University of Groningen

Genetics of myocardial interstitial fibrosis in the human heart and association with disease

Nauffal, Victor; Di Achille, Paolo; Klarqvist, Marcus D.R.; Cunningham, Jonathan W.; Hill, Matthew C.; Pirruccello, James P.; Weng, Lu Chen; Morrill, Valerie N.; Choi, Seung Hoan; Khurshid, Shaan

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

[Link to publication in University of Groningen/UMCG research database](#)

Genetics of myocardial interstitial fibrosis in the human heart and association with disease

Victor Nauffal [icon] [1,2,11], Paolo Di Achille [icon] [3,11], Marcus D. R. Klarqvist [icon] [3,11], Jonathan W. Cunningham [icon] [1,2,11], Matthew C. Hill [2,4,11], James P. Pirruccello [icon] [2,5,6], Lu-Chen Weng [icon] [2,4], Valerie N. Morrill [2], Seung Hoan Choi [icon] [2], Shaan Khurshid [2,7], Samuel F. Friedman [icon] [3], Mahan Nekoui [2], Carolina Roselli [icon] [2,8], Kenney Ng [9], Anthony A. Philippakis [icon] [2,3,10], Puneet Batra [icon] [3], Patrick T. Ellinor [icon] [2,7,12] [envelope] & Steven A. Lubitz [icon] [2,7,12] [envelope]

Myocardial interstitial fibrosis is associated with cardiovascular disease and adverse prognosis. Here, to investigate the biological pathways that underlie fibrosis in the human heart, we developed a machine learning model to measure native myocardial T1 time, a marker of myocardial fibrosis, in 41,505 UK Biobank participants who underwent cardiac magnetic resonance imaging. Greater T1 time was associated with diabetes mellitus, renal disease, aortic stenosis, cardiomyopathy, heart failure, atrial fibrillation, conduction disease and rheumatoid arthritis. Genome-wide association analysis identified 11 independent loci associated with T1 time. The identified loci implicated genes involved in glucose transport (*SLC2A12*), iron homeostasis (*HFE*, *TMPRSS6*), tissue repair (*ADAMTSL1*, *VEGFC*), oxidative stress (*SOD2*), cardiac hypertrophy (*MYH7B*) and calcium signaling (*CAMK2D*). Using a transforming growth factor β1-mediated cardiac fibroblast activation assay, we found that 9 of the 11 loci consisted of genes that exhibited temporal changes in expression or open chromatin conformation supporting their biological relevance to myofibroblast cell state acquisition. By harnessing machine learning to perform large-scale quantification of myocardial interstitial fibrosis using cardiac imaging, we validate associations between cardiac fibrosis and disease, and identify new biologically relevant pathways underlying fibrosis.

The cardiac extracellular matrix (ECM) is a dynamic compartment that has key structural and regulatory roles in establishing myocardial tissue architecture and function. Pathological perturbations to homeostatic turnover of ECM components leads to the progressive development of interstitial fibrosis[1], which is the histological hallmark of several cardiac diseases[2–8]. A myriad of hemodynamic, metabolic and inflammatory stressors contribute to the accelerated development of interstitial fibrosis and associated cardiovascular diseases[9]. There is a critical need to understand the cellular and molecular mechanisms that contribute to pathological cardiac fibrosis in humans, given that their identification would enable the development of targeted antifibrotic therapies applicable across a wide range of cardiovascular diseases. However, progress has been hindered by challenges in reliable noninvasive measurement of interstitial fibrosis at scale in the human heart, and by the
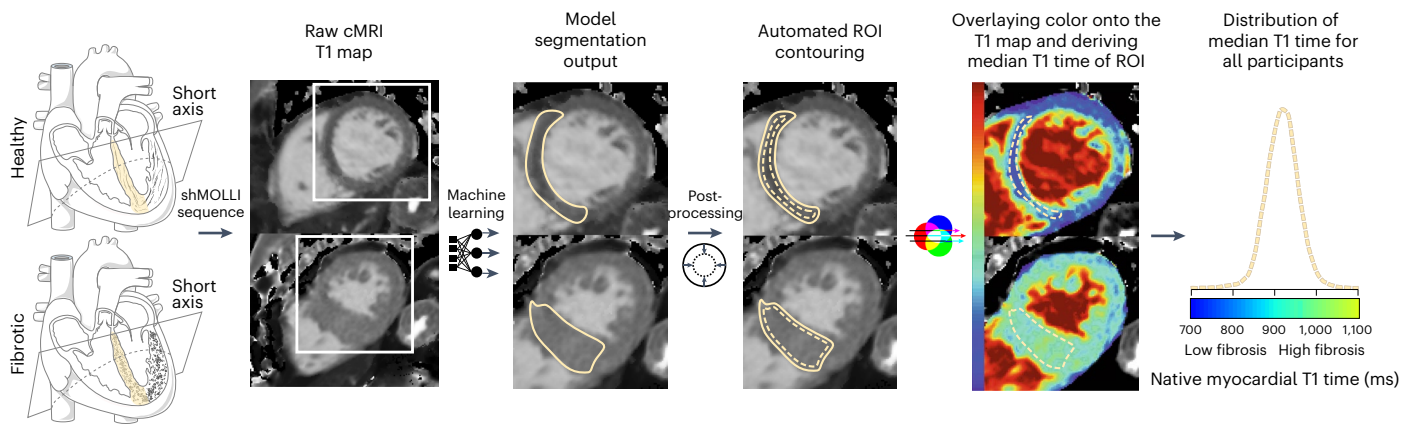
Fig. 1 | Overview of the automated pipeline for native myocardial T1 time measurement at the IVS using machine learning. A representative healthy heart and one with increased interstitial fibrosis are shown for illustration. Cardiac T1 mapping using the shMOLLI recovery sequence was performed at the mid-ventricular short axis. A machine learning model trained on the raw MRI T1 maps generated automated segmentation of the IVS (solid yellow contour) followed by selection of representative myocardial ROIs (dashed yellow contour) using morphological operations. The T1 map color legends were then used to transform pixel intensities within the ROI into T1 times. For each participant, the median T1 time by ROI was calculated and used as the representative T1 time. The heart schematics were drawn using images from Servier Medical Art, which were further modified. Servier Medical Art by Servier is licensed under a Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/). The T1 maps shown are reproduced by permission of the UKB©.

lack of adequately powered validation studies of findings from animal or in vitro tissue culture models. The advent of machine learning tools capable of generating imaging-based phenotypes at scale and large biorepositories containing deep phenotyping and genomic data offers a unique opportunity to overcome these challenges.

Native myocardial T1 time measured using cardiac magnetic resonance imaging (cMRI) is a histopathologically validated metric for quantifying interstitial fibrosis in the human heart[10,11]. The UK Biobank (UKB) is a large-scale prospective cohort with rich cMRI[12], genomic and clinical outcomes data[13]. We sought to use machine learning to quantify fibrosis in over 40,000 study participants who underwent cMRI T1 mapping, assess associations between fibrosis and clinical outcomes, and identify pathways responsible for cardiac fibrosis in humans using genetic association analyses.

Results

Machine learning enables T1 time measurement at scale
We acquired mid-ventricular short-axis cMRI T1 maps for 42,654 participants in the UKB (Fig. 1). In accordance with contemporary guidelines for the assessment of cardiac T1 mapping[14–16], we developed a machine learning model to auto-segment myocardial regions of interest (ROIs) in the interventricular septum (IVS) and quantify T1 time (Methods). T1 times measured using our machine learning model were highly correlated with manually derived T1 times in an independent validation set ($n = 100$; $r = 0.97$, 95% confidence interval (CI) = 0.95–0.98) (Supplementary Fig. 1). After T1 map image quality control, we retained measured T1 time at the IVS for 41,505 participants who constituted our study sample (Supplementary Fig. 2 and Methods).

The mean age of participants was 64.0 ± 7.7 years and 48.1% were men. The mean native myocardial T1 time of the study sample was 918.1 ± 41.5 ms (Table 1). Our results were consistent with known sex-specific patterns of higher T1 time in women than men (Supplementary Fig. 3)[17,18]. Increasing age was associated with decreased T1 time in females ($\Delta T1(ms)_{/10\,years} = -4.15 \pm 0.64$, $P = 1.9 \times 10^{-36}$) and increased T1 time in males ($\Delta T1(ms)_{/10\,years} = 2.52 \pm 0.34$, $P = 2.3 \times 10^{-13}$) after adjusting for comorbidities and medications (Supplementary Fig. 4a). Subsetting to 'healthy' participants free of cardiovascular or metabolic diseases at the time of the MRI scan did not alter the sex-specific, age-related trends in T1 time (Supplementary Fig. 4b).

Age, sex and body mass index (BMI) contributed to most of the variation in T1 time in the study sample (Supplementary Table 1).

In the sections below, we present findings from multivariable analyses adjusted for demographics, comorbidities and medications relevant to myocardial fibrosis. Additionally, in the supplementary tables, we provide results from models adjusted only for age, sex and BMI for comparison.

T1 time is associated with other cMRI-derived measures
We examined the association of myocardial interstitial fibrosis with cMRI-derived measures of left ventricular (LV) and left atrial (LA) structure and function. Before performing association testing, we rank-based inverse-normal-transformed T1 times. As such, changes in T1 time reported approximate multiples of 1 s.d. of T1 time. In the multivariable analysis, lower LV ejection fraction, higher LV mass and larger LV end-systolic volume were associated with increased T1 time. In the LA, lower ejection fraction and larger end-systolic volume were associated with increased T1 time (Supplementary Table 2). Two-sided $P < 7.1 \times 10^{-3}$ was considered statistically significant.

T1 time is associated with prevalent diseases
We then investigated whether T1 time is associated with a priori selected cardiovascular, metabolic and systemic inflammatory diseases relevant to myocardial fibrosis by comparing T1 times from participants with prevalent disease to 'healthy' participants free of cardiovascular or metabolic disorders at the time of cMRI (Methods, Fig. 2 and Supplementary Fig. 5). Two-sided $P < 3.1 \times 10^{-3}$ was considered statistically significant.

In cardiovascular diseases, higher T1 time was associated with hypertrophic cardiomyopathy ($\Delta T1_{(s.d.)} = 0.74 \pm 0.17$; $P = 1.3 \times 10^{-5}$), dilated cardiomyopathy ($\Delta T1_{(s.d.)} = 0.47 \pm 0.12$; $P = 1.3 \times 10^{-4}$), heart failure ($\Delta T1_{(s.d.)} = 0.41 \pm 0.06$; $P = 1.1 \times 10^{-10}$), atrial fibrillation (AF) ($\Delta T1_{(s.d.)} = 0.21 \pm 0.03$; $P = 7.1 \times 10^{-12}$), atrioventricular (AV) node or distal conduction disease ($\Delta T1_{(s.d.)} = 0.37 \pm 0.06$; $P = 3.7 \times 10^{-11}$) and aortic stenosis ($\Delta T1_{(s.d.)} = 0.35 \pm 0.10$; $P = 7.6 \times 10^{-4}$). Coronary artery disease (CAD) ($\Delta T1_{(s.d.)} = 0.06 \pm 0.03$; $P = 2.1 \times 10^{-2}$) and myocardial infarction (MI) ($\Delta T1_{(s.d.)} = 0.10 \pm 0.04$; $P = 1.8 \times 10^{-2}$) were associated with a trend toward higher T1 time that did not reach the threshold for statistical significance.

T1 time was not associated with hypertension ($\Delta T1_{(s.d.)} = 0.02 \pm 0.01$; $P = 0.18$), which is consistent with previous studies that showed a limited ability of T1 mapping to differentiate between individuals with hypertension and controls except among individuals with concomitant

Table 1 | Study sample characteristics at the time of first visit for cMRI

Baseline characteristic	n (%)
Participants	41,505
Age at MRI, years, mean (s.d.)	64.0 (7.7)
Male	19,956 (48.1)
BMI, kg m^{-2}, mean (s.d.)	26.5 (4.3)
Cigarette smoking[a]	
Never	25,663 (62.4)
Previous	14,009 (34.1)
Current	1,422 (3.5)
Alcohol use[a]	
None	1,948 (5.5)
Light to moderate	19,436 (54.6)
Heavy	14,198 (39.9)
Adequate physical activity[a]	26,612 (73.0)
Hypertension	12,561 (30.3)
T2D	1,409 (3.4)
T1D	162 (0.4)
Hyperlipidemia	433 (1.0)
CKD	344 (0.8)
CAD	2,457 (5.9)
MI	840 (2.0)
Dilated cardiomyopathy	61 (0.1)
Hypertrophic cardiomyopathy	29 (0.1)
Heart failure	278 (0.7)
Aortic stenosis	77 (0.2)
AF	1,197 (2.9)
Ventricular arrhythmia or history of cardiac arrest	118 (0.3)
AV node or distal conduction disease	288 (0.7)
RA	579 (1.4)
Beta blocker	2,235 (5.4)
ACE inhibitor or ARB	5,709 (13.8)
Mineralocorticoid receptor antagonist	78 (0.2)
Statins	8,024 (19.3)
Native myocardial T1 time, ms, mean (s.d.)	918.1 (41.5)
LV mass, g, mean (s.d.)	90.1 (29.6)
LV ejection fraction, %, mean (s.d.)	60 (6)
LV end-systolic volume, ml, mean (s.d.)	58.1 (18.9)
LV end-diastolic volume, ml, mean (s.d.)	143.3 (32.7)
LA end-systolic volume, ml, mean (s.d.)	33.6 (14.1)
LA end-diastolic volume, ml, mean (s.d.)	72.8 (20.2)
LA ejection fraction, %, mean (s.d.)	54.8 (8.6)

Values are presented as the number (percentage) unless otherwise specified. A subset of the study sample had data available for each cMRI parameter: LA end-systolic volume/LA end-diastolic volume/LA ejection fraction (n = 37,234); LV end-systolic volume/LV end-diastolic volume/LV ejection fraction/LV mass (n = 40,869). [a]Data for cigarette smoking, alcohol use and self-reported physical activity were available for 41,094, 35,582 and 36,462 of 41,505 study participants (the percentage reported reflects the proportion of those with non-missing data). Adequate physical activity is defined as 150 min or more of moderate or 75 min or more of vigorous activity or equivalent combination per week.

