date: August 29th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1735746/v1

Deep learning-derived cardiovascular age shares a genetic basis with other cardiac phenotypes

Julian Libiseller-Egger[1], Jody E. Phelan[1], Zachi I. Attia[2], Ernest Diez Benavente[1,3], Susana Campino[1], Paul A. Friedman[2], Francisco Lopez-Jimenez[2], David A. Leon[4,5], Taane G. Clark[1,4,*]

August 16, 2022

1 Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, UK
2 Department of Cardiovascular Medicine, Mayo Clinic College of Medicine, Rochester, MN, USA
3 Laboratory of Experimental Cardiology, University Medical Center Utrecht, Utrecht, Netherlands
4 Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK
5 Department of Community Medicine, UiT the Arctic University of Norway, Tromsø, Norway
***** Corresponding author

Abstract

Artificial intelligence (AI)-based approaches can now use electrocardiograms (ECGs) to provide expert-level performance in detecting heart abnormalities and diagnosing disease. Additionally, patient age predicted from ECGs by AI models has shown great potential as a biomarker for cardiovascular age as recent work has found differences from chronological age ("delta age") to be associated with mortality and co-morbidities. However, the genetic underpinning of delta age is unknown, but crucial for understanding underlying individual risk. By performing a genome-wide association study using UK Biobank data (n=34,432), we identified eight loci associated with delta age ($p \leq 5 \times 10^{-8}$), including genes linked to cardiovascular disease (CVD) (e.g. *SCN5A*) and (heart) muscle development (e.g. *TTN*). Our results indicate that the genetic basis of cardiovascular ageing is predominantly determined by genes directly involved with the cardiovascular system rather than those connected to more general mechanisms of ageing. Our insights inform the epidemiology of CVD, with implications for preventative and precision medicine.

Introduction

For decades it has been known that a person's electrocardiogram (ECG) changes with age [1, 2]. Therefore, in light of its non-invasiveness, ease of obtainment, and consequential ubiquity, there is great potential in using the 12-lead ECG as a biomarker for physiological changes caused by ageing [3]. These changes occur gradually and at a rate that is different between individuals, leading to substantial variation in the risk of chronic disease and mortality in older populations. In order to understand the sources of this variation, several indicators for "biological age" have been investigated, including changes in telomere length [4], the epigenome [5], blood-derived biomarkers [6], and the transcriptome [7]. Crucially, these markers have been shown to be only weakly correlated with each other [8], suggesting that they do not describe the same underlying physiological processes but rather different aspects of ageing [9]. Since cardiovascular disease (CVD) is a major source of mortality and morbidity, with drastically increasing prevalence in older age [10], an ECG-derived metric for cardiovascular age would represent a valuable addition to other "ageing" metrics, with both preventative and personalised medicine benefits.

Initial studies trying to link chronological age to the ECG signal mostly focused on human-defined ECG features, such as the QRS duration or the length of the PR interval [11]. However, the extraction of these features is not devoid of error [12] and relies on a limited number of features – as opposed to the complete signal – inevitably ignoring an unknown fraction of potentially relevant information.

Recent developments in deep learning allowed researchers to address this limitation by adapting modern convolutional artificial neural network architectures to predict patients' ages from their ECGs [13, 14]. These models can be trained "end-to-end" on the raw ECG traces from which they learn to extract (and combine in a non-linear manner) the features most suitable for a prediction task. Thus, the impact of human bias is minimised and predictive power improved as all the information in the signal is taken into account. In fact, several studies have shown that deep learning models trained on ECG traces already match and in some cases even exceed the performance of medical professionals in diagnosing certain cardiac conditions [15–17]. Given the increasing prevalence of ECG data, machine learning models of such capabilities could transform predictive medicine and cardiovascular research.

In order to use ECGs for age prediction, the neural network needs to learn how the "average" ECG for a particular age group looks. Thus, when it predicts an age considerable larger than the corresponding person's chronological age (a large delta age), this might be indicative of accelerated ageing of the cardiovascular system – with all implications for this individual's health. Indeed, large delta age has been shown to be associated with CVD, treatment outcomes, and mortality [3, 13, 14]. This observation suggests at least two principal areas of applications for the ECG-derived age (or delta age). On one hand, it could be used in the clinic as an readily obtainable prognostic tool for screening large numbers of patients. In this capacity, delta age is conceptually similar to the "excess heart age" [18], the discrepancy between a person's chronological age and their "heart age" (the age corresponding to their risk of a CVD event), which has been devised as an easily interpretable measure for CVD risk [19]. However, while the excess heart age represents the increased CVD risk due to risk factors and lifestyle choices, the delta age reflects the actual functional state of the heart. In addition to clinical use cases, ECG-derived age could also complement biomarkers used in research (e.g. telomere length or the epigenetic clock among others) on ageing in general and vascular ageing in particular.

In addition to the advances in machine learning mentioned above, the availability of genomic data (from microarrays and – more recently – whole-genome sequencing) is ever-increasing. This wealth of information has facilitated a vast number of association studies, linking biological variation in countless phenotypes to the underlying genotypes [20]. Some of these studies investigated the genetic basis of ECG-features (e.g. for the PR interval [21] or the QRS complex [22]), while others tried to determine the impact of genetic variants on the shape of the ECG traces in general [23].

In light of these converging developments, we predicted the "cardiovascular age" of 36,347 participants of the UK biobank (UKB) from their 12-lead ECGs and performed a genome-wide association study (GWAS) on the difference between predicted and chronological age (i.e. delta age). We found eight loci of genome-wide significance ($p \leq 5 \times 10^{-8}$), many of which have been associated with cardiac or muscle development (and in extension with CVD) in the past. Functional and pathway enrichment analyses confirmed this connection to the cardiovascular system. We also explored the association of delta age with specific ECG features, risk factor-derived excess heart age, and the dynamic organism state indicator (DOSI), a complementary biomarker for ageing derived from complete blood count (CBC) data. Overall, our results elucidate the genetic underpinning of this ECG-derived biomarker for cardiovascular age and validate its utility for use in research as well as the clinic.

Results

Predicting age from ECGs in the UK Biobank

We used a previously described deep learning model trained on patients of the Mayo clinic [13] to predict the age of 36,347 participants of the UKB from their 12-lead ECGs. On average, individuals were 64 years old, marginally more likely to be women (52%), and had high levels of education (tertiary education for more than 50%). They comprised a relatively healthy cohort (e.g. less than 6% had diagnosed cardiovascular conditions more severe than hypertension), commonly reporting lifestyle choices considered preventive of CVD (e.g. 63% never smoked), and showing predominantly normal ranges for body mass index (BMI), lipids, and blood pressure (**Table 1**).

As the ECGs in the UKB were noisier than those used for training the model originally [13], an initial signal filtering step was applied prior to prediction. After this pre-processing step, prediction performance on the UKB cohort was comparable to the holdout data set in the original study with a mean absolute error of 6.1 instead of 6.9 years, respectively (**Figure 1**). The Pearson correlation

coefficient between chronological and predicted age was ρ=0.53.

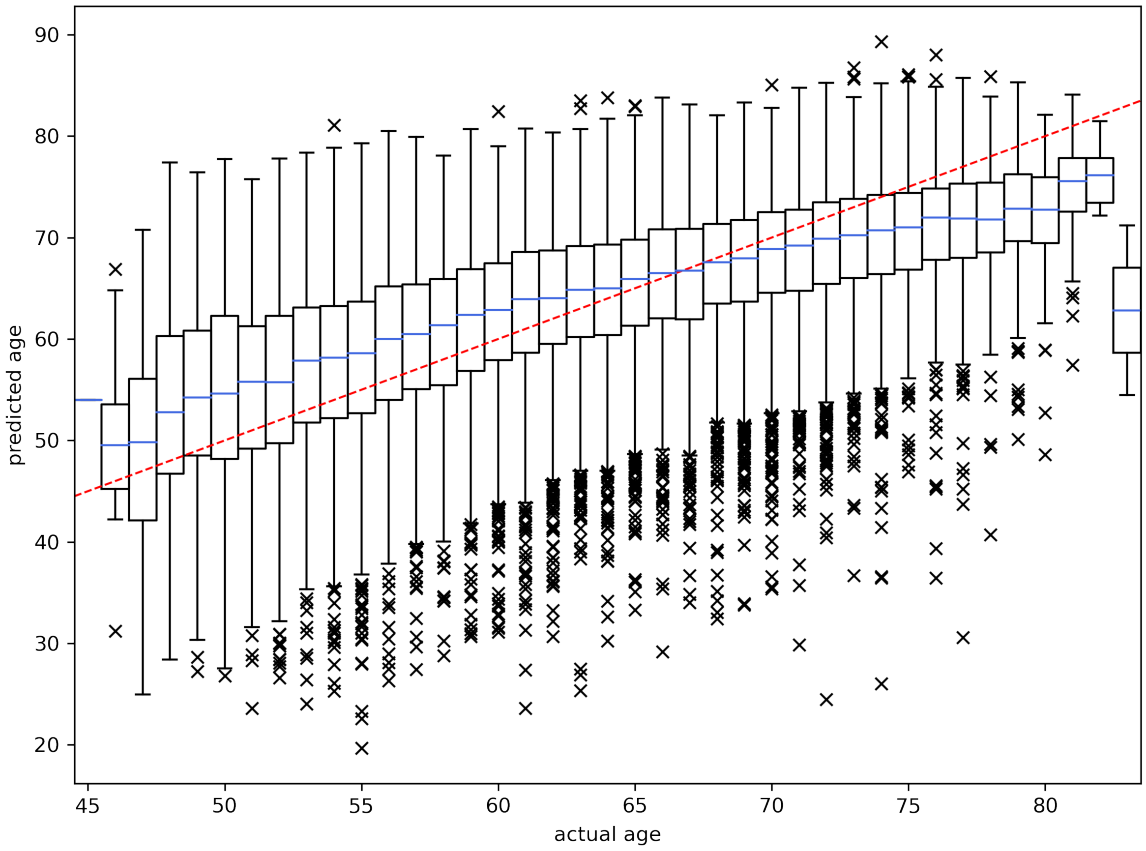


Figure 1: ECG-derived age vs. chronological age for 36,347 participants of the UKB. The Pearson correlation coefficient was 0.53.

The participants' chronological ages were then subtracted from the predicted ages to obtain the delta age (median 0.27; interquartile range -4.81–5.15 years). It was strongly associated with certain anthropometric features and cardiovascular conditions (**Table 1**), consistent with previous studies [3, 13, 14]. When adjusting for age and sex, tertiary education and physical activity were associated with a lower delta age ($p \leq 1 \times 10^{-13}$). BMI, mean arterial pressure (MAP), and low density lipoprotein (LDL), on the other hand, as well as classic cardiovascular risk factors and outcomes, such as frequently drinking alcohol, history of smoking, diagnosed diabetes, hypertension, angina, stroke, or heart attack were associated with higher delta age ($p \leq 3 \times 10^{-3}$). These findings were predominantly robust to multivariate analysis when including all mentioned variables in the model (**Table 1**). Interestingly, men had a lower delta age than women and the negative association with male sex increased when more covariates were taken into account.

Modern ECG machines automatically determine certain human-derived ECG features (e.g. PQ interval, QRS duration) when taking measurements. In the UKB data, many of these features were strongly associated with chronological age, predicted age, or both (**Supplementary Table 1**). However, only a small fraction of the variance in age could be explained by these human-derived features ($r^2 = 0.08$ for a linear regression of age on the ECG features). Interestingly, for the ages predicted by the neural network, this fraction increased almost three-fold ($r^2 = 0.22$), indicating that the model relies on information retained in these features. This insight has also been shown in a recent study which found that some features extracted by the convolutional layers of the neural net were strongly correlated with those defined by humans [24].

Table 1: Association of anthropometric features and cardiovascular risk factors in participants of the UKB with delta age.

Covariate	Info ($N_{total} = 36,349$)	Adjust for age, sex		Adjust for all	
		Effect size	P-value	Effect size	P-value
Sex (male)	17607 (48.4%)	-0.56 (-0.71, -0.41)	**4.7e-13**	-1.15 (-1.31, -0.98)	**4.1e-42**
Age	64.25 (±7.57)	-0.37 (-0.38, -0.36)	**0.0e+00**	-0.40 (-0.41, -0.39)	**0.0e+00**
Education	–	–	**4.6e-18**	–	**2.4e-06**
→ Secondary (ref. level)	14437 (40.1%)	–	–	–	–
→ Tertiary	19186 (53.3%)	-0.69 (-0.85, -0.53)	**1.7e-17**	-0.41 (-0.57, -0.25)	**9.7e-07**
→ Other	2343 (6.5%)	0.04 (-0.29, 0.36)	0.83	0.01 (-0.33, 0.34)	0.98
History of health problems:					
Diabetes	1979 (5.5%)	0.81 (0.47, 1.15)	**2.2e-06**	-0.22 (-0.58, 0.14)	0.23
Hypertension	8419 (23.2%)	1.85 (1.67, 2.04)	**3.7e-88**	0.77 (0.56, 0.97)	**2.1e-13**
Angina	727 (2.0%)	0.88 (0.34, 1.42)	**1.5e-03**	0.11 (-0.48, 0.70)	0.72
Stroke	366 (1.0%)	1.47 (0.71, 2.23)	**1.6e-04**	0.99 (0.20, 1.78)	0.014
Heart attack	524 (1.4%)	1.49 (0.85, 2.13)	**4.6e-06**	1.43 (0.75, 2.12)	**4.1e-05**
Physiological measurements:					
BMI	26.62 (±4.25)	0.24 (0.22, 0.25)	**3.6e-149**	0.16 (0.14, 0.18)	**3.8e-55**
MAP	81.11 (±8.89)	0.13 (0.12, 0.14)	**9.9e-173**	0.10 (0.09, 0.11)	**3.0e-80**
LDL [mM]	3.58 (±0.82)	0.15 (0.05, 0.24)	**2.6e-03**	0.03 (-0.07, 0.13)	0.52
Lifestyle:					
Smoking	–	–	**6.2e-11**	–	**1.6e-04**
→ Never / rarely smoked (ref. level)	22477 (62.5%)	–	–	–	–
→ Active smoker	1300 (3.6%)	0.50 (0.09, 0.92)	0.017	0.41 (-0.01, 0.84)	0.056
→ Smoked in the past	12212 (33.9%)	0.56 (0.40, 0.72)	**2.3e-11**	0.34 (0.17, 0.51)	**7.9e-05**
Alcohol at least 3x per week	16405 (45.2%)	0.24 (0.09, 0.39)	**2.3e-03**	0.33 (0.17, 0.49)	**6.3e-05**
Days of moderate PA per week	3.72 (±1.87)	-0.16 (-0.20, -0.11)	**8.4e-14**	-0.020 (-0.071, 0.030)	0.43
Days of vigorous PA per week	1.93 (±1.58)	-0.25 (-0.30, -0.20)	**1.1e-24**	-0.16 (-0.22, -0.10)	**2.0e-07**

The "Info" column lists the number of corresponding participants for categorical features (with the percentage of the total population in parentheses) or the mean value for numerical features (with the standard deviation in parentheses). P-values and effect sizes in the left double-column are adjusted for age and sex (or only sex for the age-row and vice versa). In the right double-column, the adjustment also includes all other parameters listed in the table. In the "Effect size" columns, values in parentheses denote the lower and upper bounds of the 95% confidence interval. P-values smaller than the Bonferroni-corrected threshold ($0.05/19 = 0.0026$) are highlighted in **bold**. BMI, body mass index; MAP, mean arterial pressure; LDL, low-density lipoprotein; PA, physical activity.

GWAS on delta age

To understand the genetic underpinning of delta age, association tests were performed on ∼6.4 million autosomal variants in 34,432 individuals (after filtering and quality control) while adjusting for age, sex, genotyping array, and UKB assessment centre (**Figure 2**). This analysis revealed eight loci of genome-wide significance ($p \leq 5 \times 10^{-8}$) and another seven loci of suggestive significance ($p \leq 1 \times 10^{-6}$; **Table 2**.

The variants with the strongest association with delta age were detected on chromosome 14 in the gene *SIPA1L1*, which has been linked to ECG features and other cardiac traits according to the GWAS Catalog [26]. Recently, *SIPA1L1* has also been found to be associated with heart trabeculation [27] and it is involved in the regulation of water transport in the kidney [28]. It might thus have

Table 2: Fifteen loci were found to be associated with delta age with at least suggestive significance ($p \leq 1 \times 10^{-6}$).

Chr.	Gene	rsID	Pos.	Ref.	Alt.	AF	Effect size	*P*-value
14	*SIPA1L1*	rs35866366	71849185	A	G	0.25	0.52 (0.39, 0.64)	**1.1e-15**
6	*VGLL2*	rs6901720	117510203	G	T	0.47	0.43 (0.32, 0.54)	**2.8e-14**
4	*CAMK2D*	rs35430511	114387138	T	C	0.26	0.49 (0.36, 0.61)	**3.1e-14**
8	*DEFB136*	rs4240678	11802426	C	T	0.40	0.47 (0.32, 0.62)	**4.9e-10**
2	*TTN*	rs11902709	179608207	C	T	0.05	0.78 (0.52, 1.03)	**3.0e-09**
3	*SCN5A*	rs6773331	38684397	A	T	0.98	1.24 (0.82, 1.66)	**9.1e-09**
3	*SCN10A*	rs6801957	38767315	T	C	0.59	-0.32 (-0.43, -0.21)	**2.1e-08**
5	*PKD2L2*	rs10076361	137252940	G	A	0.18	0.41 (0.27, 0.55)	**2.3e-08**
8	*EXT1*	rs57237854	118860126	ATCTTG	A	0.18	0.40 (0.25, 0.54)	5.3e-08
10	*AGAP5*	rs147790633	75447582	T	C	0.14	-0.43 (-0.59, -0.27)	8.7e-08
10	*CTNNA3*	rs72799115	68008504	G	A	0.21	0.35 (0.22, 0.49)	2.0e-07
12	*TBX3*	rs1896329	115357432	C	T	0.69	-0.31 (-0.42, -0.19)	3.9e-07
2	*SPTBN1*	rs1802889	54756740	C	T	0.68	-0.30 (-0.42, -0.19)	4.4e-07
12	*SOX5*	rs12826024	24776799	G	A	0.15	-0.39 (-0.54, -0.24)	6.1e-07
16	*CHD9*	rs75778953	52906677	C	T	0.01	-1.25 (-1.74, -0.76)	6.2e-07

The second column lists the protein-coding gene closest to the respective lead variant. Positions correspond to the GRCh37 human genome assembly [25]. Values in parentheses denote the lower and upper bounds of the 95% confidence interval of the effect size estimate. *P*-values with genome-wide significance ($p \leq 5 \times 10^{-8}$) are highlighted in bold. Chr., Chromosome; Pos., Position; Ref., Reference allele; Alt., Alternative allele; AF, frequency of the alternative allele.

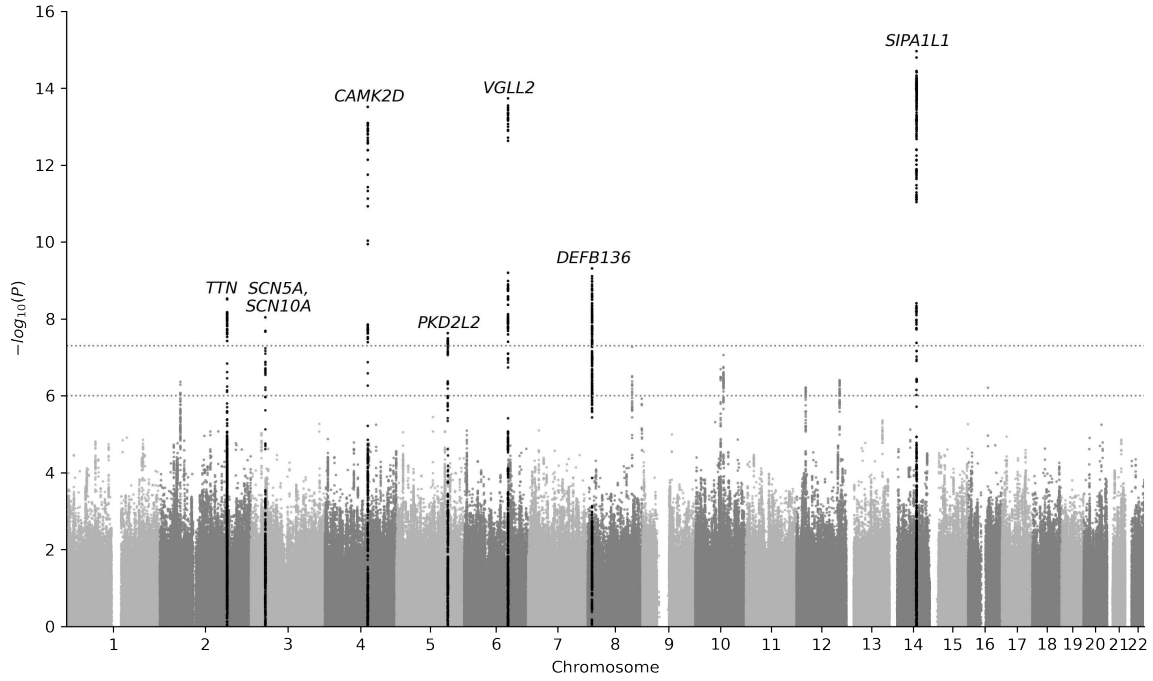


Figure 2: Manhattan plot of the association analysis (n=34,432) adjusted for age, sex, genotyping array, and UKB assessment centre. Horizontal lines mark the thresholds of genome-wide and suggestive significance ($p \leq 5 \times 10^{-8}$ and $p \leq 1 \times 10^{-6}$, respectively).

an impact on the cardiovascular system via kidney function or control of blood volume. However, instead of altering *SIPA1L1*, the causal variant in this locus could alternatively affect the expression levels of *RGS6*, which lies ∼200 kb downstream. *RGS6* is listed in the GWAS Catalog as associated with systolic blood pressure, heart rate, and heart rate variability, for which there is also mechanistic evidence [29].

Another strong association signal was found 30–100 kb upstream of *VGLL2* on chromosome 6. *VGLL2* plays a role in the development of skeletal muscle [30], but, to our knowledge, has not been directly linked to CVD so far. Nonetheless, the GWAS Catalog lists associations with relevant traits like ECG morphology, blood pressure, and atrial fibrillation, but also BMI and waist circumference. Interestingly, *VGLL2* has also been shown to be associated with an age-dependent response to sepsis in the hearts of mice [31]. However, *VGLL2* is not the only protein-coding gene in the region. The next closest (∼100 kb) is *ROS1*, a variant of which has been associated with pathological vascular remodelling [32].

Variants in *CAMK2D* also showed a strong association with delta age. *CAMK2D* encodes the δ chain of the Ca_2^+/calmodulin-dependent protein kinase II, which phosphorylates (in addition to itself) a wide variety of targets involved in a multitude of cellular functions, including neuroplasticity and memory formation [33]. It also plays a role in cardiac Ca_2^+ homeostasis and constitutive activation can lead to CVD and heart failure [34].

The next notable locus was found on chromosome 8 and many of the variants associated with delta age within this locus have also been associated with essential hypertension in the GWAS Catalog. It was located between a group of three genes for β-defensins (*DEFB136*, *DEFB135*, *DEFB134* – with *DEFB136* being the closest) and *CTSB*. Being antimicrobial peptides, β-defensins are an integral part of the innate immune system, but they also have a range of other functions [35]. *CTSB*, located ∼50 kb downstream of the variants associated with delta age, codes for cathepsin B, a protease relevant for proteolysis of intracellular proteins as well as constituents of the extracellular matrix [36]. It has been associated with a large number of diseases, including different types of cancer [37], cardiac remodelling and hypertrophy [38], as well as atherosclerosis [39]. Interestingly, cathepsin B activity has also been shown to increase with age [40].

On chromosome 2, variants in *TTN* were associated with delta age. *TTN* codes for the giant protein titin, responsible for passive mechanical properties of muscle (elasticity and stiffness) and sarcomere structure [41]. Mutations in *TTN* (especially when causing truncations) have been linked to dilated cardiomyopathy (DCM) [42] and the GWAS Catalog mapped a variety of cardiovascular phenotypes and ECG traits to *TTN*, ranging from atrial fibrillation to the PR interval and left ventricular ejection fraction.

SCN5A and the neighbouring *SCN10A* (both on chromosome 3) harboured two independent groups of variants at genome-wide significance. Both genes encode subunits of sodium channels (most prevalent in the myocardium [43] and neurons – including intracardiac ganglia [44], respectively). Variants in *SCN5A* have been linked to multiple cardiac disorders and mutations in both genes can cause the arrhythmia-inducing Brugada syndrome [45, 46].

The last locus of genome-wide significance stretched across ∼400 kb and six protein-coding genes (*KLHL3*, *HNRNPA0*, *MYOT*, *PKD2L2*, *FAM13B*, and *WNT8A*) on chromosome 5. The gene product of *KLHL3* causes the ubiquitination of substrate proteins and is involved in regulating kidney function [47]. It has been associated with a rare hereditary condition with hypertension (familial hyperkalaemic hypertension) [48] and other forms of congenital heart disease in the past [49]. *FAM13B* encodes a GTPase-activating protein, low expression levels of which have been linked to atrial fibrillation [50]. However, if we assume that there is only one causal variant at this locus, it is most likely to be found in *MYOT*, which codes for myotilin, a component of the Z-disc complex in skeletal and cardiac muscle [51]. Myotilin variants can cause myofibrillar myopathy, which sometimes also affects the heart [52]. We did not find any connections with cardiovascular phenotypes for the other three genes, but the GWAS Catalog lists associations with dysrhythmias and atrial fibrillation across the whole 400 kb-spanning locus and beyond.