LV hypertrophy (LVH) (Fig. 2 and Supplementary Fig. 5)[19]. Accordingly, we stratified participants with hypertension based on the presence of concomitant LVH. We found that hypertension was associated with higher T1 times only among participants with concomitant LVH ($\Delta T1_{(s.d.)}$ = 0.15 ± 0.03; P = 4.0 × 10^{-8}) but not among those without LVH ($\Delta T1_{(s.d.)}$ = 4.0 × 10^{-3} ± 0.01; P = 0.75) (Supplementary Fig. 6).

Metabolic disorders, including type 1 diabetes (T1D) ($\Delta T1_{(s.d.)}$ = 0.40 ± 0.07; P = 4.0 × 10^{-8}), type 2 diabetes (T2D) ($\Delta T1_{(s.d.)}$ = 0.27 ± 0.03; P = 7.7 × 10^{-20}), hyperlipidemia ($\Delta T1_{(s.d.)}$ = 0.18 ± 0.05; P = 6.4 × 10^{-5}) and chronic kidney disease (CKD) ($\Delta T1_{(s.d.)}$ = 0.20 ± 0.05; P = 8.6 × 10^{-5}), were associated with higher T1 time (Fig. 2 and Supplementary Fig. 5). Of the systemic inflammatory diseases examined, rheumatoid arthritis (RA) was associated with increased T1 time ($\Delta T1_{(s.d.)}$ = 0.13 ± 0.04; P = 6.1 × 10^{-4}) (Fig. 2 and Supplementary Fig. 5).

We performed a sensitivity analysis comparing disease cases to non-cases (that is, 'healthy' controls and non-cases with other cardiovascular and metabolic diseases), further adjusting for comorbidities and medications, and found overall similar findings (Supplementary Table 3 and Supplementary Fig. 7). Stratification by sex yielded consistent disease associations with T1 time in males and females (Supplementary Fig. 8).

We additionally examined the association of serum biomarkers and electrocardiogram (ECG) intervals relevant to the examined prevalent diseases with T1 time. Concordant with the association of diabetes mellitus with increased T1 time, individuals with hemoglobin A1c (HgbA1c) levels in the pre-diabetes ($\Delta T1_{(s.d)}$ = 0.06 ± 0.02; P = 4.6 × 10^{-4}) and diabetes ($\Delta T1_{(s.d.)}$ = 0.27 ± 0.04; P = 5.0 × 10^{-13}) ranges[20] had higher T1 times compared to those with normal HgbA1c levels. An increase in T1 time associated with impaired renal function was only apparent with moderate-to-severe reductions in estimated glomerular filtration rate[21] (<45 ml min^{-1} per 1.73 m^2; $\Delta T1_{(s.d.)}$ = 0.22 ± 0.07; P = 2.6 × 10^{-3}). Among the examined ECG intervals, a prolonged QRS interval (>120 ms) was associated with increased T1 time ($\Delta T1_{(s.d.)}$ = 0.28 ± 0.03; P = 8.9 × 10^{-19}) (Supplementary Table 4 and Supplementary Fig. 9). A two-sided P < 4.6 × 10^{-3} was used for the biomarker and ECG interval analysis.

T1 time is associated with incident cardiovascular disease

We compared the incidence of cardiovascular disease among individuals in the top 20th percentile of the T1 time distribution to that in the bottom 80th percentile over a median follow-up of 2.54 years (interquartile range = 1.63–3.88 years). We analyzed individual diseases and defined two composite endpoints including major arrhythmia (consisting of incident sustained ventricular arrhythmia, cardiac arrest or implantable cardioverter-defibrillator implantation) and major adverse cardiovascular events (MACE) (consisting of incident MI, heart failure, all-cause mortality, AF or major arrhythmia). Participants in the top 20th percentile of T1 time had a higher risk of incident heart failure (hazard ratio (HR) = 1.66, 99% CI = 1.04–2.67; P = 5.4 × 10^{-3}), AF (HR = 1.62, 99% CI = 1.18–2.23; P = 9.5 × 10^{-5}) and AV node or distal conduction disease (HR = 1.93, 99% CI = 1.32–2.83; P = 8.8 × 10^{-6}) compared to those in the lower 80th percentile. Additionally, the incidence of MACE (HR = 1.74, 99% CI = 1.38–2.20; P = 1.23 × 10^{-9}) was higher among participants in the top 20th percentile of T1 time compared to those in the lower 80th percentile (Fig. 3, Supplementary Table 5 and Supplementary Fig. 10). There was a trend toward a higher incidence of the composite major arrhythmia endpoint (HR = 1.75, 99% CI = 0.87–3.55; P = 4.0 × 10^{-2}) in the top 20th percentile of T1 time that did not reach the threshold for statistical significance (Supplementary Table 5 and Supplementary Fig. 10). A two-sided P < 0.01 was considered statistically significant. A sex-stratified sensitivity analysis showed consistent findings for males and females (Supplementary Fig. 11).

Lifestyle factors are associated with T1 time

We investigated the impact of lifestyle factors on myocardial interstitial fibrosis stratified by sex and adjusted for diseases associated with T1 time. In both males and females, light-to-moderate alcohol use ($\Delta T1_{(s.d.)}$ = − 0.21 ± 0.04, P_{males} = 1.0 × 10^{-9}; $\Delta T1_{(s.d.)}$ = − 0.15 ± 0.03,
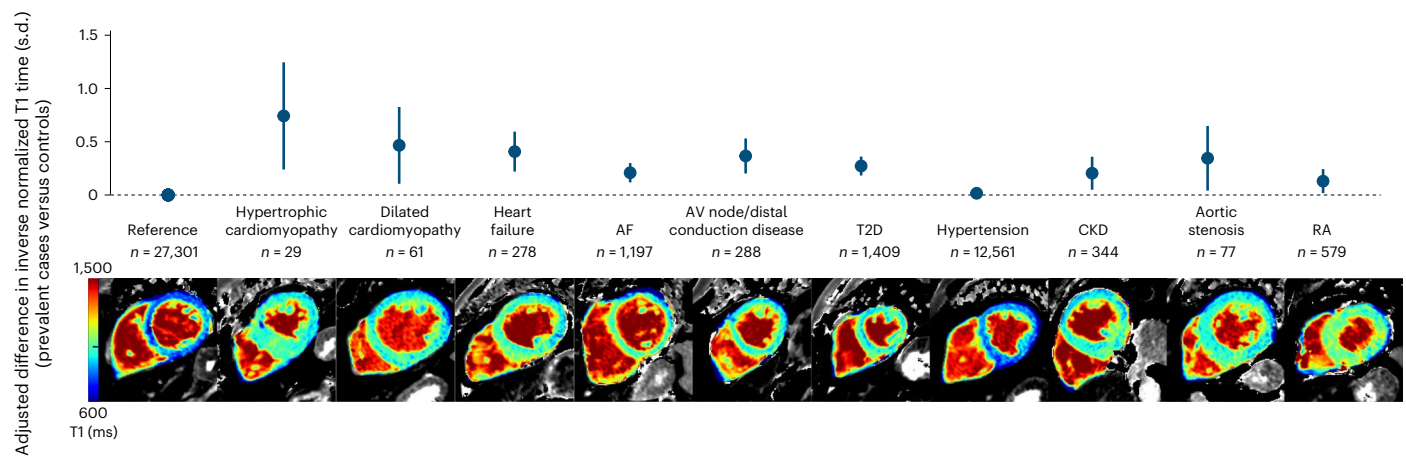
Fig. 2 | Change in native myocardial T1 time associated with prevalent cardiovascular, metabolic and systemic inflammatory diseases compared to healthy controls. Healthy controls free of prevalent dilated cardiomyopathy, hypertrophic cardiomyopathy, heart failure, AF, AV node or distal conduction disease, hypertension, diabetes mellitus, aortic stenosis, CKD, hemochromatosis and RA constituted the reference group. The numbers of controls or cases with available native myocardial T1 time are shown below each category. For each disease, a representative T1 map of a case is provided from the study sample. Multiple linear regression was implemented and a two-sided P value threshold adjusted for multiple testing of $P < 3.1 \times 10^{-3}$ was used to define statistically significant associations. Data are presented as the mean-adjusted change in T1 time along with $(1−\alpha) \times 100$ (%) CIs. CIs were calculated using the adjusted two-sided α (3.1×10^{-3}) for multiple testing. The T1 maps shown are reproduced by permission of the UKB©.

$P_{females} = 6.9 \times 10^{-8}$) and heavy alcohol use ($\Delta T1_{(s.d.)} = - 0.22 \pm 0.04$, $P_{males} = 1.7 \times 10^{-10}$; $\Delta T1_{(s.d.)} = - 0.24 \pm 0.03$, $P_{females} = 5.4 \times 10^{-18}$) were associated with lower T1 time compared to nondrinkers. Current cigarette smoking across both sexes ($\Delta T1_{(s.d.)} = 0.26 \pm 0.03$, $P_{males} = 1.3 \times 10^{-14}$; $\Delta T1_{(s.d.)} = 0.24 \pm 0.04$, $P_{females} = 7.5 \times 10^{-11}$) was associated with increased T1 time compared to nonsmokers. We found a salutary effect of adequate self-reported physical activity defined as meeting guideline recommendations of 150 min or more of moderate or 75 min or more of vigorous activity per week or the equivalent combination on myocardial interstitial fibrosis[22]. Physical activity meeting the guideline recommendations was associated with decreased T1 time in both males and females ($\Delta T1_{(s.d.)} = - 0.07 \pm 0.02$, $P_{males} = 1.6 \times 10^{-6}$; $\Delta T1_{(s.d.)} = - 0.08 \pm 0.02$, $P_{females} = 3.1 \times 10^{-8}$). Among participants with a BMI of < 30 kg m^{-2}, increasing BMI was associated with lower T1 time in both males and females ($\Delta T1_{(s.d.)/kg/m^2} = -0.08 \pm 0.01$, $P_{males} = 1.4 \times 10^{-162}$; $\Delta T1_{(s.d.)/kg/m^2} = -0.06 \pm 0.01$, $P_{females} = 1.0 \times 10^{-162}$). The trend was reversed in obese male participants with a BMI of > 30 kg m^{-2}. Increasing BMI was associated with increased myocardial fibrosis and T1 time in obese males $\Delta T1_{(s.d.)/kg/m^2} = 0.03 \pm 0.01$, $P_{males} = 2.4 \times 10^{-7}$) but was not associated with T1 time in obese females ($\Delta T1_{(s.d.)/kg/m^2} = 4.0 \times 10^{-3} \pm 4.6 \times 10^{-3}$, $P_{females} = 0.35$) (Supplementary Table 6 and Supplementary Fig. 12). A two-sided $P < 1.3 \times 10^{-2}$ in each stratum was considered statistically significant.