The seven extra loci found at suggestive significance ($p \leq 1 \times 10^{-6}$) are described in more detail in the Supplementary Information. Most of them were also in the vicinity of genes related to muscle development or the cardiovascular system, but more statistical power (e.g. through larger sample size) will be needed to confirm their associations.

To assess the robustness of our results, the GWAS was repeated with a more extensive suite of covariates (including history of CVD, exercise, and diet; for details see Methods section) and additionally with only those participants that reported a White British ethnic background (**Supplementary Figure 2**). All three analyses showed very similar results qualitatively, with a total of 17 loci reaching at least suggestive significance (**Supplementary Table 6**).

Heritability

The variant-based heritability (h^2_{g}) of delta age was estimated to be ∼12%, being robust to adjustment of cardiovascular risk factors ($12.6 \pm 1.7\%$ for regular adjustment and $11.8 \pm 1.8\%$ for extended adjustment). This magnitude is similar to other ECG traits or cardiac phenotypes, such as PR interval (18.2% [21]), long QT syndrome (14.8% [53]), or atrial fibrillation (9.6% [54]). Interestingly, the 15 loci that reached at least suggestive significance only accounted for ∼15% of the heritability estimate ($h^2_{\mathrm{g,\ top15}} = 1.8 \pm 0.3\%$ and $1.9 \pm 0.3\%$ for regular and extended adjustment, respectively), indicating that there are likely to be many variants with lower significance that are also relevant.

Functional analysis and pathway enrichment

As described above, many loci associated with ECG-derived delta age were found in the vicinity of genes involved in cardiac development or that have been linked to CVD in the past. Application of the DEPICT enrichment analysis tool [55] to the 15 loci with at least suggestive significance ($p \leq 1 \times 10^{-6}$; see **Table 2**) revealed that the GO-term with the strongest signal was "intercalated discs", which are physical connections between cardiomyocytes. The KEGG [56] pathways with the strongest association were mostly linked to calcium signalling and cardiac afflictions, which was also the case with the Mammalian Phenotype Ontology [57] gene sets (**Supplementary Data 1**). We further used DEPICT to test for tissue enrichment. All results with P-values smaller than 0.05 were either connective tissues or part of the cardiovascular system (**Supplementary Table 2**). When including all 179 loci with $p \leq 1 \times 10^{-4}$, geneset and tissue enrichment were both dominated by the cardiovascular system (**Supplementary Data 2**, **Supplementary Table 3**), reinforcing the robustness of our observations.

To confirm these results, we also employed the gProfiler functional enrichment analysis tool [58] on both sets of loci (**Supplementary Table 4**, **Supplementary Table 5**). The results were broadly categorised into heparin / heparan sulfate synthesis, muscle components, calcium / calmodulin signalling, and heart contractions. Like the previous analysis, the strength of the enrichment increased when more loci were included.

Association of variants in telomere-length and longevity-related genes

Interestingly, genes associated with other forms of biological ageing (e.g. telomere length) were mostly absent from the loci found by our analysis. In order to further investigate this surprising result, we scanned the vicinity of loci found by recent GWAS – they had also been performed on the UKB and used longevity [59] and leukocyte telomere lengths [60] as phenotypes – for variants associated with delta age. We found that none of the loci associated with longevity and only two of those associated with telomere length (rs12615793 in *ACYP2* and rs12369950 close to *SOX5*) were within one 1 Mb of variants with at least suggestive significance according to our analysis (**Supplementary Data 3**). In the first case, the lead variant of the locus we discovered was located ∼280 kb downstream of rs12615793 and in *SPTBN1*, which is required for heart development [61]. In the second case, rs12369950 was indeed part of the same locus we found to be associated with delta age.

Further analyses

In order to further investigate the main results described above, we performed statistical tests to detect whether the effects of the genomic variants were mediated via one of the covariates most strongly associated with delta age (BMI, MAP, and diagnosed hypertension), but did not find strong evidence for mediation. Additionally, we ascertained that most of the lead variants have been shown to have a significant impact on the actual shape of the ECG in a recent study [23]. We also calculated the risk factor-based "heart age" [19] and the whole blood counts-derived DOSI biomarker for ageing [62] to contrast both with the ECG-derived cardiovascular age. We found that, while the association with delta age was substantial for the "excess" heart age ($p = 3.0 \times 10^{-78}$), it was weak for the "excess" DOSI ($p \geq 1.4 \times 10^{-3}$). These findings are described in greater detail in the Supplementary Information.

Discussion

We used a deep neural network to predict the age of 36,347 individuals in the UKB from their 12-lead ECGs and observed that – similar to what has been shown in other populations [3, 13, 14] – the discrepancy to their chronological age was correlated with cardiovascular risk factors like blood pressure, BMI, and smoking status. In addition to these covariates, we also found 15 genetic loci of at least suggestive significance ($p \leq 1 \times 10^{-6}$), eight of which reached genome-wide significance ($p \leq 5 \times 10^{-8}$) in a GWAS adjusted for age, sex, genotyping array, and UKB assessment centre. We evaluated the robustness of these results by repeating the GWAS with a more extensive set of covariates including past CVD diagnoses and lifestyle variables, such as diet or the amount of physical exercise. We also carried out another round of association tests with only the subset of individuals of European ethnic origin. All three analyses yielded very similar results. Overall, about 12% of the variation in delta age could be explained by the genomic data, which is comparable to other cardiac phenotypes (e.g. 9.6% for atrial fibrillation [54]).

In order to determine whether the associations of the lead variants with the phenotype were direct and not mediated via an intermediate factor, we performed tests for mediation for the covariates most strongly associated with delta age (MAP, BMI, and diagnosed hypertension). There appeared to be weak mediating effects for some of the variants, but the signal was not strong enough to remain significant after correcting for multiple tests ($p \geq 0.024$). However, some metadata entries in the UKB were recorded a considerable amount of time before the imaging visit when the ECG was taken and some of the covariates might have changed in the intervening period. Because of this limitation and given the large number of (genetic and environmental) factors influencing cardiovascular health and ECG morphology, it is possible that stronger mediating effects might have been missed in the present

study. More research will be required in order to disentangle the network of interactions between genetic and non-genetic variables affecting cardiovascular age and its impact on the ECG.

Most of the loci discovered in our GWAS analysis have either been associated with CVD in the past or were located in the vicinity of genes involved in cardiovascular function. Functional analyses with the DEPICT enrichment analysis tool [55] found significant over-representation of gene sets related to cardiac and muscle development as well as of genes expressed in the corresponding tissues. These associations were confirmed with an alternative method (gProfiler [58]) and grew stronger and more robust when variants with weaker association with delta age were included in the analysis (i.e. when using P-value cutoffs of $p \leq 1 \times 10^{-5}$ or $p \leq 1 \times 10^{-4}$). Similarly, only a small fraction ($\sim$15%) of the heritability we found could be explained by the 15 top loci. Together, these two findings suggest that many of the variants with only moderate significance might also be potential components of the genetic basis of delta age, but larger studies will be needed to verify their signal.

In addition to their links to CVD, the lead variants in most loci of genome-wide significance have also been associated with the actual shape of the ECG in a recent study [23]. This is a promising sign as it might help to illuminate the "black box" character of the neural network used for age prediction. In general, the knowledge about the effects of age on the ECG and the impact of genetic variants should be combined in order to aid in the interpretation of results produced by opaque deep learning models in the medical domain.

Several different biomarkers for ageing have been proposed in the last two decades, with telomere length and the epigenetic clock arguably receiving the most attention. Despite each being a good predictor for mortality, these metrics were shown to only correlate weakly with each other, implying that they are governed by different aspects of the mechanisms of ageing [8, 9]. We observed something similar as we did not find a strong association of variants previously linked to ageing [59] or telomere length [60] with delta age. We also calculated the DOSI, a blood counts-derived marker for biological ageing and physiological resilience [62], for our cohort and – as opposed to the risk factor-derived "excess" heart age – correlation of the "excess" DOSI with delta age was inconclusive. More research relating different markers of biological ageing with delta age is needed, but the available evidence suggests that genetic variants associated with more general forms of ageing (e.g. in *APOE*, *FOXO3*, *TERT*, *LMNA*) have little impact on cardiovascular age compared to genes involved in the development and function of the cardiovascular system itself.

Viewed in their entirety, our findings corroborate that the ECG-derived age reflects the physiological state of the heart and that it can be used to assess cardiovascular ageing and health. Interestingly, for two of the loci with the strongest association with delta age (*SIPA1L1* and *VGLL2*), the connection to cardiovascular phenotypes in the literature was not as clear as for many others. They therefore represent promising targets for deeper mechanistic investigation in future work. Additionally, efforts on fine-mapping will be needed to identify individual causal variants and also to confirm relevant genes since variants in linkage disequilibrium with the lead variant spanned hundreds of kilobases for some of the loci found in this study. This raises the opportunity of narrowing down the range of potential causal variants with association studies in populations of non-European ancestry.

Our work shows that genetic factors underlying cardiovascular ageing and its effect on the ECG should be incorporated into prediction models in order to improve their accuracy and interpretability. In a future of personalised medicine with readily available genomic information, the non-invasive ECG (including from wearable devices), combined with an easily obtainable measure of ECG-derived delta age, will be a valuable instrument in the clinicians' toolkit for assessing heart health at routine examinations and monitoring treatment outcomes. Moreover, resources like the UKB, hosting an ever-increasing wealth of genomic, epigenetic, and transcriptomic data, will facilitate better comparisons as well as deeper understanding of the individual biomarkers for ageing, their underlying mechanisms, and how they complement one another. Ultimately, large-scale analysis of such data, combined with artificial intelligence (AI) methodologies, will translate patient-level genomic and ECG information into preventative medicine and public health measures, leading to earlier detection of CVD and a longer healthspan.

Methods

Study population

This work has been conducted using data from the UK Biobank, which recruited 500,000+ people aged between 40-69 years in 2006-2010 from across the United Kingdom. With their consent, they provided detailed information about their lifestyle, had physical measures taken as well as blood, urine and saliva samples collected and stored for future analysis. We used the 10-second 12-lead ECG traces and CVD-related metadata of 37,520 participants. The ECGs were recorded during the first imaging visit (after 2014) and the metadata questionnaires were completed during the initial and first repeat assessment visits (2006–2010 and 2012–2013, respectively). All methods were performed in accordance with relevant guidelines and regulations posed by the UKB and approved by the London School of Hygiene & Tropical Medicine ethics committee. The project application reference was 54050 (www.ukbiobank.ac.uk).

Deep learning model, ECG pre-processing, and age prediction

The architecture and training procedure of the deep learning model used in this study are described in detail elsewhere [13]. In brief, 499,727 10-second 12-lead ECGs of patients of the Mayo clinic were used to train a convolutional neural network to predict patient age and a holdout dataset of 275,056 patients was used for testing model performance. Due to the ECGs in the UKB being noisier than the training data, they had to undergo a filtering step prior to prediction. This was achieved using a four-pole Butterworth filter allowing frequencies from 0.5 to 100 Hz to pass. After this pre-processing step, ECG-derived age was predicted for 36,347 individuals in the UKB.

Metadata processing

Whenever multiple measurements of a relevant variable were available for a given sample, the mean or the value with the smallest time gap to the ECG recording was used for continuous and categorical data, respectively. MAP was calculated from systolic (SBP) and diastolic blood pressure (DBP) measurements using the equation (SBP + 2 · DBP) / 3. These MAP values were then averaged with the MAP measurements derived from Pulse Wave Analysis to give the final values. The UKB contains a host of diet variables ranging from the amount of raw vegetables eaten per day to the type of fat used for cooking. We performed principal component analysis (PCA) on a selection of 24 of these variables and included the first three principal components (accounting for ∼25% of the total variation) as covariates in the GWAS with extended adjustment (see below).

Association testing

Pre-processing of genotype data and association testing were carried out using PLINK (v. 2.00) [63]. For quality control, we removed variants that either (i) were missing in more than 1% of samples, (ii) showed a minor allele frequency under 1%, (iii) were not in Hardy-Weinberg Equilibrium ($P < 1 \times 10^{-6}$), or (iv) had an imputation score below 0.8. Samples with more than 2% missing genotypes or that were outside of three standard deviations from the mean heterozygosity were dropped. Additionally, one sample from each closely related pair (first or second degree relations as determined by KING robust kinship inference [64]) was removed. The dimension of the final genotype matrix was 34,432 samples times 6,357,764 autosomal variants. PCA [65] was performed on this matrix and the first 10 principal components were retained for use as covariates in the association tests.