Causal effect of T1D on cardiac fibrosis
We explored the causal contribution of cardiovascular risk factors and diseases to myocardial fibrosis using two-sample Mendelian randomization (MR). The inverse variance weighted (IVW) method constituted our primary analysis. We used the MR-Egger method as a sensitivity analysis to account for potential horizontal pleiotropy[23]. We found consistent evidence from the IVW and MR-Egger methods supporting a potential causal effect of T1D ($\Delta T1_{IVW(s.d.)/log\ odds\ of\ genetic\ predisposition\ to\ T1D} = 0.007 \pm 0.002$, $P = 1.0 \times 10^{-3}$; $\Delta T1_{MR\text{-}Egger(s.d.)/log\ odds\ of\ genetic\ predisposition\ to\ T1D} = 0.01 \pm 0.003$, $P = 1.0 \times 10^{-3}$) on myocardial interstitial fibrosis. This was also supported by inspecting the scatter plot of the association of individual genetic variants comprising the T1D genetic instrument with both T1 time and T1D (Supplementary Fig. 13). A causal effect of diastolic blood pressure on myocardial T1 time was suggested by the IVW method, but this association was not supported by the MR-Egger

method or the scatter plot of variant effect sizes. We did not find evidence for a causal effect of BMI, systolic blood pressure, CKD, estimated glomerular filtration rate, T2D, CAD, AF or serum lipoproteins on myocardial fibrosis (Supplementary Table 7). A two-sided $P < 3.9 \times 10^{-3}$ was considered statistically significant.

Genetic association analysis highlights loci relevant to fibrosis
Next, we sought to determine the genetic basis of interstitial fibrosis by performing genetic association analysis of T1 time. The SNP heritability (h_g^2) of T1 time was 0.13, which was lower than that of other cMRI phenotypes in the UKB, such as LV mass (0.26), LV end-diastolic volume (0.40) and LV end-systolic volume (0.31)[24]. Interestingly, in contrast to the strong association of measured T1 time with other cMRI measures of LV and LA structure and function, we observed limited genetic correlation of T1 time with these cMRI measures, suggesting that while shared exposures drive the association between these MRI measures and T1 time, distinct biological pathways contribute to the development of myocardial interstitial fibrosis (Supplementary Fig. 14).

We performed a genome-wide association study (GWAS) and discovered 11 genome-wide-significant loci (Fig. 4a and Supplementary Table 8). There was no evidence of inflation in our GWAS results ($\lambda_{GC} = 1.053$, linkage disequilibrium (LD) score regression intercept = 1.0002) (Supplementary Fig. 15). Regional association plots for genome-wide-significant SNPs are shown in Supplementary Fig. 16.

A regulatory region variant in a promoter-flanking region upstream of the solute carrier *SLC2A12* gene was the most significant lead SNP (rs2627230_T; $P = 8.1 \times 10^{-14}$) and was associated with increased T1 time. *SLC2A12* encodes glucose transporter type 12 (GLUT12), a basal and insulin-independent glucose transporter in the heart[25] with previously reported associations with idiopathic dilated cardiomyopathy in humans[26], as well as insulin resistance[27] and kidney disease[28] in animal models. Lead SNPs near *SOD2* (rs9457699_G; $P = 2.0 \times 10^{-11}$) and *VEGFC* (rs365843_T; $P = 3.2 \times 10^{-9}$), two genes with established roles in cardiac hypertrophy and fibrosis in animal models[29,30], were associated with decreased T1 times. Another lead SNP associated with T1 time is an intronic variant in *ADAMTSL1* (rs1576900_A; $P = 3.6 \times 10^{-11}$), a gene encoding an ADAMTS-like protein thought to modulate the function of ADAMTS metalloproteinases with integral roles in ECM turnover[31,32]. Additionally, rs6120777_A, an intronic
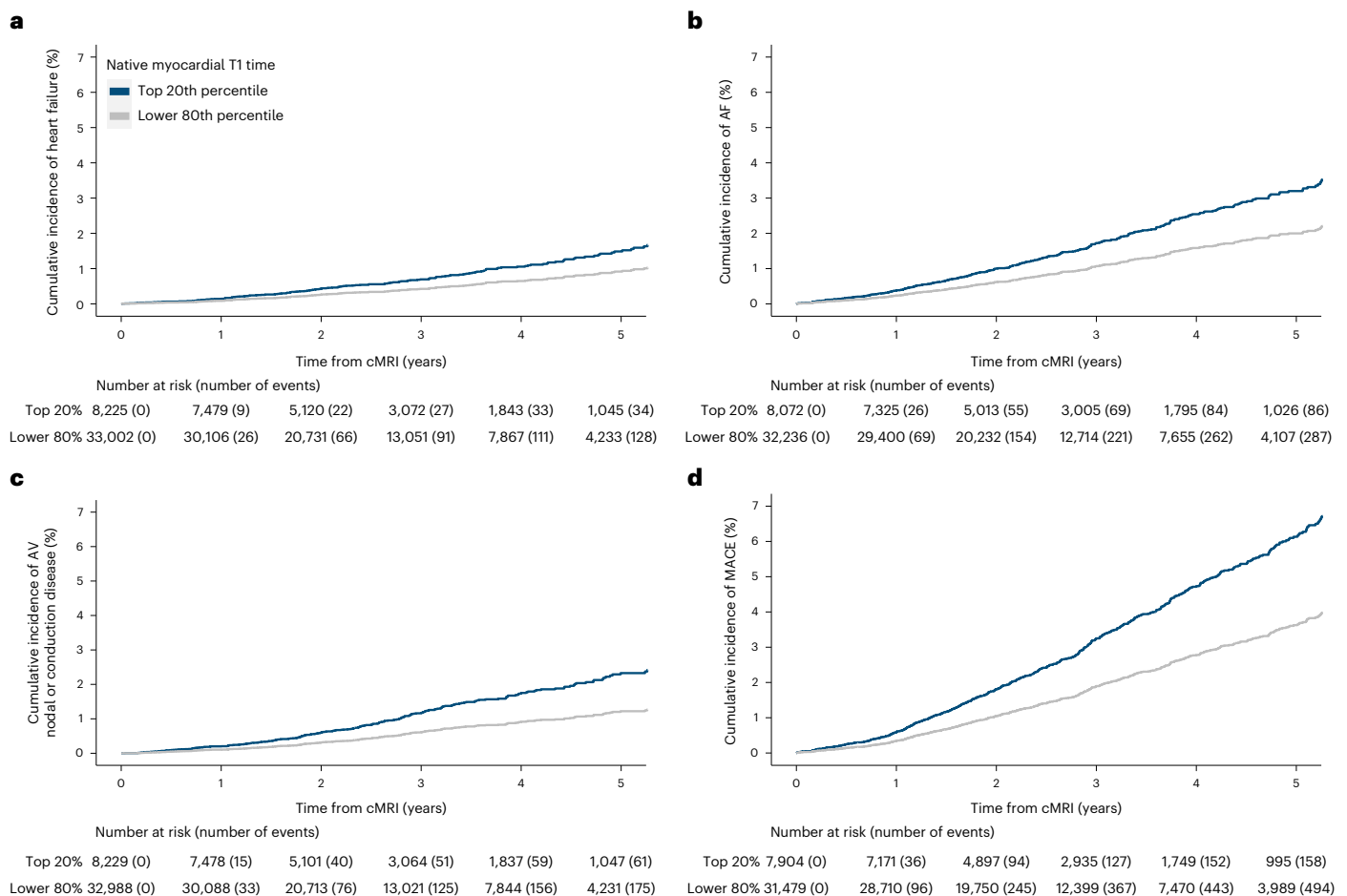
Fig. 3 | Adjusted cumulative incidence of cardiovascular events stratified by native myocardial T1 time. Plots demonstrate the cumulative incidence of incident (a) heart failure, (b) AF, (c) AV node or distal conduction disease, and (d) MACE stratified by top 20th percentile and lower 80th percentile of native myocardial T1 time.

variant in *MYH7B*, was among the lead genome-wide-significant variants ($P = 9.2 \times 10^{-11}$). *MYH7B* has been associated with familial hypertrophic cardiomyopathy[33], but no associations between rs6120777_A and hypertrophic cardiomyopathy have yet been reported. Notably, the association of rs6120777_A with T1 time persisted even after the exclusion of hypertrophic cardiomyopathy cases in a sensitivity analysis (Supplementary Table 9).

Interestingly, we also identified rs115740542_C ($P = 2.7 \times 10^{-10}$), a variant near the *HFE* gene in perfect LD with rs1800562_A, which leads to the missense change p.Cys282Tyr and is the most common cause of hereditary hemochromatosis, an iron overload disorder associated with cardiomyopathy[34]. Additionally, rs855791_A, a missense variant in *TMPRSS6* with previously reported associations with iron homeostasis[35], was associated with T1 time ($P = 5.7 \times 10^{-10}$).

We identified a variant in *CAMK2D* (rs55754224_T; $P = 1.4 \times 10^{-9}$) previously associated with AF[36]. The remaining genome-wide-significant variants were located near genes associated with cardiac arrhythmias[37–39], cardiac remodeling[40] and myocyte cytoskeletal proteins[41], including *PPP2R3A*, *PIM1* and *KANK1*, respectively.

In conditional analysis, no additional independent genome-wide-significant SNPs were identified. We performed a sensitivity analysis after excluding participants with diseases known to be associated with focal replacement fibrosis (MI, heart failure, and dilated and hypertrophic cardiomyopathy) and identified ten genome-wide-significant loci, nine of which overlapped with the main GWAS loci. The subthreshold variant rs2271426_T, an intronic variant in *PIK3C2B*, reached genome-wide significance ($P = 4.1 \times 10^{-8}$) in the

GWAS sensitivity analysis (Supplementary Figs. 17 and 18 and Supplementary Table 9).

Excluding participants with prevalent hereditary hemochromatosis ($n = 62$) did not alter the association with T1 time of the two loci, *H2BC4/HFE* and *TMPRSS6*, associated with iron homeostasis (Supplementary Figs. 19 and 20 and Supplementary Table 10). We then performed a sex-stratified GWAS and identified no new or sex-specific loci (Supplementary Figs. 21 and 22 and Supplementary Table 11).

ADAMTSL1 and SLC2A12 expression is associated with fibrosis

Of the 11 lead SNPs identified in the GWAS of T1 time, 8 (or their proxies, $r^2 > 0.8$) were expression quantitative trait loci (eQTLs) in the LV or right atrial (RA) appendage (Supplementary Table 12). Notably, rs1576900_A, the lead GWAS intronic variant in *ADAMTSL1*, was also the top eQTL variant for *ADAMTSL1* in both the LV and RA appendage and was associated with decreased *ADAMTSL1* expression and lower T1 time (Supplementary Figs. 23 and 24). In addition, rs2627230_T, the lead GWAS variant in a regulatory region upstream of *SLC2A12*, was the top eQTL variant for *SLC2A12* in the LV and was associated with decreased *SLC2A12* expression and higher T1 time (Supplementary Fig. 23).

In the transcriptome-wide association analysis, increased expression of *ADAMTSL1* in the LV tissue ($P = 5.3 \times 10^{-9}$) and RA appendage ($P = 1.3 \times 10^{-7}$) was associated with an increase in myocardial interstitial fibrosis as measured by T1 time. On the other hand, increased expression of *SLC2A12* in the LV ($P = 1.9 \times 10^{-10}$) was associated with a decrease in myocardial interstitial fibrosis (Fig. 4b and Supplementary Tables 13 and 14).
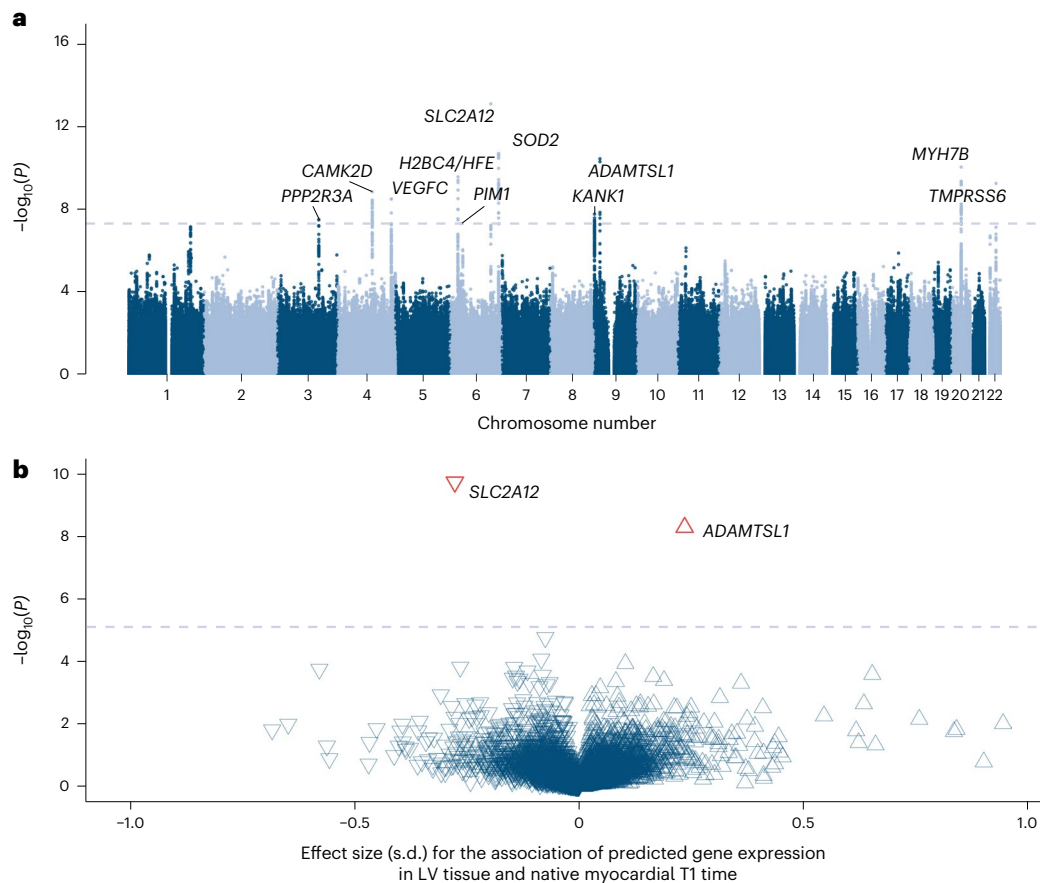
Fig. 4 | Genome-wide and transcriptome-wide association analyses. a, Native myocardial T1 time genome-wide association results across 22 autosomes. Nearest genes were used for annotation. A fixed-effect multiple linear regression model was implemented. The dashed gray line represents the threshold for genome-wide significance (two-sided $P < 5 \times 10^{-8}$ adjusted for multiple testing). **b,** Volcano plot depicting transcriptome-wide association results for native myocardial T1 time using human LV tissue gene expression from GTEx v.8 and S-PrediXcan. The upward-facing triangles reflect increased T1 time associated with increased gene expression in LV tissue. The downward-facing triangles reflect decreased T1 time associated with increased gene expression in LV tissue. The dashed gray line represents the threshold for transcriptome-wide significance (two-sided $P < 7.5 \times 10^{-6}$ adjusted for multiple testing).