In total, four GWAS with delta age as phenotype were carried out. The main analysis included all participants remaining after filtering and adjusted for age, sex, genotyping array, and UKB assessment centre. Additionally, in order to assess the robustness of the results, the association tests were repeated with an extended set of covariates: education (secondary, tertiary, other); smoking status (current smoker, past smoker, never / rarely smoked); alcohol consumption three or more times per week; having been diagnosed with diabetes, hypertension, angina, stroke, or heart attack in the past; BMI; MAP; LDL concentration; days of moderate exercise per week; days of vigorous exercise per week; three

principal components derived from a PCA of 24 diet variables available in the UKB. Both analyses were then repeated with the subset of participants with white British as ethnic background ($N = 31,971$).

Heritability estimation and pathway enrichment analysis

The variant-based heritability of delta age was estimated using GREML-LDMS [66] implemented in GCTA (v. 1.93.2) [67] while stratifying the variants based on linkage disequilibrium (four bins) and minor allele frequency (MAF) (two bins with MAF = 0.05 as boundary). The analysis was carried out with both sets of covariates and later repeated with the subsets of variants found within the 15 loci of at least suggestive significance in order to also calculate the heritability of the top hits found by the GWAS. Genomic position ranges of the individual loci were calculated as part of the DEPICT workflow. DEPICT [55] and gProfiler [58] were used for pathway and tissue enrichment analyses. DEPICT was run on the GWAS summary statistics with $p = 1 \times 10^{-6}$ and $p = 1 \times 10^{-4}$ as thresholds. It uses PLINK internally to determine independent loci based on the P-value threshold and a 500 kb clumping window before testing for gene set and tissue enrichment relying on data from the following databases: Gene Ontology [68], KEGG [56], Reactome [69], InWeb [70], Mouse Genome Database [71], and Gene Expression Omnibus [72]. The coordinates of the loci found by DEPICT were additionally pasted into the gProfiler web tool, which tested for enrichment based on the Gene Ontology, KEGG, Reactome, WikiPathways [73], TRANSFAC [74], miRTarBase [75], Human Protein Atlas [76], CORUM [77], and Human Phenotype Ontology [78] databases.

Data availability

All data is available from the UKB (www.ukbiobank.ac.uk).

Code availability

All scripts used open source software (see Methods).

Acknowledgements

T.G.C. was funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and a Wellcome Trust Strategic Award (Grant no. 100217/Z/12/A). D.A.L. was funded by a Wellcome Trust Strategic Award (Grant no. 100217/Z/12/A).The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

D.A.L and T.G.C conceived the project and applied for access to the UK Biobank data. E.D.B assisted with drafting the UK Biobank application. Z.I.A generated the predicted age using a convolutional neural network, with the support of P.A.F and F.L.-J. J.L.-E and J.E.P performed the data processing and analysis, under the supervision of T.G.C, with feedback on results from Z.I.A, S.C, F.L.-J and D.A.L. J.L.-E. and T.G.C wrote the first draft of the manuscript. All authors commented on versions of the manuscript and approved the final manuscript.

Competing interest statement

P.A.F., Z.I.A., and F.L.-J. have filed intellectual property related to the AI algorithm used here to detect biological age from the ECG. The remaining authors declare no competing interests.

Materials and Correspondence

Prof. Taane G. Clark (taane.clark@lshtm.ac.uk)

References

1. Simonson, E. The effect of age on the electrocardiogram. *The American journal of cardiology* **29,** 64–73 (1972).
2. Vicent, L. & Martínez-Sellés, M. Electrocardiogeriatrics: ECG in advanced age. *Journal of electrocardiology* **50,** 698–700 (2017).
3. Ladejobi, A. O. *et al.* The 12-lead electrocardiogram as a biomarker of biological age. *European Heart Journal-Digital Health* (2021).
4. Blackburn, E. H., Epel, E. S. & Lin, J. Human telomere biology: a contributory and interactive factor in aging, disease risks, and protection. *Science* **350,** 1193–1198 (2015).
5. Fransquet, P. D., Wrigglesworth, J., Woods, R. L., Ernst, M. E. & Ryan, J. The epigenetic clock as a predictor of disease and mortality risk: a systematic review and meta-analysis. *Clinical epigenetics* **11,** 1–17 (2019).
6. Levine, M. E. Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age? *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* **68,** 667–674 (2013).
7. Peters, M. J. *et al.* The transcriptional landscape of age in human peripheral blood. *Nature communications* **6,** 1–14 (2015).
8. Belsky, D. W. *et al.* Eleven telomere, epigenetic clock, and biomarker-composite quantifications of biological aging: do they measure the same thing? *American Journal of Epidemiology* **187,** 1220–1230 (2018).
9. Jylhävä, J., Pedersen, N. L. & Hägg, S. Biological age predictors. *EBioMedicine* **21,** 29–36 (2017).
10. Yazdanyar, A. & Newman, A. B. The burden of cardiovascular disease in the elderly: morbidity, mortality, and costs. *Clinics in geriatric medicine* **25,** 563 (2009).
11. Ball, R. L., Feiveson, A. H., Schlegel, T. T., Starc, V. & Dabney, A. R. Predicting heart age using electrocardiography. *Journal of personalized medicine* **4,** 65–78 (2014).
12. Shah, A. P. & Rubin, S. A. Errors in the computerized electrocardiogram interpretation of cardiac rhythm. *Journal of electrocardiology* **40,** 385–390 (2007).
13. Attia, Z. I. *et al.* Age and sex estimation using artificial intelligence from standard 12-lead ECGs. *Circulation: Arrhythmia and Electrophysiology* **12,** e007284 (2019).
14. Lima, E. M. *et al.* Deep neural network estimated electrocardiographic-age as a mortality predictor. *medRxiv* (2021).
15. Hannun, A. Y. *et al.* Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine* **25,** 65–69 (2019).
16. Ribeiro, A. H. *et al.* Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature communications* **11,** 1–9 (2020).
17. Kwon, J.-M. *et al.* Comparing the performance of artificial intelligence and conventional diagnosis criteria for detecting left ventricular hypertrophy using electrocardiography. *EP Europace* **22,** 412–419 (2020).
18. Yang, Q. *et al.* Vital signs: predicted heart age and racial disparities in heart age among US adults at the state level. *Morbidity and Mortality Weekly Report* **64,** 950–958 (2015).
19. DAgostino Sr, R. B. *et al.* General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* **117,** 743–753 (2008).
20. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Communications biology* **2,** 1–11 (2019).
21. Ntalla, I. *et al.* Multi-ancestry GWAS of the electrocardiographic PR interval identifies 202 loci underlying cardiac conduction. *Nature communications* **11,** 1–12 (2020).
22. Norland, K. *et al.* Sequence variants with large effects on cardiac electrophysiology and disease. *Nature communications* **10,** 1–10 (2019).

23. Verweij, N. *et al.* The genetic makeup of the electrocardiogram. *Cell systems* **11,** 229–238 (2020).

24. Attia, Z. I., Lerman, G. & Friedman, P. A. Deep neural networks learn by using human-selected electrocardiogram features and novel features. *European Heart Journal-Digital Health* **2,** 446–455 (2021).

25. Church, D. M. *et al.* Modernizing reference genome assemblies. *PLoS biology* **9,** e1001091 (2011).

26. Buniello, A. *et al.* The NHGRI-EBIws GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research* **47,** D1005–D1012 (2019).

27. Meyer, H. V. *et al.* Genetic and functional insights into the fractal structure of the heart. *Nature* **584,** 589–594 (2020).

28. Wang, P.-J. *et al.* Vasopressin-induced serine 269 phosphorylation reduces Sipa1l1 (signal-induced proliferation-associated 1 like 1)-mediated aquaporin-2 endocytosis. *Journal of Biological Chemistry* **292,** 7984–7993 (2017).

29. Yang, J. *et al.* RGS6, a modulator of parasympathetic activation in heart. *Circulation research* **107,** 1345–1349 (2010).

30. Honda, M. *et al.* Vestigial-like 2 contributes to normal muscle fiber type distribution in mice. *Scientific reports* **7,** 1–12 (2017).

31. Checchia, P. A. *et al.* Myocardial transcriptional profiles in a murine model of sepsis: Evidence for the importance of age. *Pediatric Critical Care Medicine* **9,** 530–535 (2008).

32. Ali, Z. A. *et al.* Oxido-reductive regulation of vascular remodeling by receptor tyrosine kinase ROS1. *The Journal of clinical investigation* **124,** 5159–5174 (2014).

33. Bayer, K. U. & Schulman, H. CaM kinase: still inspiring at 40. *Neuron* **103,** 380–394 (2019).

34. Mattiazzi, A. *et al.* Chasing cardiac physiology and pathology down the CaMKII cascade. *American Journal of Physiology-Heart and Circulatory Physiology* **308,** H1177–H1191 (2015).

35. Shelley, J. R., Davidson, D. J. & Dorin, J. R. The dichotomous responses driven by β-Defensins. *Frontiers in immunology* **11,** 1176 (2020).

36. Yadati, T., Houben, T., Bitorina, A. & Shiri-Sverdlov, R. The ins and outs of cathepsins: physiological function and role in disease management. *Cells* **9,** 1679 (2020).

37. Aggarwal, N. & Sloane, B. F. Cathepsin B: multiple roles in cancer. *PROTEOMICS–Clinical Applications* **8,** 427–437 (2014).

38. Blondelle, J., Lange, S., Greenberg, B. H. & Cowling, R. T. Cathepsins in heart disease–chewing on the heartache? *American Journal of Physiology-Heart and Circulatory Physiology* **308,** H974–H976 (2015).

39. Mareti, A. *et al.* Cathepsin B expression is associated with arterial stiffening and atherosclerotic vascular disease. *European journal of preventive cardiology* **27,** 2288–2291 (2020).

40. Wyczalkowska-Tomasik, A. & Paczek, L. Cathepsin B and L activity in the serum during the human aging process: cathepsin B and L in aging. *Archives of gerontology and geriatrics* **55,** 735–738 (2012).

41. LeWinter, M. M. & Granzier, H. Cardiac titin: a multifunctional giant. *Circulation* **121,** 2137–2145 (2010).

42. Tharp, C. A., Haywood, M. E., Sbaizero, O., Taylor, M. R. & Mestroni, L. The giant protein titins role in cardiomyopathy: genetic, transcriptional, and post-translational modifications of TTN and their contribution to cardiac disease. *Frontiers in Physiology,* 1436 (2019).

43. Remme, C. *et al.* The cardiac sodium channel displays differential distribution in the conduction system and transmural heterogeneity in the murine ventricular myocardium. *Basic research in cardiology* **104,** 511–522 (2009).

44. Verkerk, A. O. *et al.* Functional Nav1.8 channels in intracardiac neurons: the link between SCN10A and cardiac electrophysiology. *Circulation research* **111,** 333–343 (2012).

45. Li, W. *et al.* SCN5A Variants: Association With Cardiac Disorders. *Frontiers in Physiology* **9,** 1372. ISSN: 1664-042X. https://www.frontiersin.org/article/10.3389/fphys.2018.01372 (2018).

46. Hu, D. *et al.* Mutations in SCN10A are responsible for a large fraction of cases of Brugada syndrome. *Journal of the American College of Cardiology* **64,** 66–79 (2014).

47. Gong, Y. *et al.* KLHL3 regulates paracellular chloride transport in the kidney by ubiquitination of claudin-8. *Proceedings of the National Academy of Sciences* **112,** 4340–4345 (2015).

48. Glover, M. *et al.* Detection of mutations in KLHL3 and CUL3 in families with FHHt (familial hyperkalaemic hypertension or Gordon's syndrome). *Clinical Science* **126,** 721–726 (2014).

49. Wang, L., Lai, G., Chu, G., Liang, X. & Zhao, Y. cMyBP-C was decreased via KLHL3-mediated proteasomal degradation in congenital heart diseases. *Experimental cell research* **355,** 18–25 (2017).

50. Hsu, J. *et al.* Genetic control of left atrial gene expression yields insights into the genetic susceptibility for atrial fibrillation. *Circulation: Genomic and Precision Medicine* **11,** e002107 (2018).

51. Wang, J., Dube, D. K., Mittal, B., Sanger, J. M. & Sanger, J. W. Myotilin dynamics in cardiac and skeletal muscle cells. *Cytoskeleton* **68,** 661–670 (2011).

52. Olivé, M., Kley, R. A. & Goldfarb, L. G. Myofibrillar myopathies: new developments. *Current opinion in neurology* **26,** 527 (2013).

53. Lahrouchi, N. *et al.* Transethnic genome-wide association study provides insights in the genetic architecture and heritability of long QT syndrome. *Circulation* **142,** 324–338 (2020).

54. Nielsen, J. B. *et al.* Genome-wide study of atrial fibrillation identifies seven risk loci and highlights biological pathways and regulatory elements involved in cardiac development. *The American Journal of Human Genetics* **102,** 103–115 (2018).

55. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nature communications* **6,** 1–9 (2015).