Multi-omic assessment of T1 time-associated loci

To further evaluate the T1 time genome-wide-significant loci, we implemented a cellular assay for fibroblast activation whereby primary human cardiac fibroblasts are stimulated with transforming growth factor β1 (TGFβ1) (Fig. 5a). We profiled the transcriptional and epigenomic characteristics of these TGFβ1-stimulated fibroblasts temporally using RNA sequencing (RNA-seq) and assay for transposase-accessible chromatin with sequencing (ATAC–seq). Principal component analysis (PCA) of the transcriptomic data revealed that the TGFβ1-stimulated cardiac fibroblasts followed a distinct trajectory (Fig. 5b). Differential expression analysis (false discovery rate (FDR) < 0.01 and log$_2$ fold change > 1) comparing unstimulated controls to cells treated with TGFβ1 for 72 h (Supplementary Table 15) uncovered several regulators of cardiac fibrosis, including *MEOX1* (Fig. 5c)[42]. To interrogate our GWAS loci, we examined the genes nearest to the lead SNPs within a locus in addition to nonoverlapping genes within these loci that were identified in our eQTL lookup. Notably, *MYH7B*, *TMPRSS6*, *NCK1-AS1*, *U91328.19* and *PNLDC1* are not expressed in cardiac fibroblasts and could not be assessed. Of the 14 genes expressed in cardiac fibroblasts, 5 genes in 5 distinct loci showed evidence of differential gene expression with TGFβ1 treatment at FDR < 0.01 (Fig. 5d).

Next, we interrogated the open chromatin landscapes in cardiac myofibroblasts. Consistent with the efficacy of our assay, we observed differential chromatin accessibility at canonical downstream TGFβ targets, including *IGFBP3* (Fig. 5e). Overall, the epigenetic and transcriptional trajectories were similar (Fig. 5f). Differential chromatin

accessibility analysis (FDR < 1 × 10^{-10}; Supplementary Table 16) across stimulated fibroblasts uncovered 15,428 peaks that separated into six clusters (Fig. 5g). Among these peaks, we identified an enrichment for the SMAD and TEAD binding sites, which have recently been implicated as regulators of fibroblast-to-myofibroblast cell state transitions using de novo motif enrichment analysis (Fig. 5g)[43]. Of the 19 prioritized genes, 7 genes within 7 distinct loci had annotated differentially accessible peaks with TGFβ1 treatment (FDR < 1 × 10^{-10}, log$_2$ fold change > 1) (Fig. 5h). Intersecting our gene set with the differentially expressed genes (DEGs) and ATAC peaks, we identified three overlapping genes including *VEGFC*, *KANK1* and *PIM1* (Fig. 5h). Consistent with a decrease in mRNA expression, we observed a decrease in promoter accessibility in *KANK1*, as well as additional changes in chromatin accessibility (Fig. 5i).

Using our prioritized gene set, 9 of the 11 genome-wide-significant loci consisted of genes that showed evidence of differential transcriptional or epigenetic signatures with TGFβ1 stimulation. Expanding our gene set to include all genes within 250 kb from the lead variants in our loci (n$_{genes}$ = 139), we identified an additional 7, 13 and 2 genes that with TGFβ1 treatment were differentially expressed only, had annotated differentially accessible peaks only or both, respectively (Supplementary Fig. 25). Using the expanded gene set, all 11 loci included genes that showed responsiveness to TGFβ1 treatment.

Discussion

We developed an automated machine learning model to measure myocardial interstitial fibrosis in over 40,000 participants in the UKB.

We identified associations between myocardial fibrosis and diabetes mellitus, renal disease, aortic stenosis, cardiomyopathy, AF, conduction disease and RA. MR analysis provided evidence for a causal effect of T1D on myocardial interstitial fibrosis. Furthermore, greater myocardial fibrosis at the time of cMRI was associated with incident cardiovascular disease over a median follow-up of 2.5 years. In a large-scale GWAS of T1 time in the human heart, we identified 11 independent loci implicating genes involved in biological pathways relevant to fibrosis, including glucose transport (*SLC2A12*), iron homeostasis (*HFE*, *TMPRSS6*), tissue repair (*ADAMTSL1*, *VEGFC*), oxidative stress (*SOD2*), cardiac hypertrophy (*MYH7B*) and calcium signaling (*CAMK2D*). Using a cellular assay for TGFβ1-mediated cardiac fibroblast activation, we demonstrated that the GWAS loci were enriched for genes that exhibited transcriptional and epigenetic changes after treatment with TGFβ1, further supporting their functional relevance to myocardial fibrosis. Overall, the heritability of myocardial interstitial fibrosis as measured by native myocardial T1 time was relatively low, emphasizing the important contribution of nongenetic environmental and lifestyle factors to cardiac fibrosis.

Our findings have several major implications. First, our results highlight the role of glucose homeostasis and diabetes in myocardial fibrosis and pinpoint potential additional pathways for further interrogation. Our strongest GWAS variant fell within a regulatory region upstream of *SLC2A12*, which encodes an insulin-independent glucose transporter (GLUT12) highly expressed in the heart[25]. Additionally, in the MR analysis, we found evidence supporting a potential causal association between T1D and increased myocardial T1 time. GLUT12 knockout in zebrafish leads to the development of heart failure and a diabetic phenotype[27], which is consistent with the results of our transcriptome-wide association study, suggesting that decreased expression of *SLC2A12* in cardiac tissue was associated with increased interstitial fibrosis.

Second, the pathways involved in tissue repair were associated with myocardial fibrosis. Increased human cardiac expression of *ADAMTSL1*, which encodes an ADAMTS-like protein lacking catalytic activity and is thought to modulate the function of ADAMTS metalloproteinases with integral roles in ECM turnover[31,32], was associated with higher myocardial interstitial fibrosis in this study. The exact effect of ADAMTSL-1 on ADAMTS metalloproteinases is unknown; however, homology between mammalian ADAMTSL-1 and invertebrate papilin, a known inhibitor of ADAMTS-2, has been reported[44,45]. Studies in mice with cardiac-specific overexpression of *Adamts2* showed an abrogated pressure-overload-induced hypertrophic response[46]. Thus, potential inhibition of ADAMTS-2 activity may explain the increased myocardial fibrosis associated with increased expression of *ADAMTSL1* in the human myocardium. *VEGFC*- and *VEGFD*-mediated lymphangiogenesis has been associated with cardiac repair; knockout zebrafish models go on to develop severe cardiac hypertrophy and myocardial interstitial fibrosis[30]. In this study, the lead SNP rs365843_T tagging *VEGFC* was associated with increased expression of *VEGFC* in human RA appendage tissue, and with lower T1 time reflecting decreased myocardial interstitial fibrosis. In line with the findings described above, TGFβ1 treatment of human cardiac fibroblasts was associated with decreased expression of *VEGFC* in our cellular assay. Thus, our current findings extend those from animal models and suggest a role for reparative pathways involving *ADAMTS* and *VEGFC* in reducing myocardial fibrosis in the human heart.

Third, our results shed light on the role of myocardial oxidative stress in the development of myocardial interstitial fibrosis. The lead SNP rs9457699_G is an eQTL for *SOD2* in the human LV and was associated with increased expression of *SOD2* and lower T1 time. This is congruent with findings from *Sod2*-knockout mice, which exhibit increased levels of reactive oxygen species with associated myocardial fibrosis and development of dilated cardiomyopathy[29].

Fourth, several established pathways involved in myocardial fibrosis were implicated in our results. We provide further evidence for the role of *CAMK2D* in pathological cardiac remodeling and fibrosis[47,48] via association of rs55754224_T, an intronic variant in *CAMK2D*, with increased myocardial T1 time. rs55754224_T has been associated with increased risk of AF[36], an atrial arrhythmia in which atrial remodeling and myocardial fibrosis are central pathological features[7]. *PIM1* has a role in antagonizing cell senescence[49] and reducing myocardial infarct size[50] and was implicated in our results.

Fifth, we identified two genes, *HFE* and *TMPRSS6*, which were associated with iron homeostasis in our GWAS even after exclusion of participants with prevalent hemochromatosis. Iron deposition in the heart alters the magnetic properties of myocardial tissue and is associated with lower T1 time and with the development of iron overload cardiomyopathy[51].

Sixth, we found an association between modifiable lifestyle behaviors and burden of myocardial interstitial fibrosis. Our findings emphasize the importance of abstaining from cigarette smoking and regular physical activity for cardiovascular health. Extremes of weight gain or loss were associated with greater myocardial fibrosis, particularly in males. The J-shaped association between BMI and cardiovascular outcomes has been increasingly recognized in epidemiological studies[52]. Alcohol use was associated with decreased T1 time in both males and females, which has been reported in previous studies[53]. Prolonged heavy alcohol consumption (more than 50 drinks per week for more than 10 years) is associated with alcohol-induced cardiomyopathy[54]; however, this population is not well represented in the UKB.

Our study has several limitations. First, contrast agents were not used in the UKB cMRI protocol, which did not allow the assessment of late gadolinium enhancement and extracellular volume fraction. Second, the T1 maps in the UKB were obtained at a single mid-ventricular short-axis slice; thus, we cannot be certain that a single slice is representative of myocardial fibrosis throughout the LV. Third, the presence of tissue edema and paramagnetic ions may alter T1 time independently of interstitial fibrosis. Fourth, UKB participants are predominantly of European ancestry; thus, findings from our genetic analysis may not apply to other ancestries. As new multi-ancestry biorepositories with deep phenotyping emerge, further validation of our study findings in an independent multi-ancestry sample will

Fig. 5 | Multi-omic examination of human cardiac fibroblast activation. a, Schematic created with BioRender.com describing the cardiac fibroblast activation experiments. **b**, PCA of control and TGFβ1-treated cardiac fibroblast RNA-seq. **c**, Volcano plot displaying DEGs between control (0 h) and stimulated (72 h of TGFβ1 treatment) cardiac fibroblasts assessed using a generalized linear model implementing a negative binomial distribution. Expected differentially expressed cardiac fibrosis regulator genes are labeled. The red dots indicate significant DEGs (FDR < 0.01) with log₂ fold change > 1.0 associated with TGFβ1 treatment. **d**, Heatmap of normalized expression levels for prioritized genes associated with GWAS loci and expressed in cardiac fibroblasts. The red font indicates that the gene is significantly differentially expressed (FDR < 0.01). **e**, Genome browser tracks showing expected enhanced chromatin accessibility in cardiac fibroblasts around *IGFBP1* and *IGFBP3*, which are known to mediate TGFβ1-induced cardiac fibroblast activation. The labels indicate hours after TGFβ1 treatment. **f**, PCA of ATAC–seq data from activated cardiac fibroblasts. **g**, Left: heatmap displaying differential chromatin accessibility analysis presented as normalized accessibility counts. The columns represent average accessibility for three replicates. Right: de novo motif enrichment analysis carried out on the clusters of differentially accessible ATAC–seq peaks. **h**, Venn diagram showing the intersection of significant DEGs from RNA-seq, differentially accessible peaks enriched in TGFβ1-treated fibroblasts (FDR < 1 × 10⁻¹⁰) and prioritized genes associated with GWAS loci. **i**, Representative genome browser track of ATAC–seq data depicting decreased chromatin accessibility in a promoter-flanking region upstream of the *KANK1* locus after TGFβ1 treatment, which is concordant with the RNA-seq data demonstrating decreased expression with TGFβ1 treatment. Genomic coordinates are based on the GRCh38 reference genome.