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

57. Smith, C. L. & Eppig, J. T. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **1,** 390–399 (2009).

58. Raudvere, U. *et al.* g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research* **47,** W191–W198 (2019).

59. Pilling, L. C. *et al.* Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging (Albany NY)* **9,** 2504 (2017).

60. Codd, V. *et al.* Polygenic basis and biomedical consequences of telomere length variation. *Nature genetics* **53,** 1425–1433 (2021).

61. Yang, P. *et al.* βII spectrin (SPTBN1): biological function and clinical potential in cancer and other diseases. *International journal of biological sciences* **17,** 32 (2021).

62. Pyrkov, T. V. *et al.* Longitudinal analysis of blood markers reveals progressive loss of resilience and predicts human lifespan limit. *Nature communications* **12,** 1–10 (2021).

63. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** s13742–015 (2015).

64. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010).

65. Galinsky, K. J. *et al.* Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *The American Journal of Human Genetics* **98,** 456–472 (2016).

66. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature genetics* **47,** 1114–1120 (2015).

67. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88,** 76–82 (2011).

68. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature genetics* **25,** 25–29 (2000).

69. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research* **39,** D691–D697 (2010).

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

71. Blake, J. A. *et al.* The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic acids research* **42,** D810–D817 (2014).

72. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data setsupdate. *Nucleic acids research* **41,** D991–D995 (2012).

73. Slenter, D. N. *et al.* WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic acids research* **46,** D661–D667 (2018).

74. Matys, V. *et al.* TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic acids research* **34,** D108–D110 (2006).

75. Chou, C.-H. *et al.* miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic acids research* **46,** D296–D302 (2018).

76. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347,** 1260419 (2015).

77. Giurgiu, M. *et al.* CORUM: the comprehensive resource of mammalian protein complexes2019. *Nucleic acids research* **47,** D559–D563 (2019).

78. Köhler, S. *et al.* Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic acids research* **47,** D1018–D1027 (2019).

Supplementary Information

GWAS results of suggestive significance

The genome-wide association study (GWAS) with regular adjustment found eight loci associated with delta age at genome-wide significance ($p \leq 5 \times 10^{-8}$; for details see Results section in the main text) and seven loci at suggestive significance ($p \leq 1 \times 10^{-6}$) which are described in more detail below.

EXT1 on chromosome 8 was just short of reaching genome-wide significance. It encodes a glycosyltransferase involved in the synthesis of heparan sulfate, a component of the extracellular matrix with a wide array of functions [1], including a crucial role in early development [2]. *EXT1* itself has also been linked to the development of the heart [3]. On chromosome 10, two loci with lead variants in *AGAP5* and *CTNNA3* were identified. *CTNNA3* codes for αT-catenin, a component of cell junctions called intercalated discs which mechanically connect myocytes [4]. The other locus comprised a large number of variants in linkage disequilibrium, spanning ∼200 kb and several genes (*MYOZ1*, *SYNPO2L*, *AGAP5*, *SEC24C*, *FUT11*, *CHCHD1*, *ZSWIM8*, *AC022400.5*, *NDST2*, and *CAMK2G*). Out of these, *MYOZ1* and *SYNPO2L* are relevant for cardiac processes. *MYOZ1* binds to Z-disc proteins, is involved in calcineurin signalling [5], and mutations in it have been linked to atrial fibrillation [6] as well as heart failure [7]. Similarly, *SYNPO2L* also interacts with Z-disc proteins and it too has been shown to be associated with atrial fibrillation [8, 9]. Moreover, the GWAS Catalog lists associations for *SYNPO2L* with electrocardiogram (ECG) morphology, hypertrophic cardiomyopathy, and blood pressure, among other cardiac phenotypes.

Variants in *SOX5* on chromosome 12 were also associated with delta age at suggestive significance, with links to ECG morphology, atrial fibrillation, and blood pressure, in the GWAS Catalog. *SOX5* is involved in heart development in *D. melanogaster* [10] and has been found to be associated with dilated cardiomyopathy (DCM) [11]. The second locus on chromosome 12 was found ∼225 kb upstream of *TBX3*, which was the closest protein-coding gene. It is a transcription factor involved in the development of multiple tissues, including within the heart [12]. Variants upstream of *TBX3* have been associated with pacemaker cell function [13]. The GWAS Catalog also lists associations with multiple cardiac phenotypes for *TBX3*, including blood pressure, ECG morphology, PR interval, heart failure, and others.

In addition to *TTN*, which reached genome-wide significance, another locus on chromosome 2 was found to be associated with delta age. *SPTBN1* encodes II spectrin, which is required for normal cardiac development as well as excitability [14]. In the GWAS Catalog it has been linked to a long list of phenotypes, some of which relate to the cardiovascular system (e.g. ECG morphology, PR interval, or glomerular filtration rate).

The final locus reaching suggestive significance was located on chromosome 16 with the closest protein-coding gene being *CHD9*, a chromatin remodelling protein. We did not find a direct connection of *CHD9* to heart development or cardiovascular disease (CVD) and no cardiac phenotypes were linked in the GWAS Catalog. Also, only a single rare (minor allele frequency (MAF) =1.2%) variant crossed the significance threshold at this locus and it was not in strong linkage disequilibrium with any other variants in the region. Moreover, it lost suggestive significance in the dataset with only white British ethnic background (**Supplementary Table 6**). Taken together, these findings might indicate a spurious association. Thus, in order to confirm the signal at this locus, future studies with larger and more ethnically diverse datasets are needed.

Mediation analyses

Whilst most GWAS results were robust with respect to the differing levels of adjustment for potential confounders, there were some lead variants with considerable quantitative (but not qualitative) differences in significance. This observation may arise from the variants affecting delta age (at least partially) via mediation of one or more of the covariates (e.g. a variant not influencing delta age directly but rather blood pressure which itself then impacts ECG-derived age). Therefore, we ran mediation analyses for mean arterial pressure (MAP), body mass index (BMI), and hypertension as individual mediators while adjusting for the regular covariates. They were carried out using the Pingouin Python package [15] and with 5,000 bootstrap iterations per test. The results suggested minor mediating effects for some variant-mediator combinations, but none of them remained significant when

taking multiple testing into account ($p_{min} = 0.024$; **Supplementary Data 4**), indicating that the impact of the genomic variants on delta age was indeed direct.

Impact of lead variants on ECG morphology

In a recent study, 3-lead ECGs were aligned and averaged to generate one representative time trace of a single heartbeat for more than 70,000 participants in the UK biobank (UKB) [16]. The aligned signal was then split into 500 data points per individual and 500 GWAS were performed to identify variants associated with each of these subsections of the ECG. This resulted in 300+ significant loci, many of which had not been reported in a cardiac context before, and revealed novel insights into diseases like DCM and early repolarisation. A webtool (`www.ecgenetics.org`) was developed for browsing the results and to look up the impact of a given variant on the shape of the ECG. We used this tool to assess our GWAS results. Seven out of eight loci with genome-wide significant association with delta age also exhibited genome-wide significance ($p = 1 \times 10^{-10}$ when correcting for 500 tests) for at least one section of the ECG in the study of Verweij et *al.* [16] (**Supplementary Table 6**). Out of all lead variants found by our analyses, those in *SCN5A*, *SCN10A*, and *TBX3* were shown to have the strongest impact on ECG morphology, all three of which are connected to the electrophysiology of the heart (for details see above and the Results section in the main text).

Comparison with risk factor-derived "heart age"

The "heart age" of a person is a function of a list of risk factors and is defined as the age of somebody who has the same risk of CVD as the original person, but for whom all these risk factors are at normal levels [17]. Thus, with the exception of being derived exclusively from risk factors, the "excess heart age" (the difference between heart age and chronological age) is conceptually similar to delta age. In order to calculate the heart age of the individuals in our cohort, we extracted data on the relevant risk factors from the UKB and adapted the source code (available at `https://github.com/Antbits/heartage`) of an online tool [18] of the National Health Service, which relies on a model for heart age fitted on a British population [19]. We then subtracted the heart age from the chronological age of the participants to determine the excess heart age and compared it to delta age. Overall, excess heart age was strongly correlated with delta age ($p = 3.0 \times 10^{-78}$ when adjusting for chronological age; delta age increased by 0.108–0.134 for each year of excess heart age). This was not surprising as many of the the risk factors comprising the heart age were individually correlated with delta age (**Table 1**), but it reaffirms the connection of structural / functional features of the heart – as picked up by the ECG-derived delta age – with cardiovascular risk factors and lifestyle choices.

Association with the Dynamic Organism State Indicator

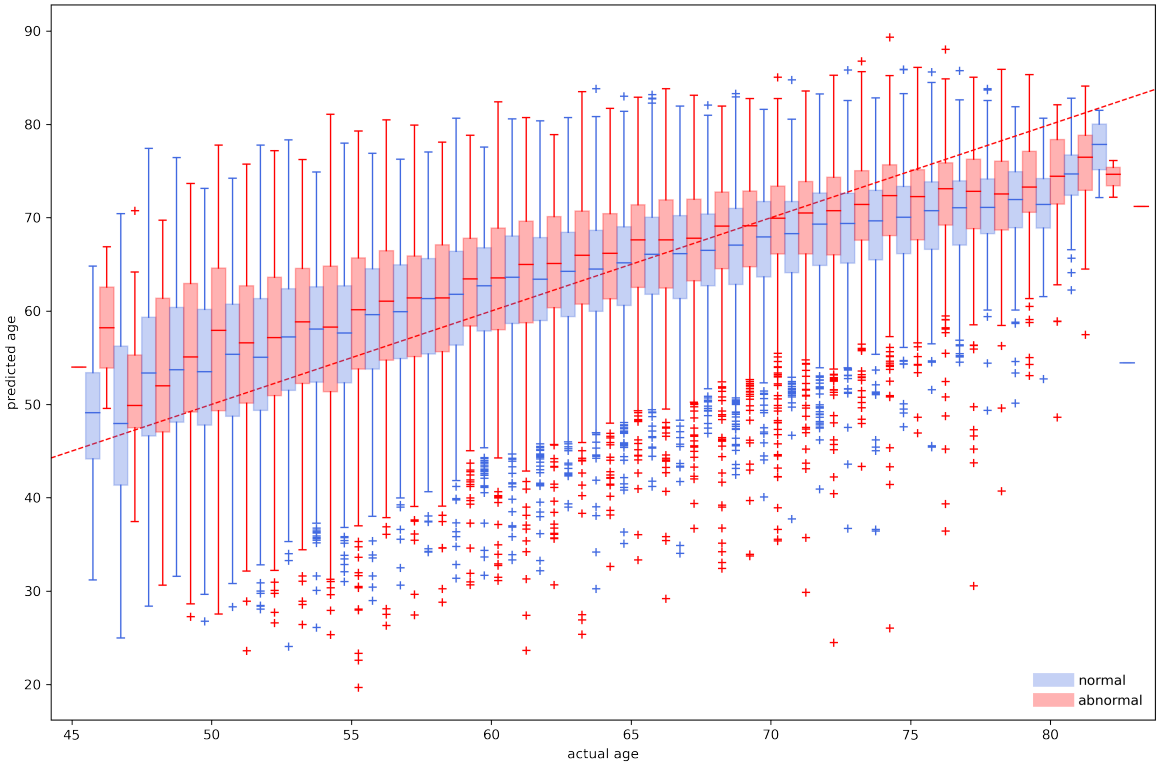
The relevant blood count data were extracted and used to determine the dynamic organism state indicator (DOSI) metric [20]. Then, the "excess" DOSI was calculated by fitting cubic splines to the DOSI against age for each sex and subsequently subtracting the value of the spline for the corresponding age from the individual DOSI values. When adjusting for age, sex, and the UKB assessment centre, we found neither the DOSI nor the excess DOSI to be associated with delta age. However, when including all other covariates which were also corrected for in the GWAS with extended adjustment, there was a weak association with both metrics ($p \leq 4 \times 10^{-3}$; see **Supplementary Table 7**).