be possible. Fifth, longer follow-up and continued imaging of UKB participants will allow for better-powered analyses examining the prognostic role of myocardial interstitial fibrosis for cardiovascular disease in the future. Notably, our findings of increased risk of incident major arrhythmia associated with increased interstitial fibrosis should be interpreted with caution given the relatively limited number of events ($n = 75$) examined. Sixth, phenotypic characterization

of the UKB was performed using disease and procedure codes, and self-report using surveys, which may be subject to misclassification bias. Seventh, our cellular assay for cardiac fibrosis was based on TGFβ1 stimulation of cardiac fibroblasts; however, multiple other cell types and non-TGFβ1-dependent profibrotic pathways contribute to myocardial interstitial fibrosis, which could not be assessed using this assay.
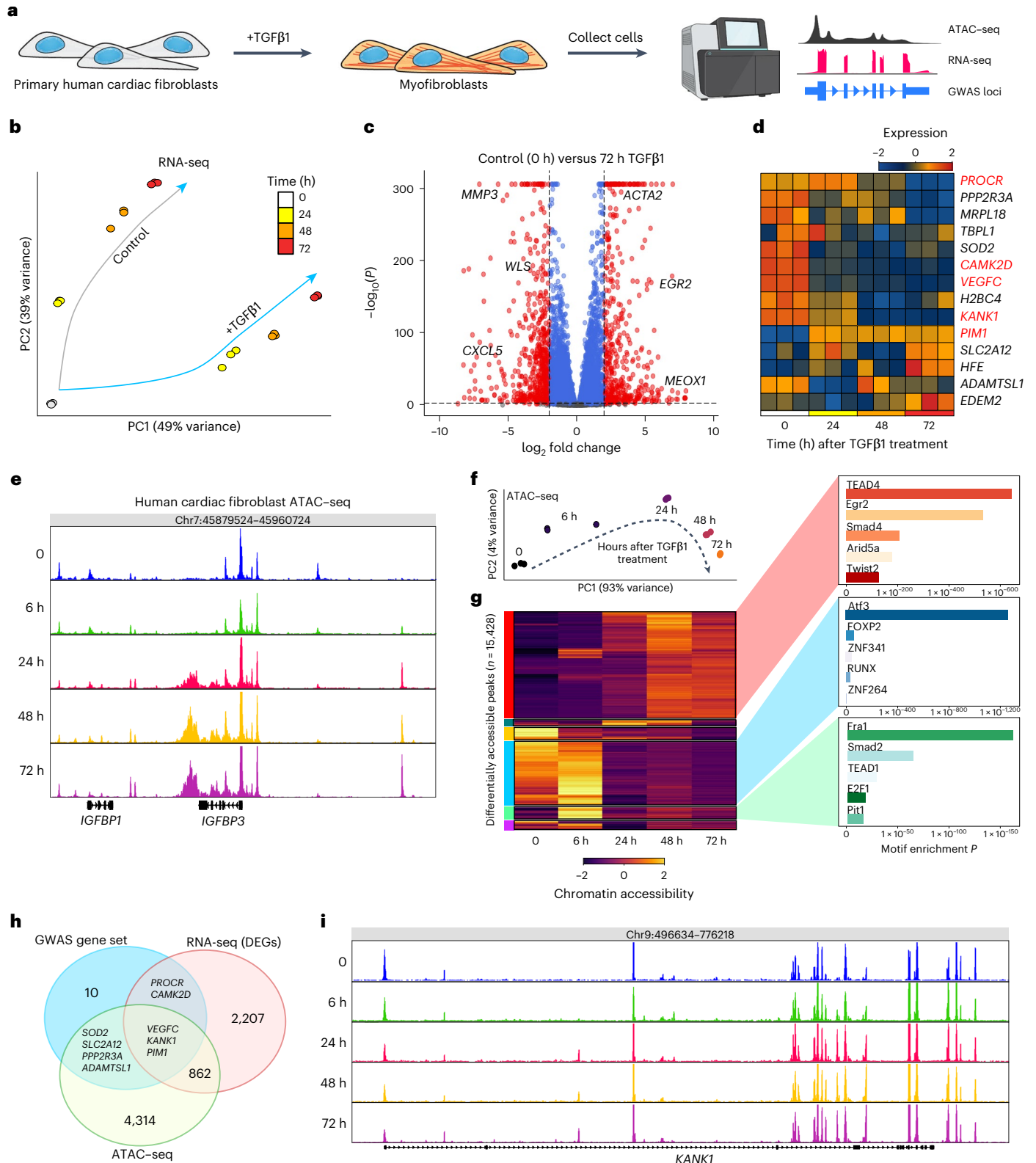
In conclusion, machine learning enables the quantification of myocardial interstitial fibrosis at scale. Our study yielded insights into new biological pathways underlying cardiac fibrosis and prioritized several pathways relevant to myocardial fibrosis for further investigation.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01371-5.

References

1. Del Monte-Nieto, G., Fischer, J. W., Gorski, D. J., Harvey, R. P. & Kovacic, J. C. Basic biology of extracellular matrix in the cardiovascular system, part 1/4: *JACC* Focus Seminar. *J. Am. Coll. Cardiol.* **75**, 2169–2188 (2020).
2. aus dem Siepen, F. et al. T1 mapping in dilated cardiomyopathy with cardiac magnetic resonance: quantification of diffuse myocardial fibrosis and comparison with endomyocardial biopsy. *Eur. Heart J. Cardiovasc. Imaging* **16**, 210–216 (2015).
3. Frangogiannis, N. G. & Kovacic, J. C. Extracellular matrix in ischemic heart disease, part 4/4: *JACC* Focus Seminar. *J. Am. Coll. Cardiol.* **75**, 2219–2235 (2020).
4. Díez, J., González, A. & Kovacic, J. C. Myocardial interstitial fibrosis in nonischemic heart disease, part 3/4: *JACC* Focus Seminar. *J. Am. Coll. Cardiol.* **75**, 2204–2218 (2020).
5. Bing, R. et al. Imaging and impact of myocardial fibrosis in aortic stenosis. *JACC Cardiovasc. Imaging* **12**, 283–296 (2019).
6. Nguyen, T. P., Qu, Z. & Weiss, J. N. Cardiac fibrosis and arrhythmogenesis: the road to repair is paved with perils. *J. Mol. Cell. Cardiol.* **70**, 83–91 (2014).
7. Ling, L.-H. et al. Diffuse ventricular fibrosis in atrial fibrillation: noninvasive evaluation and relationships with aging and systolic dysfunction. *J. Am. Coll. Cardiol.* **60**, 2402–2408 (2012).
8. Chen, Z. et al. Myocardial tissue characterization by cardiac magnetic resonance imaging using T1 mapping predicts ventricular arrhythmia in ischemic and non-ischemic cardiomyopathy patients with implantable cardioverter-defibrillators. *Heart Rhythm* **12**, 792–801 (2015).
9. Kong, P., Christia, P. & Frangogiannis, N. G. The pathogenesis of cardiac fibrosis. *Cell. Mol. Life Sci.* **71**, 549–574 (2014).
10. Mewton, N., Liu, C. Y., Croisille, P., Bluemke, D. & Lima, J. A. C. Assessment of myocardial fibrosis with cardiovascular magnetic resonance. *J. Am. Coll. Cardiol.* **57**, 891–903 (2011).
11. Diao, K.-Y. et al. Histologic validation of myocardial fibrosis measured by T1 mapping: a systematic review and meta-analysis. *J. Cardiovasc. Magn. Reson.* **18**, 92 (2017).
12. Petersen, S. E. et al. UK Biobank's cardiovascular magnetic resonance protocol. *J. Cardiovasc. Magn. Reson.* **18**, 8 (2015).
13. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
14. Messroghli, D. R. et al. Clinical recommendations for cardiovascular magnetic resonance mapping of T1, T2, T2* and extracellular volume: a consensus statement by the Society for Cardiovascular Magnetic Resonance (SCMR) endorsed by the European Association for Cardiovascular Imaging (EACVI). *J. Cardiovasc. Magn. Reson.* **19**, 75 (2017).
15. Rogers, T. et al. Standardization of T1 measurements with MOLLI in differentiation between health and disease—the ConSept study. *J. Cardiovasc. Magn. Reson.* **15**, 78 (2013).
16. Puntmann, V. O., Peker, E., Chandrashekhar, Y. & Nagel, E. T1 mapping in characterizing myocardial disease: a comprehensive review. *Circ. Res.* **119**, 277–299 (2016).
17. Liu, C.-Y. et al. Evaluation of age-related interstitial myocardial fibrosis with cardiac magnetic resonance contrast-enhanced T1 mapping: MESA (Multi-Ethnic Study of Atherosclerosis). *J. Am. Coll. Cardiol.* **62**, 1280–1287 (2013).
18. Roy, C. et al. Age and sex corrected normal reference values of T1, T2 T2* and ECV in healthy subjects at 3T CMR. *J. Cardiovasc. Magn. Reson.* **19**, 72 (2017).
19. Treibel, T. A. et al. Extracellular volume quantification in isolated hypertension—changes at the detectable limits? *J. Cardiovasc. Magn. Reson.* **17**, 74 (2015).
20. American Diabetes Association. 2. Classification and diagnosis of diabetes: *standards of medical care in diabetes—2020*. *Diabetes Care* **43**, S14–S31 (2020).
21. Stevens, P. E. et al. Evaluation and management of chronic kidney disease: synopsis of The Kidney Disease: Improving Global Outcomes 2012 clinical practice guideline. *Ann. Intern. Med.* **158**, 825–830 (2013).
22. Arnett, D. K. et al. ACC/AHA guideline on the primary prevention of cardiovascular disease: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on clinical practice guidelines. *Circulation* **140**, e563–e595 (2019).
23. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
24. Pirruccello, J. P. et al. Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. *Nat. Commun.* **11**, 2254 (2020).
25. Waller, A. P. et al. GLUT12 functions as a basal and insulin-independent glucose transporter in the heart. *Biochim. Biophys. Acta* **1832**, 121–127 (2013).
26. Heidecker, B. et al. The gene expression profile of patients with new-onset heart failure reveals important gender-specific differences. *Eur. Heart J.* **31**, 1188–1196 (2010).
27. Jiménez-Amilburu, V., Jong-Raadsen, S., Bakkers, J., Spaink, H. P. & Marín-Juez, R. GLUT12 deficiency during early development results in heart failure and a diabetic phenotype in zebrafish. *J. Endocrinol.* **224**, 1–15 (2015).
28. Linden, K. C. et al. Renal expression and localization of the facilitative glucose transporters GLUT1 and GLUT12 in animal models of hypertension and diabetic nephropathy. *Am. J. Physiol. Renal Physiol.* **290**, F205–F213 (2006).
29. Sharma, S. et al. SOD2 deficiency in cardiomyocytes defines defective mitochondrial bioenergetics as a cause of lethal dilated cardiomyopathy. *Redox Biol.* **37**, 101740 (2020).
30. Vivien, C. J. et al. Vegfc/d-dependent regulation of the lymphatic vasculature during cardiac regeneration is influenced by injury context. *NPJ Regen. Med.* **4**, 18 (2019).
31. Perrucci, G. L., Rurali, E. & Pompilio, G. Cardiac fibrosis in regenerative medicine: destroy to rebuild. *J. Thorac. Dis.* **10**, S2376–S2389 (2018).
32. Kelwick, R., Desanlis, I., Wheeler, G. N. & Edwards, D. R. The ADAMTS (A Disintegrin and Metalloproteinase with Thrombospondin motifs) family. *Genome Biol.* **16**, 113 (2015).
33. Chen, P. et al. *MYH7B* variants cause hypertrophic cardiomyopathy by activating the CaMK-signaling pathway. *Sci. China Life Sci.* **63**, 1347–1362 (2020).
34. Alexander, J. & Kowdley, K. V. *HFE*-associated hereditary hemochromatosis. *Genet. Med.* **11**, 307–313 (2009).
35. Ramsay, A. J., Hooper, J. D., Folgueras, A. R., Velasco, G. & López-Otín, C. Matriptase-2 (TMPRSS6): a proteolytic regulator of iron homeostasis. *Haematologica* **94**, 840–849 (2009).
36. Roselli, C. et al. Multi-ethnic genome-wide association study for atrial fibrillation. *Nat. Genet.* **50**, 1225–1233 (2018).