GWAS on abnormal versus normal ECGs

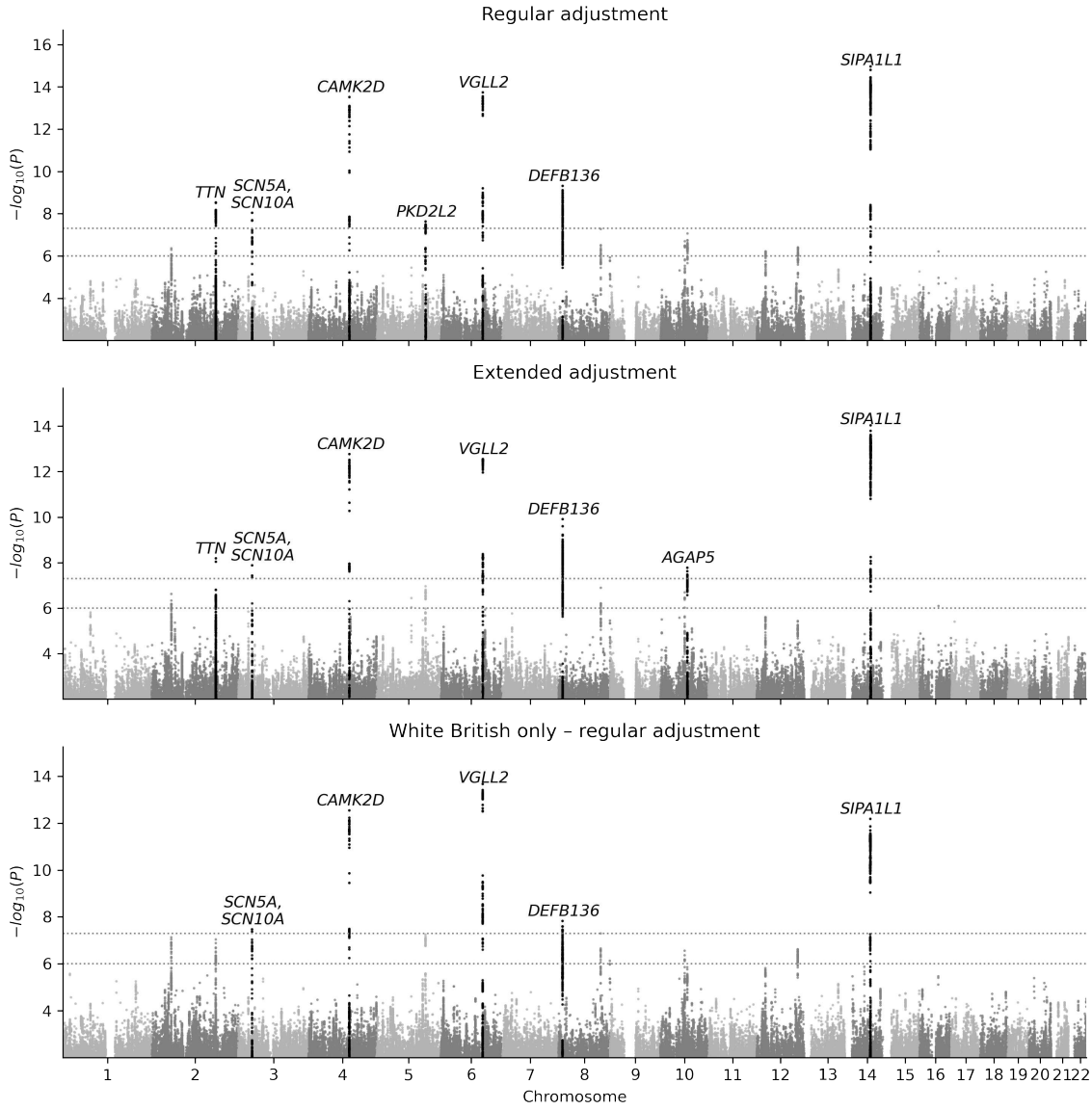
Many ECG machines give automatic diagnoses when taking a measurement, including reporting a "normal" or "abnormal" status. Age predictions on such "abnormal" ECGs were consistently higher on average compared to "normal" ECGs (**Supplementary Figure 1**). In addition to the delta age analysis, we ran two extra GWAS (using regular and extended adjustment like for the main analysis) on the abnormal-vs-normal phenotype to see if we could replicate the findings of a recent study done on a relatively small Chinese cohort ($n = 1006$) [21]. However, no variants reached genome-wide significance ($p \leq 5 \times 10^{-8}$) and the variants reaching suggestive significance ($p = 1 \times 10^{-6}$) adjusting

for either set of covariates were different from those reported by the study (**Supplementary Figure 3**, **Supplementary Table 8**).‘

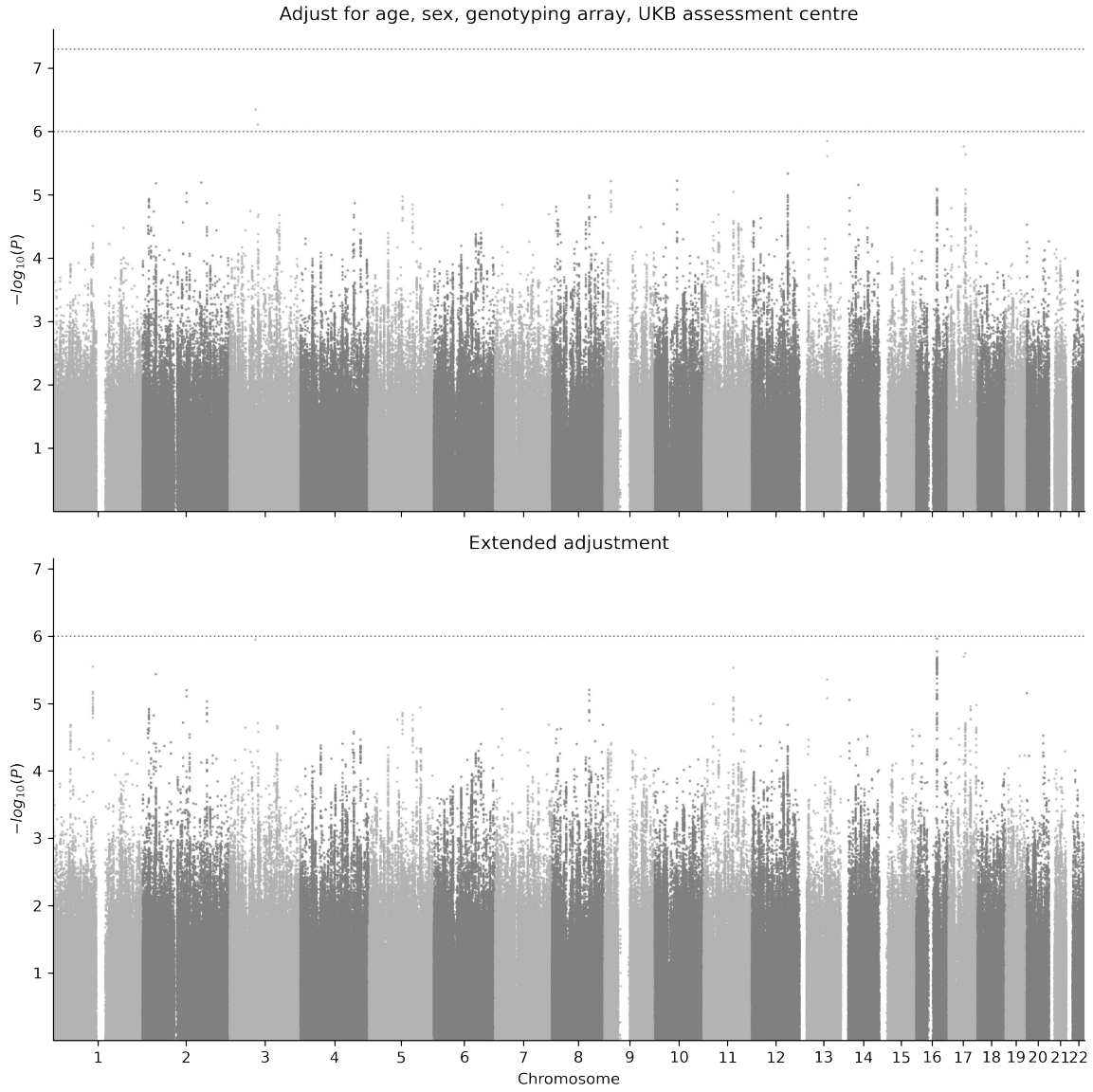
Supplementary Figures



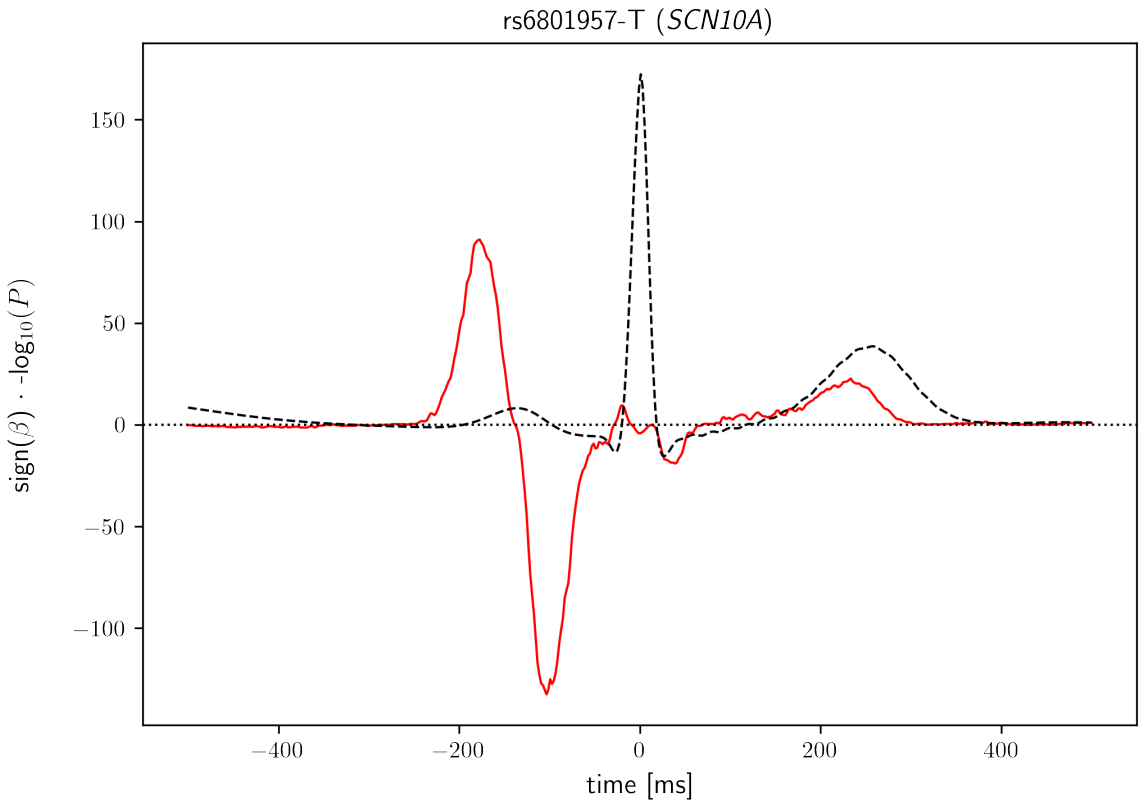
Supplementary Figure 1: Age predicted from 12-lead ECGs by a deep convolutional neural network vs. chronological age for 36,347 participants of the UKB. Predictions from "abnormal" ECG s (as automatically determined by the ECG machine) are shown in red; predictions from other ECGs in blue.



Supplementary Figure 2: Manhattan plots for contrasting the GWAS results using "regular" adjustment (age, sex, genotyping array, assessment centre) and "extended" adjustment (see Methods section for a full list of covariates) of the whole dataset as well as a subset including only participants with an ethnic background of "White British".



Supplementary Figure 3: Manhattan plots of the association tests for "regular" (age, sex, genotyping array, UKB assessment centre) and extended (see Methods for the full list of covariates) adjustment with abnormal vs. normal ECG as binary phenotype. Dotted lines show the thresholds for genome-wide and suggestive significance ($p = 5 \times 10^{-8}$ and $p = 1 \times 10^{-6}$, respectively).



Supplementary Figure 4: Impact of the lead variant in *SCN10A* (rs6801957) on ECG morphology as determined by Verweij *et al.* (red line) [16]. The y-axis represents the signed logarithm of the P-values for each of the 500 sections into which the averaged ECG signal was divided. The dotted black line depicts the mean ECG from [16] for reference with the P-wave at -150 ms, the QRS complex at around 0 ms, and the T-wave at +300 ms.

Supplementary Tables

Supplementary Table 1: Association of human-defined ECG features with chronological and ECG-derived age.

		Univariate		Multivariate (without age)	
		Effect size	*P*-value	Effect size	*P*-value
Ventricular rate	Age	0.49 (0.41, 0.57)	**2.1e-34**	0.24 (0.16, 0.32)	**3.6e-9**
	Pred. age	0.51 (0.42, 0.60)	**2.0e-27**	-0.06 (-0.16, 0.03)	0.19
P duration	Age	0.31 (0.23, 0.39)	**6.2e-14**	0.19 (0.12, 0.27)	**1.0e-6**
	Pred. age	0.55 (0.46, 0.64)	**1.1e-31**	0.49 (0.40, 0.58)	**7.7e-26**
PP interval	Age	-0.57 (-0.66, -0.47)	**5.1e-30**	-0.27 (-0.37, -0.17)	**4.7e-8**
	Pred. age	-0.63 (-0.75, -0.52)	**6.3e-27**	-0.03 (-0.15, 0.09)	0.64
PQ interval	Age	1.27 (1.17, 1.36)	**6.2e-144**	1.07 (0.97, 1.16)	**9.7e-108**
	Pred. age	2.36 (2.25, 2.47)	**0.0**	2.32 (2.21, 2.43)	**0.0**
QRS num	Age	0.39 (0.29, 0.49)	**1.1e-14**	0.11 (0.01, 0.21)	0.025
	Pred. age	0.37 (0.25, 0.48)	**6.4e-10**	-0.23 (-0.36, -0.11)	**1.5e-4**
QRS duration	Age	0.68 (0.60, 0.76)	**2.7e-63**	0.41 (0.33, 0.49)	**9.3e-24**
	Pred. age	0.61 (0.52, 0.71)	**1.5e-39**	0.46 (0.36, 0.55)	**9.4e-21**
QT interval	Age	0.13 (0.03, 0.23)	9.8e-3	0.34 (0.24, 0.43)	**7.2e-12**
	Pred. age	0.51 (0.39, 0.63)	**1.2e-17**	0.87 (0.75, 0.99)	**2.9e-47**
QTC interval	Age	0.96 (0.86, 1.07)	**6.0e-73**	0.81 (0.71, 0.92)	**2.2e-52**
	Pred. age	1.57 (1.45, 1.70)	**4.8e-138**	1.14 (1.01, 1.26)	**1.1e-67**
RR interval	Age	-0.50 (-0.60, -0.41)	**1.9e-24**	-0.22 (-0.32, -0.12)	**1.2e-5**
	Pred. age	-0.55 (-0.67, -0.44)	**3.5e-21**	0.08 (-0.04, 0.20)	0.20
P axis	Age	0.62 (0.52, 0.71)	**6.4e-35**	0.61 (0.52, 0.71)	**8.5e-37**
	Pred. age	0.89 (0.78, 1.00)	**8.9e-53**	1.11 (0.99, 1.22)	**1.6e-80**
R axis	Age	-1.21 (-1.30, -1.11)	**4.2e-137**	-0.95 (-1.05, -0.86)	**1.8e-86**
	Pred. age	-3.19 (-3.29, -3.08)	**0.0**	-2.87 (-2.98, -2.76)	**0.0**
T axis	Age	0.33 (0.23, 0.42)	**2.4e-11**	0.31 (0.22, 0.40)	**1.1e-10**
	Pred. age	-0.41 (-0.53, -0.30)	**1.4e-12**	-0.30 (-0.41, -0.18)	**3.4e-7**

To allow for comparable effect sizes, the ECG features were divided by their standard deviations prior to running the regression tests. Linear models used to generate the results shown in the 'Multivariate' column included all covariates listed for "extended adjustment" in the Methods section except for age. Values in parentheses denote the lower and upper bounds of the 95% confidence interval. *P*-values smaller than the Bonferroni-corrected threshold ($0.05/24 = 0.0021$) are highlighted in bold.

Supplementary Table 2: Tissue enrichment analysis results generated by DEPICT using a GWAS P-value threshold of 10^{-6} (showing all entries with nominal P-values $\leq$ 0.05).

MeSH term	Name	MeSH first level term	Nominal P value	False discovery rate
A10.165.450.300	Cicatrix	Tissues	0.02	≥ 0.20
A10.165.450	Granulation Tissue	Tissues	0.02	≥ 0.20
A07.541	Heart	Cardiovascular System	0.03	≥ 0.20
A07.541.560	Heart Ventricles	Cardiovascular System	0.04	≥ 0.20
A07.231.114	Arteries	Cardiovascular System	0.04	≥ 0.20

MeSH, Medical Subject Heading

Supplementary Table 3: Tissue enrichment analysis results generated by DEPICT using a GWAS P-value threshold of 10^{-4} (showing all entries with nominal P-values ≤ 0.05).

MeSH term	Name	MeSH first level term	Nominal P value	False discovery rate
A07.541.358	Heart Atria	Cardiovascular System	3.6e-05	<0.01
A07.541.358.100	Atrial Appendage	Cardiovascular System	4.5e-05	<0.01
A07.541	Heart	Cardiovascular System	7.5e-05	<0.01
A07.541.560	Heart Ventricles	Cardiovascular System	2.6e-04	<0.01
A10.690	Muscles	Tissues	3.8e-03	<0.20
A05.360.319.679.690	Myometrium	Urogenital System	4.9e-03	<0.20
A11.329.830	Stromal Cells	Cells	5.8e-03	<0.20
A11.329.228	Fibroblasts	Cells	6.1e-03	<0.20
A10.690.552.500	Muscle Skeletal	Tissues	6.7e-03	<0.20
A10.690.552	Muscle Striated	Tissues	6.7e-03	<0.20
A07.231.114	Arteries	Cardiovascular System	7.3e-03	<0.20
A02.633.567.850	Quadriceps Muscle	Musculoskeletal System	8.3e-03	<0.20
A07.541.510.110	Aortic Valve	Cardiovascular System	8.6e-03	<0.20
A07.541.510	Heart Valves	Cardiovascular System	8.6e-03	<0.20
A11.329	Connective Tissue Cells	Cells	9.8e-03	<0.20
A11.329.629	Osteoblasts	Cells	0.02	≥ 0.20
A10.690.467	Muscle Smooth	Tissues	0.03	≥ 0.20
A06.407.071.140	Adrenal Cortex	Endocrine System	0.03	≥ 0.20
A10.165.450.300.425	Keloid	Tissues	0.04	≥ 0.20
A06.407.071	Adrenal Glands	Endocrine System	0.04	≥ 0.20
A11.872.580	Mesenchymal Stem Cells	Cells	0.04	≥ 0.20
A02.835.583.443	Joint Capsule	Musculoskeletal System	0.05	≥ 0.20
A02.835.583	Joints	Musculoskeletal System	0.05	≥ 0.20
A02.835.583.443.800	Synovial Membrane	Musculoskeletal System	0.05	≥ 0.20
A10.165.450.300	Cicatrix	Tissues	0.05	≥ 0.20
A10.165.450	Granulation Tissue	Tissues	0.05	≥ 0.20

MeSH, Medical Subject Heading

Supplementary Table 4: Results of gProfiler [22] including all loci found in the GWAS with $p \leq 1 \times 10^{-6}$.

Source	Name	ID	*P*-value
GO:MF	cytoskeletal protein binding	GO:0008092	8.0e-04
	calmodulin binding	GO:0005516	1.8e-03
	[heparan sulfate]-glucosamine N-sulfotransferase activity	GO:0015016	5.2e-03
	actin binding	GO:0003779	6.1e-03
GO:BP	heparin biosynthetic process	GO:0030210	6.9e-04
	heart contraction	GO:0060047	7.8e-04
	heart process	GO:0003015	9.9e-04
	cardiac muscle contraction	GO:0060048	1.1e-03
	muscle system process	GO:0003012	1.3e-03
	cell communication involved in cardiac conduction	GO:0086065	1.4e-03
	bundle of His cell to Purkinje myocyte communication	GO:0086069	1.4e-03
	heparin metabolic process	GO:0030202	1.7e-03
	cardiac muscle cell action potential	GO:0086001	3.5e-03
	regulation of cardiac muscle contraction	GO:0055117	3.9e-03
	striated muscle contraction	GO:0006941	4.3e-03
	muscle contraction	GO:0006936	6.3e-03
	heparan sulfate proteoglycan biosynthetic process	GO:0015012	8.0e-03
	regulation of striated muscle contraction	GO:0006942	9.7e-03
	cardiac conduction	GO:0061337	1.1e-02
	regulation of heart rate	GO:0002027	1.2e-02
	heparan sulfate proteoglycan metabolic process	GO:0030201	1.5e-02
	regulation of muscle system process	GO:0090257	1.6e-02
	regulation of cardiac muscle cell contraction	GO:0086004	1.7e-02
	regulation of actin filament-based movement	GO:1903115	2.8e-02
	regulation of heart rate by cardiac conduction	GO:0086091	2.8e-02
	regulation of membrane depolarization	GO:0003254	3.2e-02
	action potential	GO:0001508	4.5e-02
GO:CC	sarcomere	GO:0030017	1.4e-03
	myofibril	GO:0030016	2.3e-03
	contractile fiber	GO:0043292	2.8e-03
	calcium- and calmodulin-dependent protein kinase complex	GO:0005954	3.3e-03
	Z disc	GO:0030018	4.6e-03
	I band	GO:0031674	6.7e-03
	voltage-gated sodium channel complex	GO:0001518	4.5e-02
KEGG	Melanogenesis	KEGG:04916	1.7e-02
	Glycosaminoglycan biosynthesis - heparan sulfate / heparin	KEGG:00534	2.6e-02
REAC	Phase 0 - rapid depolarisation	REAC:R-HSA-5576892	5.5e-05
	Muscle contraction	REAC:R-HSA-397014	3.8e-03
	Interaction between L1 and Ankyrins	REAC:R-HSA-445095	5.0e-03
	Cardiac conduction	REAC:R-HSA-5576891	1.4e-02
	CaMK IV-mediated phosphorylation of CREB	REAC:R-HSA-111932	3.4e-02
HP	Ventricular tachycardia	HP:0004756	1.3e-02
	Abnormality of synovial bursa morphology	HP:0025231	4.6e-02
	Bursitis	HP:0025232	4.6e-02

GO, Gene Ontology; MF, Molecular Function; BP, Biological Process; CC, Cellular Component; REAC, Reactome Pathway Database; HP, Human Phenotype Ontology

Supplementary Table 5: Results of gProfiler [22] including all loci with $p \leq 1 \times 10^{-4}$.

Source	Name	ID	*P*-value
GO:MF	cytoskeletal protein binding	GO:0008092	6.4e-05
	transmembrane transporter binding	GO:0044325	5.9e-04
	actin binding	GO:0003779	1.1e-03
	actin filament binding	GO:0051015	1.5e-02
	heparan sulfate sulfotransferase activity	GO:0034483	2.0e-02
GO:CC	sarcoplasm	GO:0016528	2.7e-04
	sarcomere	GO:0030017	2.8e-04
	sarcoplasmic reticulum membrane	GO:0033017	3.0e-04
	Z disc	GO:0030018	3.6e-04
	sarcolemma	GO:0042383	5.7e-04
	myofibril	GO:0030016	8.7e-04
	I band	GO:0031674	9.4e-04
	cytoskeleton	GO:0005856	1.3e-03
	sarcoplasmic reticulum	GO:0016529	1.3e-03
	contractile fiber	GO:0043292	1.4e-03
	cytoplasm	GO:0005737	1.9e-03
	cell junction	GO:0030054	1.6e-02
	supramolecular fiber	GO:0099512	4.2e-02
	supramolecular polymer	GO:0099081	4.8e-02
	calcium ion-transporting ATPase complex	GO:0090534	5.0e-02
KEGG	Adrenergic signaling in cardiomyocytes	KEGG:04261	3.9e-06
	Dilated cardiomyopathy	KEGG:05414	1.0e-03
	Oxytocin signaling pathway	KEGG:04921	8.8e-03
	cAMP signaling pathway	KEGG:04024	9.5e-03
	Arrhythmogenic right ventricular cardiomyopathy	KEGG:05412	1.3e-02
	Cardiac muscle contraction	KEGG:04260	2.8e-02
	Hypertrophic cardiomyopathy	KEGG:05410	3.4e-02
	Amphetamine addiction	KEGG:05031	5.0e-02
REAC	Cardiac conduction	REAC:R-HSA-5576891	4.2e-02
WP	Calcium regulation in cardiac cells	WP:WP536	4.2e-02
HP	Heart block	HP:0012722	1.6e-04
	Ventricular tachycardia	HP:0004756	1.7e-04
	Supraventricular arrhythmia	HP:0005115	8.5e-04
	Cardiac conduction abnormality	HP:0031546	2.3e-03
	Ventricular arrhythmia	HP:0004308	2.6e-03
	Sudden death	HP:0001699	2.8e-03
	Sudden cardiac death	HP:0001645	5.7e-03
	Autosomal dominant inheritance	HP:0000006	1.2e-02
	Dilated cardiomyopathy	HP:0001644	1.3e-02
	Cardiac arrest	HP:0001695	1.6e-02
	Atrioventricular block	HP:0001678	1.8e-02
	Sick sinus syndrome	HP:0011704	1.9e-02
	Reduced systolic function	HP:0006673	2.0e-02
	Abnormal atrioventricular conduction	HP:0005150	3.1e-02

GO:BP and CORUM were skipped for space reasons. GO, Gene Ontology; MF, Molecular Function; BP, Biological Process; CC, Cellular Component; KEGG, Kyoto Encyclopedia of Genes and Genomes; REAC, Reactome Pathway Database; WP, WikiPathways; HP, Human Phenotype Ontology

Supplementary Table 6: Loci found by our GWAS analyses to be associated with delta age and corresponding P-values of association with ECG morphology as determined by Verweij *et al.* [16].