37. Heijman, J., Ghezelbash, S., Wehrens, X. H. T. & Dobrev, D. Serine/threonine phosphatases in atrial fibrillation. *J. Mol. Cell. Cardiol.* **103**, 110–120 (2017).

38. Lubbers, E. R. & Mohler, P. J. Roles and regulation of protein phosphatase 2A (PP2A) in the heart. *J. Mol. Cell. Cardiol.* **101**, 127–133 (2016).

39. Ramirez, A. H. et al. Novel rare variants in congenital cardiac arrhythmia genes are frequent in drug-induced torsades de pointes. *Pharmacogenomics J.* **13**, 325–329 (2013).

40. Zhu, N. et al. Pim-1 kinase phosphorylates cardiac troponin I and regulates cardiac myofilament function. *Cell. Physiol. Biochem.* **45**, 2174–2186 (2018).

41. Pan, W. et al. Structural insights into ankyrin repeat-mediated recognition of the kinesin motor protein KIF21A by KANK1, a scaffold protein in focal adhesion. *J. Biol. Chem.* **293**, 1944–1956 (2018).

42. Alexanian, M. et al. A transcriptional switch governs fibroblast activation in heart disease. *Nature* **595**, 438–443 (2021).

43. Xiao, Y. et al. Hippo pathway deletion in adult resting cardiac fibroblasts initiates a cell state transition with spontaneous and self-sustaining fibrosis. *Genes Dev.* **33**, 1491–1505 (2019).

44. Wight, T. N. The ADAMTS proteases, extracellular matrix, and vascular disease: waking the sleeping giant(s)! *Arterioscler. Thromb. Vasc. Biol.* **25**, 12–14 (2005).

45. Hirohata, S. et al. Punctin, a novel ADAMTS-like molecule, ADAMTSL-1, in extracellular matrix. *J. Biol. Chem.* **277**, 12182–12189 (2002).

46. Wang, X. et al. Critical role of ADAMTS2 (a disintegrin and metalloproteinase with thrombospondin motifs 2) in cardiac hypertrophy induced by pressure overload. *Hypertension* **69**, 1060–1069 (2017).

47. Willeford, A. et al. CaMKIIδ-mediated inflammatory gene expression and inflammasome activation in cardiomyocytes initiate inflammation and induce fibrosis. *JCI Insight* **3**, 97054 (2018).

48. Ling, H. et al. Requirement for Ca^{2+}/calmodulin-dependent kinase II in the transition from pressure overload-induced cardiac hypertrophy to heart failure in mice. *J. Clin. Invest.* **119**, 1230–1240 (2009).

49. Ebeid, D. E. et al. PIM1 promotes survival of cardiomyocytes by upregulating c-Kit protein expression. *Cells* **9**, 2001 (2020).

50. Muraski, J. A. et al. Pim-1 regulates cardiomyocyte survival downstream of Akt. *Nat. Med.* **13**, 1467–1475 (2007).

51. Torlasco, C. et al. Role of T1 mapping as a complementary tool to T2* for non-invasive cardiac iron overload assessment. *PLoS ONE* **13**, e0192890 (2018).

52. Song, X. et al. Cardiovascular and all-cause mortality in relation to various anthropometric measures of obesity in Europeans. *Nutr. Metab. Cardiovasc. Dis.* **25**, 295–304 (2015).

53. Voskoboinik, A. et al. Relation of alcohol consumption to left ventricular fibrosis using cardiac magnetic resonance imaging. *Am. J. Cardiol.* **123**, 460–465 (2019).

54. Fernández-Solà, J. Cardiovascular risks and benefits of moderate and heavy alcohol consumption. *Nat. Rev. Cardiol.* **12**, 576–587 (2015).

[1]Cardiovascular Division, Brigham and Women's Hospital, Boston, MA, USA. [2]Cardiovascular Disease Initiative, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [5]Cardiology Division, Massachusetts General Hospital, Boston, MA, USA. [6]Division of Cardiology, University of California San Francisco, San Francisco, CA, USA. [7]Demoulas Center for Cardiac Arrhythmias, Massachusetts General Hospital, Boston, MA, USA. [8]Medical Center Groningen, University of Groningen, Groningen, the Netherlands. [9]Center for Computational Health, IBM Research, Cambridge, MA, USA. [10]Eric and Wendy Schmidt Center, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [11]These authors contributed equally: Victor Nauffal, Paolo Di Achille, Marcus D. R. Klarqvist, Jonathan W. Cunningham, Matthew C. Hill. [12]These authors jointly supervised this work: Patrick T. Ellinor, Steven A. Lubitz. ✉e-mail: ellinor@mgh.harvard.edu; slubitz@mgh.harvard.edu

Methods

We trained a machine learning model to segment cardiac T1 maps from the UKB and measure native myocardial T1 time at the IVS. We examined the associations between T1 time and cardiometabolic risk factors, cardiovascular diseases, serum biomarkers, ECG intervals and lifestyle factors. We performed a two-sample MR analysis to examine the causal effect of cardiovascular risk factors and disease on myocardial interstitial fibrosis. We then performed genome- and transcriptome-wide association analyses of native myocardial T1 time. Lastly, we explored the relevance of the identified T1 time-associated loci to myocardial fibrosis by profiling the transcriptional and open chromatin characteristics of genes within these loci using a cellular assay for TGFβ1-mediated cardiac fibroblast activation.

Study design and sample

The UKB is a prospective cohort of 502,629 individuals from the United Kingdom enrolled between 2006 and 2010 with deep phenotyping, imaging and multiple genomic data types. Cohort design has been described previously[13,55]. Briefly, around 9.2 million individuals aged 40–69 years living in England, Scotland and Wales were invited to participate in the study, and 5.4% agreed to participate. Extensive questionnaire data, physical measures and biological samples were collected at baseline, with ongoing data collection in large subsets of the cohort, including repeated assessments and multimodal imaging. Starting in 2014, 42,654 participants returned for the first multimodal imaging visit, including cMRI with T1 mapping, allowing for the assessment of myocardial interstitial fibrosis[12]. All study participants are followed longitudinally for health-related outcomes through linkage to national health-related datasets.

Use of UKB data was performed under application no. 7089 and was approved by the local Massachusetts General Hospital institutional review board. All participants signed written informed consent before participating in the UKB.

cMRI T1 mapping protocol and image quality control

A standardized non-contrast-enhanced cMRI protocol using a clinical wide-bore 1.5-T scanner (MAGNETOM Aera, Syngo Platform VD13A, Siemens Healthcare) was performed on all cMRI substudy participants. The scanner is equipped with 48 receiver channels, a 45 mT m^{-1} and 200 T m^{-1} s^{-1} gradient system, an 18-channel anterior body surface coil used in combination with 12 elements of an integrated 32-element spine coil and ECG gating for cardiac synchronization. The imaging protocol included: three long-axis cines, one short-axis cine, a phase contrast sequence at the LV outflow tract, three-segment short-axis tagging and mid-ventricular short-axis T1 mapping. Native T1 mapping within a single breath hold was performed using the shortened modified Look–Locker inversion recovery (shMOLLI, WIP780B) technique. The following imaging parameters for T1 mapping were implemented: field of view 360 × 236 mm, voxel size 0.9 × 0.9 × 8.0, flip angle 35 degrees and repetition time/echo time 2.6/1.07 ms^{12}. T1 maps were generated online and stored in the UKB imaging database.

To date, the UKB MRI core laboratory has only released raw T1 maps to UKB researchers. As such, we developed our own automated pipeline to measure native myocardial T1 time from raw T1 maps (Fig. 1). First, we set up an automatic procedure to identify raw T1 map series among the several files provided by the UKB under the category 'experimental shMOLLI sequence images' (UKB field ID: 20214). Preliminary explorations indicated that the T1 map series names contained the 't1map' keyword and explicitly mentioned the MRI 'sax (short-axis)' view. Therefore, we discarded all series with names not containing either of the two keywords. Then, we set up a standardized quality control process using a custom online tool to streamline the review of the 42,654 selected T1 maps by four experienced MRI reviewers (V.N., M.D.R.K., P.D.A. and J.W.C.; Supplementary Fig. 26). Native myocardial T1 time was measured at the IVS[14]. All images were reviewed and assessed for overall image quality and artifacts involving myocardial segments (IVS and LV free wall). A cardiologist (V.N.) reviewed all images that were flagged by any of the four reviewers and made a final ascertainment on image quality and the extent of artifacts involving myocardial segments. Off-axis images and those with severe distortion of overall image pixel intensity were excluded (Supplementary Fig. 27). An artifact within a myocardial segment was deemed major if it affected at least one-third of the segment of interest (Supplementary Fig. 27). T1 maps with major artifacts involving the IVS (n = 1,149) were excluded from the final T1 time study sample (Supplementary Fig. 2). T1 maps with major artifacts restricted to the LV free wall were included and contributed to the final T1 time study sample (Supplementary Fig. 27). Artifacts were threefold higher within the LV free wall segment than within the IVS. Current consensus guidelines recommend using the IVS to avoid susceptibility to artifacts from lung, liver or veins and for accurate and reproducible measurement of native myocardial T1 time[14–16].

Semantic segmentation and T1 time measurement

Six hundred (500 training, 100 validation) T1 maps were randomly selected and used to develop our machine learning model (Supplementary Fig. 2). To train a machine learning model to segment the IVS, two cardiologists (V.N. and J.W.C.) labeled all cardiac structures within the short-axis T1 maps (350 V.N. and 250 J.W.C.). Fifty T1 maps were labeled by both readers to allow for assessment of inter-reader reliability. Cardiac structures that were labeled included IVS, LV free wall, papillary muscles and trabeculae, LV blood pool, right ventricle (RV) free wall and RV blood pool. Additionally, an ROI encompassing the mid-myocardium within the IVS and excluding the blood pool on either side was delimited (Supplementary Fig. 28). Pixel intensity values were transformed to T1 times using the accompanying T1 map legend. Native myocardial T1 time was measured as the median T1 time for all pixels within the corresponding IVS ROI.

The manual tracing procedure, called semantic segmentation, displayed high inter-reader concordance at the IVS between the cardiologist-labeled segmentations, as measured in 50 overlapping cMRI acquisitions (Sørensen–Dice coefficient = 0.84, 95% CI = 0.76–0.92). Additionally, IVS-derived T1 times were highly correlated between the two readers (r = 0.95, 95% CI = 0.92–0.97) (Supplementary Fig. 29).

We then trained a machine learning model to identify the IVS using the cardiologist-segmented data as truth labels. The machine learning model had high accuracy for segmenting the IVS when tested in the validation set (Sørensen–Dice coefficient = 0.82, 95% CI = 0.70–0.94). Model predictions were then post-processed. First, we subtracted the blood pool, papillary muscles and trabeculae, LV and RV free walls, and any noncardiac structures and retained the auto-segmented IVS. Second, to minimize the potential of contamination by residual blood pool, trabeculae or noncardiac structures, we applied a sequence of morphological operations (that is, skeletonization followed by dilation with a three-pixel kernel) to the IVS segment and generated representative mid-myocardial ROIs within the IVS. In the validation set, automatically generated T1 times were highly correlated with T1 times derived from manually traced ROIs (r = 0.97, 95% CI = 0.95–0.98) (Supplementary Fig. 1).

Next, the machine learning model was used to segment the IVS in the remaining 42,054 cMRIs not used for model training or validation, followed by automated selection of ROIs and measurement of T1 time. To maximize the quality of the generated T1 times, we manually reviewed all 42,654 T1 maps to exclude low-quality acquisitions and major artifacts affecting the IVS (see 'cMRI T1 mapping protocol and image quality control' section). After quality control, we retained measured T1 time at the IVS for 41,505 participants who constituted our study sample (Supplementary Fig. 2). The mean T1 time of the study sample was 918.1 ± 41.5 ms (Table 1). These values are consistent with previously reported T1 times in a smaller study from the UKB[56] including 11,882 cMRI scans.

Machine learning model development

For segmenting cardiac structures in cMRI T1 maps, we employed the DenseNet-121 architecture[57] as the base encoder model in a U-Net model[58] that was pretrained on ImageNet[59]. DenseNets are constructed with two principal building blocks: (1) dense blocks consisting of batch normalization, the nonlinear ReLU activation function and 3×3 convolutions of increasing number of channels that are propagated from previous layers to enable efficient gradient flow; and (2) transition blocks that compress the number of channels by half-using channel-wise convolutions (1×1), and perform a spatial reduction by a factor of 2 by using an average pooling layer of stride 2 and pool size 2. The U-Net architecture contains long-range skip connections that allow for pixel-accurate segmentation by sharing feature information along a contracting–expansive path. This is achieved by concatenating features at each downsampling in the encoder with the corresponding features at each upsampling step. These 'skip connections' preserve contextual and spatial information.