Chr.	Gene	rsID	Pos.	Ref.	Alt.	AF	Type	Effect size	P-value (assoc. with delta age)	P-value (assoc. with ECG morph.)
2	*SPTBN1*	rs1802889	54,756,740	C	T	0.68	regular	-0.30 (-0.42, -0.19)	4.4e-07*	**2.3e-13**
							extended	-0.32 (-0.43, -0.20)	2.4e-07*	
							white British	-0.34 (-0.46, -0.21)	7.5e-08*	
	TTN	rs12464215	179,690,968	G	T	0.46	regular	0.31 (0.20, 0.42)	**2.9e-08**	**8.0e-25**
							extended	0.33 (0.22, 0.44)	**6.5e-09**	
							white British	0.31 (0.20, 0.42)	9.2e-08*	
3	*SCN5A*	rs6773331	38,684,397	A	T	0.98	regular	1.24 (0.82, 1.66)	**9.1e-09**	**5.7e-71**
							extended	1.24 (0.81, 1.67)	**1.3e-08**	
							white British	1.22 (0.79, 1.66)	**3.4e-08**	
	SCN10A	rs6801957	38,767,315	T	C	0.59	regular	-0.32 (-0.43, -0.21)	**2.1e-08**	**2.8e-133**
							extended	-0.29 (-0.40, -0.17)	6.3e-07*	
							white British	-0.32 (-0.44, -0.21)	**4.3e-08**	
4	*CAMK2D*	rs35430511	114,387,138	T	C	0.26	regular	0.49 (0.36, 0.61)	**3.1e-14**	**2.9e-22**
							extended	0.48 (0.35, 0.61)	**1.7e-13**	
							white British	0.48 (0.35, 0.61)	**2.9e-13**	
5	*RGMB*	rs59839451	97,514,975	G	T	0.02	regular	-0.92 (-1.31, -0.53)	3.6e-06	7.8e-03
							extended	-1.03 (-1.42, -0.63)	3.7e-07*	
							white British	-0.87 (-1.27, -0.46)	2.5e-05	
	PKD2L2	rs10076361	137,252,940	G	A	0.18	regular	0.41 (0.27, 0.55)	**2.3e-08**	1.6e-03
							extended	0.40 (0.25, 0.54)	1.1e-07*	
							white British	0.41 (0.26, 0.56)	5.7e-08*	
6	*VGLL2*	rs6901720	117,510,203	G	T	0.47	regular	0.43 (0.32, 0.54)	**2.8e-14**	**1.0e-10**
							extended	0.42 (0.31, 0.53)	**2.9e-13**	
							white British	0.44 (0.33, 0.55)	**4.5e-14**	
8	*DEFB136*	rs4240678	11,802,426	C	T	0.40	regular	0.47 (0.32, 0.62)	**4.9e-10**	**1.7e-18**
							extended	0.49 (0.34, 0.64)	**1.2e-10**	
							white British	0.44 (0.29, 0.59)	**1.5e-08**	
	EXT1	rs57237854	118,860,126	ATCTTG	A	0.18	regular	0.40 (0.25, 0.54)	5.3e-08*	5.9e-07
							extended	0.39 (0.25, 0.54)	1.3e-07*	
							white British	0.41 (0.26, 0.56)	5.1e-08*	
	ZC3H3	rs2294117	144,520,147	G	A	0.25	regular	0.31 (0.19, 0.44)	1.2e-06	1.2e-02
							extended	0.30 (0.18, 0.43)	3.6e-06	
							white British	0.33 (0.20, 0.46)	7.5e-07*	
10	*CTNNA3*	rs72799115	68,008,504	G	A	0.21	regular	0.35 (0.22, 0.49)	2.0e-07*	8.9e-07
							extended	0.36 (0.22, 0.49)	2.1e-07*	
							white British	0.36 (0.22, 0.50)	4.3e-07*	
	AGAP5	rs147790633	75,447,582	T	C	0.14	regular	-0.43 (-0.59, -0.27)	8.7e-08*	**3.5e-24**
							extended	-0.46 (-0.62, -0.30)	**1.7e-08**	
							white British	-0.40 (-0.57, -0.24)	1.4e-06	
12	*SOX5*	rs12826024	24,776,799	G	A	0.15	regular	-0.39 (-0.54, -0.24)	6.1e-07*	**1.7e-21**
							extended	-0.37 (-0.53, -0.22)	2.5e-06	
							white British	-0.38 (-0.54, -0.22)	2.4e-06	
	TBX3	rs1896329	115,357,432	C	T	0.69	regular	-0.31 (-0.42, -0.19)	3.9e-07*	**6.2e-110**
							extended	-0.28 (-0.40, -0.16)	5.4e-06	
							white British	-0.32 (-0.45, -0.20)	2.7e-07*	
14	*SIPA1L1*	rs35866366	71,849,185	A	G	0.25	regular	0.52 (0.39, 0.64)	**1.1e-15**	**2.2e-13**
							extended	0.51 (0.38, 0.63)	**9.3e-15**	
							white British	0.48 (0.35, 0.61)	**6.6e-13**	
16	*CHD9*	rs75778953	52,906,677	C	T	0.01	regular	-1.25 (-1.74, -0.76)	6.2e-07*	2.3e-03
							extended	-1.25 (-1.75, -0.75)	8.1e-07*	
							white British	-1.19 (-1.70, -0.69)	3.4e-06	

This table lists all loci and lead variants which were associated with delta age with at least suggestive significance ($p \leq 1 \times 10^{-6}$) in at least one GWAS run (i.e. with regular adjustment, extended adjustment, or regular adjustment including only participants with White British ethnic background; regular adjustment includes age, sex, genotyping array, and UKB assessment centre; see Methods for a list of covariates included in extended adjustment). For some loci, not all GWAS runs featured the same lead variant (i.e. variant with the smallest P-value). In such cases, a representative variant ranking high in all three analyses was chosen to be listed in the table. Positions correspond to the GRCh37 human genome assembly. Values in parentheses denote the lower and upper bounds of the 95% confidence interval of the effect size estimate. P-values for association with ECG morphology were retrieved from `www.ecgenetics.org` [16]. P-values with genome-wide significance ($p \leq 5 \times 10^{-8}$ for our GWAS results and $p \leq 1 \times 10^{-10}$ for the results of Verweij *et al* [16] due to multiple testing) are highlighted in bold. P-values with suggestive significance in the GWAS ($p \leq 1 \times 10^{-6}$) are highlighted with an asterisk. Chr., Chromosome; Pos., Position; Ref., Reference allele; Alt., Alternative allele; AF, frequency of the alternative allele; assoc., associated; morph., morphology.

Supplementary Table 7: Association of DOSI and excess DOSI with delta age.

Metric	Univariate		Multivariate (age, sex, assessment centre)		Multivariate (full)	
	Effect size	*P*-value	Effect size	*P*-value	Effect size	*P*-value
DOSI	-0.52 (-0.65, -0.38)	1.7e-13	0.07 (-0.06, 0.21)	0.28	0.20 (0.07, 0.34)	3.6e-03
eDOSI	0.34 (0.20, 0.49)	2.9e-06	0.09 (-0.04, 0.23)	0.17	0.22 (0.09, 0.36)	1.4e-03

Values in the parentheses denote the lower and upper bounds of the 95% confidence intervals of the effect size estimate.

Supplementary Table 8: Loci associated with "abnormal" ECGs (as automatically labelled by the ECG machine).

Chr.	Gene	rsID	Pos.	Ref.	Alt.	AF	Adj.	LOR	*P*-value
3	*RYBP*	rs9883587	72,383,495	G	A	0.76	basic	0.09 (0.06, 0.13)	4.5e-07*
							full	0.09 (0.06, 0.13)	1.1e-06
	ROBO1	rs115880984	78,675,814	G	A	0.02	basic	-0.28 (-0.40, -0.17)	7.8e-07*
							full	-0.25 (-0.37, -0.14)	2.0e-05

Two GWAS were run with abnormal vs. normal ECGs as binary phenotype. One adjusted for age, sex, genotyping array, and UKB assessment centre, whereas the other one included more covariates for extended adjustment (see Methods for a full list of covariates). Positions correspond to the GRCh37 human genome assembly. Values in parentheses denote the lower and upper bounds of the 95% confidence interval of the effect size estimate. *P*-values with suggestive significance are highlighted with an asterisk. Chr., Chromosome; Pos., Position; Ref., Reference allele; Alt., Alternative allele; AF, frequency of the alternative allele; Adj., Adjustment; LOR, Log odds ratio;

Supplementary Data

Supplementary Data 1 DEPICT geneset enrichment analysis with a P-value threshold of 1×10^{-6}.

Supplementary Data 2 DEPICT geneset enrichment analysis with a P-value threshold of 1×10^{-4}.

Supplementary Data 3 Smallest P-values of variants found by our analysis (GWAS with regular adjustment) in the vicinity (within 1 Mb) of loci associated with parental longevity [23] and leukocyte telomere lengths [24].

Supplementary Data 4 Results of mediation analyses for BMI, MAP, and diagnosed hypertension as individual mediators were determined with the Pingouin Python package [15] and with 5,000 bootstrap iterations per test.

References for Supplementary Information

1. Sarrazin, S., Lamanna, W. C. & Esko, J. D. Heparan sulfate proteoglycans. *Cold Spring Harbor perspectives in biology* **3,** a004952 (2011).
2. Poulain, F. E. & Yost, H. J. Heparan sulfate proteoglycans: a sugar code for vertebrate development? *Development* **142,** 3456–3467 (2015).
3. Zhang, R. *et al.* Heparan sulfate biosynthesis enzyme, Ext1, contributes to outflow tract development of mouse heart via modulation of FGF signaling. *PLoS One* **10,** e0136518 (2015).
4. Vite, A., Li, J. & Radice, G. L. New functions for alpha-catenins in health and disease: from cancer to heart regeneration. *Cell and tissue research* **360,** 773–783 (2015).
5. Frey, N. *et al.* Calsarcin-2 deficiency increases exercise capacity in mice through calcineurin/NFAT activation. *The Journal of clinical investigation* **118,** 3598–3608 (2008).
6. Martin, R. I. *et al.* Genetic variants associated with risk of atrial fibrillation regulate expression of PITX2, CAV1, MYOZ1, C9orf3 and FANCC. *Journal of molecular and cellular cardiology* **85,** 207–214 (2015).
7. Shah, S. *et al.* Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nature communications* **11,** 1–12 (2020).
8. Clausen, A. G., Vad, O. B., Andersen, J. H. & Olesen, M. S. Loss-of-Function Variants in the SYNPO2L Gene Are Associated With Atrial Fibrillation. *Frontiers in cardiovascular medicine* **8** (2021).
9. Ellinor, P. T. *et al.* Meta-analysis identifies six new susceptibility loci for atrial fibrillation. *Nature genetics* **44,** 670–675 (2012).
10. Li, A. *et al.* Silencing of the Drosophila ortholog of SOX5 in heart leads to cardiac dysfunction as detected by optical coherence tomography. *Human molecular genetics* **22,** 3798–3806 (2013).
11. Liu, Y. *et al.* High expression levels and localization of Sox5 in dilated cardiomyopathy. *Molecular Medicine Reports* **22,** 948–956 (2020).
12. Washkowitz, A. J., Gavrilov, S., Begum, S. & Papaioannou, V. E. Diverse functional networks of Tbx3 in development and disease. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **4,** 273–283 (2012).
13. Van Eif, V. W. *et al.* Genome-wide analysis identifies an essential human TBX3 pacemaker enhancer. *Circulation research* **127,** 1522–1535 (2020).
14. Yang, P. *et al.* βII spectrin (SPTBN1): biological function and clinical potential in cancer and other diseases. *International journal of biological sciences* **17,** 32 (2021).
15. Vallat, R. Pingouin: statistics in Python. *J. Open Source Softw.* **3,** 1026 (2018).
16. Verweij, N. *et al.* The genetic makeup of the electrocardiogram. *Cell systems* **11,** 229–238 (2020).
17. DAgostino Sr, R. B. *et al.* General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* **117,** 743–753 (2008).
18. National Health Service. *What's your heart age?* `https://www.nhs.uk/conditions/nhs-health-check/check-your-heart-age-tool/` (2022).
19. Hippisley-Cox, J., Coupland, C., Robson, J. & Brindle, P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. *Bmj* **341** (2010).
20. Pyrkov, T. V. *et al.* Longitudinal analysis of blood markers reveals progressive loss of resilience and predicts human lifespan limit. *Nature communications* **12,** 1–10 (2021).
21. Wang, M., Gao, J., Shi, Y. & Zhao, X. A genome-wide association and polygenic risk score study on abnormal electrocardiogram in a Chinese population. *Scientific reports* **11,** 1–11 (2021).
22. Raudvere, U. *et al.* g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research* **47,** W191–W198 (2019).

718 23. Pilling, L. C. *et al.* Human longevity: 25 genetic loci associated in 389,166 UK biobank partici-
719 pants. *Aging (Albany NY)* **9,** 2504 (2017).
720 24. Codd, V. *et al.* Polygenic basis and biomedical consequences of telomere length variation. *Nature*
721 *genetics* **53,** 1425–1433 (2021).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- suppdata.tar.gz