The inputs for this model were the cMRI T1 maps with size $288 \times 384 \times 3$. The models were trained with the Adam optimizer[60] with a learning rate set to a cosine decay policy, decaying from 0.0001 to 0 over 100 epochs, weight decay of 0.0001, categorical cross-entropy as the loss function and a batch size of 16. No additional hyperparameter search or ablation studies were performed.

For all training data, the following augmentations (random permutations of the training images) were applied: random shifts in the xy plane by up to plus or minus 16 pixels and rotations by up to plus or minus five degrees around its center axis.

Phenotype derivation and association with T1 time

Prevalent cardiometabolic, cardiovascular and systemic inflammatory diseases at the time of the first visit for cMRI and incident cardiovascular events were ascertained using International Classification of Diseases and Related Health Problems, 9th and 10th revisions, codes and OPCS Classification of Interventions and Procedures v.4 codes as well as self-report using surveys (Supplementary Table 17). Derived myocardial T1 times were rank-based and inverse-normal-transformed; thus, reported changes in T1 times are dimensionless and reflect approximately multiples of 1 s.d. of the underlying quantitative trait. Multiple linear regression was used to assess the association of prevalent cardiometabolic, systemic inflammatory and cardiovascular diseases selected a priori and relevant to myocardial fibrosis at the time of MRI with T1 time by comparing disease cases to 'healthy' controls free of cardiovascular and metabolic disorders. Healthy participants were selected if free of prevalent dilated cardiomyopathy, hypertrophic cardiomyopathy, heart failure, AF, AV node and distal conduction disease, hypertension, diabetes mellitus, aortic stenosis, CKD, hemochromatosis and RA. Multivariable models were adjusted for age at MRI visit, sex, BMI, MRI scanner, beta-blocker use, angiotensin-coverting enzyme (ACE) or angiotensin receptor blocker (ARB) use, statin use and mineralocorticoid receptor antagonist use. We observed a J-shaped relationship between BMI and T1 time with higher BMI associated with lower T1 time until the obesity threshold (BMI > 30) was exceeded, at which point BMI was associated with increased T1 time. As such, we modeled BMI as a linear spline with a knot at 30 kg m^{-2}. We performed three sensitivity analyses. First, we performed an analysis comparing cases to non-cases and further adjusting for comorbidities including MI, T1D, T2D and CKD. Second, we performed a sex-stratified analysis to examine trends of association across males and females. Third, to examine the impact of the presence of LVH on the association of hypertension with T1 time, we performed a stratified analysis by LVH status. LVH was defined as LV mass indexed to body surface area greater than 72 g m^{-2} in males and greater than 55 g m^{-2} in females using recently defined reference values in the UKB[61]. $P < 3.1 \times 10^{-3}$ (0.05/16) was used to determine statistically significant associations after adjusting for multiple testing.

We leveraged biomarker and ECG data in the UKB to examine the association of T1 time with serum biomarkers and ECG intervals associated with the examined prevalent diseases. Examined serum biomarkers were measured at the time of enrollment in the UKB and included HbA1c, lipoproteins, C-reactive protein, creatinine, cystatin C and insulin-like growth factor I. Of 41,505 participants with T1 time measured, 33,241 had serum biomarker data available. The examined ECG intervals included automated measurements of the QRS interval, P-wave duration and PR interval (Marquette 12SL ECG analysis program, GE Healthcare) from supine-resting 12-lead ECGs performed at the time of MRI. We excluded ECGs from participants with a diagnosis of Wolff–Parkinson–White syndrome, a paced rhythm, AF or atrial flutter at the time of the ECG, second- or third-degree AV block, digoxin use or class I or III antiarrhythmic drug use. Of 41,505 participants with T1 time measured, 28,602 had ECG data that met our inclusion criteria. Estimated glomerular filtration was calculated using the CKD-EPI equation incorporating both serum creatinine and cystatin C[62]. We categorized biomarkers based on clinically established thresholds where possible[20,21,63,64]. C-reactive protein and insulin-like growth factor I were categorized into top decile versus lower 90th percentile. Multiple linear regression was used to assess the association of biomarkers and ECG intervals with T1 time, further adjusting for age at MRI, sex, BMI, MRI scanner, MI, beta-blocker use, ACE inhibitor or ARB use, statin use and mineralocorticoid receptor antagonist use. A two-sided $P < 4.6 \times 10^{-3}$ (0.05/11) was considered statistically significant.

A time-to-event analysis was performed to assess the association of T1 time with incident cardiovascular events. Follow-up time was defined as the time from MRI visit to first occurrence of the outcome of interest, death or last follow-up (30 April 2020). For each incident disease analysis, study participants with prevalent disease at the time of MRI were excluded because they were not at risk for the outcome of interest. We then stratified the cohort into the upper 20th and lower 80th percentiles of T1 time. Using a multivariable Cox proportional-hazards model adjusted for age at MRI, sex, BMI, MRI scanner, MI, T1D, T2D, CKD, beta-blocker use, ACE inhibitor and ARB use, statin use and mineralocorticoid receptor antagonist use, we examined the association of T1 time with incident cardiovascular events. Adjusted Kaplan–Meier curves were constructed to compare the incidence rate of cardiovascular events between the two groups. The validity of the proportional-hazards assumption was verified by examining the Schoenfeld residuals. A two-sided $P < 0.01$ (0.05/5) was considered statistically significant. All statistical tests were performed using R v.4.0.2 (R Foundation for Statistical Computing).

Association analysis of lifestyle factors with T1 time

Lifestyle factors including alcohol use, cigarette smoking, physical activity and obesity were examined for their association with T1 time. Alcohol use data were based on self-report in the UKB. All participants completed a touchscreen questionnaire at their initial assessment containing a series of questions pertaining to alcohol use. Participants reporting weekly drinking were shown a chart with common alcoholic beverages and corresponding units of alcohol and asked to report the average number of drinks consumed per week by category. Participants reporting drinking once to three times a month or on special occasions were asked to report the average number of drinks consumed per month. We aggregated these data and converted the number of drinks to grams of alcohol based on units defined by the UK National Health Service (https://www.nhs.uk/live-well/alcohol-advice/calculating-alcohol-units/; for example, one standard glass of wine = 2.1 units = 16.8 g alcohol). Participants who responded with 'prefer not to answer,' 'I do not know' or with incomplete responses were excluded. According to the National Institute on Alcohol Abuse and Alcoholism, light-to-moderate alcohol use, termed 'drinking in moderation', was defined as 98 g per week or less for women and 196 g per week or less for men, with one standard US drink containing 14 g alcohol. Heavy alcohol

use was defined as more than 98 g per week for women and more than 196 g per week for men (https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/moderate-binge-drinking). Cigarette smoking status was categorized into nonsmokers, previous smokers and current smokers. Physical activity was assessed using self-reported activity questionnaires. Adequate physical activity was defined as self-reported activity that met standard American Heart Association guidelines[22] (150 min or more of moderate-intensity activity or 75 min or more of vigorous-intensity activity or the equivalent combination per week). Finally, for BMI and given the observed J-shaped relationship with T1 time, we report the association of BMI with T1 time separately for those with BMI < 30 and BMI ≥ 30 kg m^{-2}. We examined the association of lifestyle factors with T1 time stratified by sex and adjusted for age at MRI, MI, T1D, T2D, heart failure, AF, CKD, ACE inhibitor use, ARB use, beta-blocker use, mineralocorticoid receptor antagonist use and statin use. We additionally adjusted for BMI modeled as a linear spline with the knot at 30 kg m^{-2} when the BMI was not the examined outcome. Within each sex stratum, a two-sided $P < 1.3 \times 10^{-2}$ (0.05/4) was considered statistically significant.

Genomic data, imputation, sample and variant quality control

In total, 488,377 UKB participants were genotyped using either one of two overlapping arrays, the UK BiLEVE Axiom array or the UKB Axiom array. Before imputation, several quality control filters were applied to the genotype data. Variants with more than 5% missing rate, a minor allele frequency (MAF) < 0.0001 and that violated the Hardy–Weinberg equilibrium ($P < 1 \times 10^{-12}$) were excluded. Additionally, samples that were identified as outliers for genotype missingness rate (more than 5%) and heterozygosity were also excluded. These filters resulted in a genotype dataset that included 670,730 autosomal variants in 487,442 samples. Imputation into the Haplotype Reference Consortium and UK10K and 1000 Genomes phase 3 reference panels was carried out using IMPUTE4. The imputation process resulted in a dataset with 93,095,623 autosomal SNPs and short indels in 487,442 individuals[13].

Of 42,654 study participants who underwent cMRI with T1 mapping, 41,635 had imputed genetic data available. Sample and variant quality control filters were applied before conducting genetic association analyses. Samples with sex chromosome aneuploidy and those with discordant genetically inferred and self-reported sex were excluded. One of each pair of third-degree relatives or closer was excluded. Variants with an imputation quality score (INFO value) <0.3 and those with an MAF <0.01 were excluded. After quality control, our dataset included 40,399 individuals with 9,853,972 SNPs and short indels. Among the 40,399 study participants with adequate-quality genetic data, 39,339 had T1 time data that passed quality control and constituted the study sample for the genome-wide association analysis (Supplementary Fig. 2).

Two-sample MR analysis

We performed a two-sample MR analysis to examine the causal effect of cardiovascular risk factors and disease on myocardial interstitial fibrosis as measured using T1 time. Exposures examined included BMI, systolic blood pressure, diastolic blood pressure, CKD, estimated glomerular filtration rate, T1D, T2D, AF, CAD, low-density lipoprotein, high-density lipoprotein, total cholesterol and triglycerides. We searched the public domain for summary statistics of the examined exposures in large study samples of predominantly European ancestry and prioritized the inclusion of study samples that did not overlap with our UKB study sample where possible. If ancestry-specific results were reported, we used the European ancestry results. Details of the included summary statistics are summarized in Supplementary Table 18.

We derived genetic instruments for each cardiovascular risk factor or disease using the publicly available summary statistics after removing strand-ambiguous variants. We first subset to genome-wide-significant variants ($P < 5 \times 10^{-8}$) and pruned the

variants (LD threshold = $r^2 < 0.01$ within a 1,000-kb window) using 1000 Genomes multi-ancestry or European phase 3 LD data, based on the ancestral composition of the available summary statistics, to identify a subset of independent variants associated with each cardiovascular risk factor or disease. Next, we harmonized the external summary statistics with our T1 time summary statistics by transforming all genomic coordinates to GRCh37, harmonizing strand orientation and aligning effect alleles and effect estimates. We excluded from the genetic instrument variants with MAF <1% in our study sample or with poor imputation quality (INFO score < 0.3) (Supplementary Table 18). We performed two-sample MR using the IVW method in the Mendelian Randomization package in R[65]. We additionally used the MR-Egger method as a sensitivity analysis. The MR-Egger directional pleiotropy test assesses for the presence of horizontal pleiotropy; the MR-Egger regression test yields pleiotropy-robust causal estimates[23]. A two-sided $P < 3.85 \times 10^{-3}$ (0.05/13) was considered statistically significant after accounting for multiple testing. Finally, scatter plots of the association of the individual genetic variants included in each genetic instrument with both T1 time and the exposure under investigation were examined to assess for plausible causal associations.

Genome-wide common variant association analysis

We performed a common variant genome-wide association analysis of T1 time using a fixed-effect linear regression model in PLINK 2.0 (ref. 66). The models were adjusted for age at MRI, sex, MRI scanner, genotyping array and the first ten principal components of genetic ancestry. Rank-based inverse-normal transformation was applied to the measured myocardial T1 times. As such, effect size estimates in the GWAS are dimensionless and reflect approximately multiples of 1 s.d. of the underlying quantitative trait. A two-sided $P < 5 \times 10^{-8}$ was used to define genome-wide-significant common variants; $5 \times 10^{-8} < P < 1 \times 10^{-6}$ denoted suggestive loci. Distinct genomic loci were defined by starting with the SNP with the lowest P value, excluding other SNPs within 500 kb and iterating until no SNPs remained. The independently significant SNPs with the lowest P value at each genomic locus were termed lead SNPs. We then performed a conditional analysis adjusting for the imputed allele dosage of each lead SNP to examine for additional independent genome-wide-significant SNPs within a locus. We performed three sensitivity analyses. First, we repeated the above GWAS of T1 time after excluding individuals with prevalent diseases at the time of MRI known to be associated with focal replacement fibrosis, including heart failure, dilated cardiomyopathy, hypertrophic cardiomyopathy and MI (n_{GWAS} = 38,339). Second, we repeated the GWAS after excluding 62 participants with prevalent hereditary hemochromatosis (n_{GWAS} = 39,277) to examine whether the identified genome-wide-significant loci associated with iron homeostasis were driven by hereditary hemochromatosis cases. Third, we performed a sex-stratified GWAS for males (n_{GWAS} = 19,025) and females (n_{GWAS} = 20,314) separately.

LD score regression analysis was performed using ldsc (v.1.0.0)[67]. With ldsc, the genomic control factor (λ_{GC}) was partitioned into components reflecting polygenicity and inflation using the software's default parameters.

Regional association plots were generated with LocusZoom (v.1.4)[68] using LD data from the 1000 Genomes phase 3 European reference panel. In instances where lead SNPs were not part of the 1000 Genomes phase 3 reference panel, in-sample LD was calculated using PLINK v.1.9.

Heritability and genetic correlation analysis

SNP heritability of T1 time was assessed using BOLT-REML (v2.3.4)[69]. We also computed genetic correlation between T1 time and other cMRI measures including LV end-diastolic and systolic volumes, LV mass, LV ejection fraction, left atrial end-diastolic and systolic volumes and left atrial ejection fraction using ldsc[70]. These cMRI-based phenotypes from the UKB have been described previously[71,72].

eQTL and transcriptome-wide association study analysis

We performed an eQTL lookup using v.8 of the Genotype-Tissue Expression (GTEx) database[73]. In-sample LD was calculated for all variants within 1 Mb of genome-wide-significant lead SNPs using PLINK v.1.9. The list of proxy SNPs for each lead SNP were generated using an LD r^2 threshold > 0.8. We searched the GTEx v.8 database for statistically significant differential gene expression in the RA appendage and LV tissues associated with the lead SNPs and their proxies. When no significant differential gene expression associated with the lead SNP was identified, significant findings from the closest proxy were reported.

We then performed a transcriptome-wide association analysis to test the mediating effects of gene expression levels in RA appendage and LV tissue on T1 time. We used precomputed transcript expression reference weights derived using elastic net models from S-PrediXcan on GTEx v.8 eQTL data for the RA appendage and LV[74]. S-PrediXcan was then run with its default settings. A Bonferroni-corrected, two-sided $P < 7.5 \times 10^{-6}$ (0.05/6,637 genes tested) was used to define significant gene expression–phenotype associations.

Cardiac fibrosis assay and profiling

We used primary human cardiac fibroblasts (catalog no. ACBRI5118, Cell Systems) for all in vitro experiments. Cells were cultured in FGM-3 Cardiac Fibroblast Growth Medium (catalog no. CC-4525, Lonza). The cardiac fibroblast activation assay was initiated by performing a medium exchange for starvation medium (CC-4525 without FCS and supplements) containing TGFβ1 (catalog no. T7039, Sigma-Aldrich) at 10 ng ml^{-1}. Controls were given starvation medium only. Total RNA was extracted using the Direct-zol RNA Miniprep kit (catalog no. R2051, Zymo Research). Standard RNA-seq libraries were generated and sequenced by Genewiz on an Illumina NovaSeq. Sequenced reads were aligned to the human genome (GRCh38) using Salmon (v.1.8.0)[75]. Differential expression analysis was performed with DESeq2 (v.1.30.1)[76].

Fifty thousand cardiac fibroblasts were used as input for ATAC–seq, according to the OMNI-ATAC–seq protocol[77]. Transposed DNA was purified with a PCR MinElute kit (catalog no. 28004, QIAGEN); the final ATAC–seq libraries were purified with 1.8× SPRI purification using SPRISelect beads (Beckman Coulter) after PCR amplification. Libraries were sequenced on an Illumina NextSeq 500. Reads were mapped to the human genome (GRCh38) using Bowtie2 with default paired-end settings (v.2.3.4.3)[78]. Next, all nonnuclear and unmapped paired reads were discarded. Duplicated reads were removed with the Picard MarkDuplicates function using default settings. Visualization of the ATAC–seq signals was done with Homer (v.4.10)[79] and all reads were normalized by read count where scores represent read count per base pair per 1×10^7 reads. Motif enrichment analysis was performed with Homer using the findMotifsGenome.pl function with default parameters. Peak calling was carried out with MACS2 (callpeak--nomodel--broad) v2.2.6 using all ATAC–seq libraries as input. Reads were counted for each condition from the comprehensive peak file by using bedtools multicov module (v.2.26.0)[80]. Quantile normalization of the ATAC–seq datasets was performed with CQN (v.1.36.0)[81]; offsets were fed into DESeq2 to quantify differential accessibility.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The UKB data are made available to researchers from research institutions with genuine research inquiries according to institutional review board and UKB approval. The genome-wide association analysis summary statistics are available from the downloads page of the Cardiovascular Disease Knowledge Portal (https://cvd.hugeamp.org/). Raw and processed next-generation sequencing data have been deposited at the NCBI Gene Expression Omnibus under accession no. GSE225336. The GRCh37 data are publicly available at https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/. The GRCh38 data are publicly available at https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/. The Genotype-Tissue Expression v.8 datasets are publicly available at https://www.gtexportal.org/home/datasets. All other data are contained in the article and its supplementary information or are available upon reasonable request from the corresponding authors. Source data are provided with this paper.

Code availability

The code used to download, quality-control and train the machine learning models is available at https://github.com/broadinstitute/ml4h under an open-source BSD license.

References

55. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).

56. Puyol-Antón, E. et al. Automated quantification of myocardial tissue characteristics from native T$_1$ mapping using neural networks with uncertainty-based quality-control. *J. Cardiovasc. Magn. Reson.* **22**, 60 (2020).

57. Huang, G., Liu, Z., Pleiss, G., van der Maaten, L. & Weinberger, K. Q. Convolutional networks with dense connectivity. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 8704–8716 (2022).

58. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention* (eds Navab, N. et al.) 234–241 (Springer, 2015).

59. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition* (eds Flynn, P. & Mortensen, E.) 248–255 (IEEE, 2009).

60. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://doi.org/10.48550/arXiv.1412.6980 (2017).

61. Petersen, S. E. et al. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort. *J. Cardiovasc. Magn. Reson.* **19**, 18 (2017).

62. Inker, L. A. et al. Estimating glomerular filtration rate from serum creatinine and cystatin C. *N. Engl. J. Med.* **367**, 20–29 (2012).

63. Stone, N. J. et al. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J. Am. Coll. Cardiol.* **63**, 2889–2934 (2014).

64. Grundy, S. M. et al. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J. Am. Coll. Cardiol.* **73**, 3168–3209 (2019).

65. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **46**, 1734–1739 (2017).

66. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

67. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

68. Boughton, A. P. et al. LocusZoom.js: interactive and embeddable visualization of genetic association study results. *Bioinformatics* **37**, 3017–3018 (2021).

69. Loh, P.-R. et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).

70. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

71. Pirruccello, J. P. et al. Deep learning of left atrial structure and function provides link to atrial fibrillation risk. Preprint at *medRxiv* https://doi.org/10.1101/2021.08.02.21261481 (2021).

72. Khurshid, S. et al. Deep learning to predict cardiac magnetic resonance-derived left ventricular mass and hypertrophy from 12-lead ECGs. *Circ. Cardiovasc. Imaging* **14**, e012281 (2021).

73. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

74. Barbeira, A. N. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).

75. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

76. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

77. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).

78. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

79. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

80. Quinlan, A. R. BEDTools: the Swiss-Army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12.1–11.12.34 (2014).

81. Hansen, K. D., Irizarry, R. A. & Wu, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**, 204–216 (2012).

Author contributions

V.N., J.W.C., P.T.E. and S.A.L. conceived the study. M.D.R.K. and P.D.A. downloaded and prepared the cMRI data. V.N., M.D.R.K., P.D.A. and J.W.C. performed the quality control. M.D.R.K. trained the machine learning models. V.N., M.D.R.K. and P.D.A. performed the main analyses. M.C.H. performed the in vitro experiments. V.N., M.D.R.K., P.D.A., J.W.C., M.C.H., P.T.E. and S.A.L. wrote the paper. J.P.P., L.-C.W., V.N.M., S.H.C., S.K., S.F.F., M.N., C.R., K.N., A.A.P. and P.B. contributed to the analysis plan or provided critical revisions.

Competing interests

Additional information

nature portfolio

Corresponding author(s): Steven A. Lubitz, MD, MPH

Last updated by author(s): 02/28/2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Code used to ingest, quality control, and train machine learning models are available at https://github.com/broadinstitute/ml4h under an open-source BSD license.
Data analysis	All statistical tests were performed using R version 4.0.2 (R Foundation for Statistical Computing, Vienna, Austria)(R, Core Team 2020). PLINK 2.0 was used to perform genome wide association analysis. PLINK 1.9 was used to calculate in-sample linkage disequilibrium (LD) parameters. LD score regression analysis and genetic correlation were performed using ldsc version 1.0.0. Regional association plots were generated with LocusZoom (version 1.4). SNP-heritability was assessed using BOLT-REML v2.3.4. Transcriptome wide association analysis was performed using S-PrediXcan on Python 3.6. Total RNA was extracted using the Direct-zol RNA Miniprep kit (Zymo Research, R2051). Standard RNA-seq libraries were generated and sequenced by Genewiz on an Illumina NovaSeq. Sequenced reads were aligned to the human genome (GRCh38) using Salmon (version 1.8.0). Differential expression analysis was performed with DESeq2 (1.30.1). Transposed DNA was purified with a Qiagen PCR MinElute kit (Qiagen 28004) and final ATAC-seq libraries were purified with a 1.8X SPRI purification using SPRISelect beads (Beckman Coulter) following PCR amplification. Libraries were sequenced on an Illumina Nextseq 500. Reads were mapped to the human genome (GRCh38) using Bowtie2 with default paired-end settings (version 2.3.4.3). Motif enrichment analysis was performed with Homer (version 4.10) using the findMotifsGenome.pl function with default parameters. Peak calling was carried out with MACS2 (callpeak—nomodel— broad) (version 2.2.6) using all ATAC-seq libraries as input. Quantile normalization of ATAC-seq data sets was performed with CQN (version 1.36.0),91 and offsets were fed into DESeq2 to quantify differential accessibility.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

UK Biobank data are made available to researchers from research institutions with genuine research inquiries, following IRB and UK Biobank approval. Genome-wide association analysis summary statistics are available from the Downloads page of the Cardiovascular Disease Knowledge Portal (broadcvdi.org). Raw and processed next-generation sequencing data have been deposited at the NCBI Gene Expression Omnibus under accession number GSE225336. Genome Reference Consortium Human Build 37 (GRCh37) data is publicly available at https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/. Genome Reference Consortium Human Build 38 (GRCh38) data is publicly available at https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/. Genotype-Tissue Expression (GTEx) version 8 datasets are publicly available at https://www.gtexportal.org/home/datasets. All other data are contained within the article and its supplementary information, or are available upon reasonable request to the corresponding author.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences   ☐ Behavioural & social sciences   ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The study sample constituted all UK Biobank participants who underwent cardiac T1 mapping during the first imaging visit (N=42,654).
Data exclusions	We excluded from our analysis participants who had cardiac T1 maps that did not pass our image quality control step.
Replication	Replication of our GWAS loci was not possible as there is no current sizeable biorepository with combined genomic and cardiac T1 mapping data, a highly specialized MRI sequence that is not routinely performed in clinical practice. To provide further data supporting the biologic relevance of our GWAS loci to fibrosis, we interrogated our GWAS loci using a cellular-based TGFB1 fibrosis assay.
Randomization	This is a genetic analysis of cardiac fibrosis in a prospective observational cohort and thus randomization of treatment or exposure was not performed and was not applicable to the study design.
Blinding	This is a genetic analysis of cardiac fibrosis in a prospective observational cohort and treatment or exposure blinding was not performed nor applicable to the study design.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Cell line source(s)	We used primary human cardiac fibroblasts (ACBRI5118, Cell Systems)
Authentication	None
Mycoplasma contamination	Not tested. Procured from Cell Systems at low passage number.
Commonly misidentified lines (See ICLAC register)	NA

Human research participants

Population characteristics	The study sample included 41,505 UK Biobank participants who underwent cardiac magnetic resonance imaging with T1 mapping during the first imaging visit and passed our imaging quality control filters. The mean age of the participants was 64.0 ± 7.7 years and 48.1% were men. 39,339/41,505 had imputed genomic data avaliable in the UK Biobank.
Recruitment	The UK Biobank is a prospective cohort of 502,629 individuals from the UK enrolled between 2006-2010. Around 9.2 million individuals 40-69 years old living in England, Scotland, and Wales were invited to participate in the study and 5.4% agreed to participate. Extensive questionnaire data, physical measures, and biological samples were collected at baseline, with ongoing data collection in large subsets of the cohort, including repeated assessments and multimodal imaging. Starting in 2014, 42,654 participants have returned for the first multi-modal imaging visit including cardiac magnetic resonance imaging with T1 mapping allowing for the assessment of myocardial interstitial fibrosis.
Ethics oversight	Use of UK Biobank data was performed under application number 7089 and was approved by the local Massachusetts General Hospital institutional review board. All participants signed informed consent prior to participation in the UKB.

Note that full information on the approval of the study protocol must also be provided in the manuscript.