nature communications

Article

https://doi.org/10.1038/s41467-026-70379-2

An integrated germline and somatic genomic model for coronary artery disease

Received: 21 February 2025

Accepted: 16 February 2026

Published online: 26 March 2026

Check for updates

A list of authors and their affiliations appears at the end of the paper

Multiple germline and somatic genomic factors are associated with risk of coronary artery disease, but there is no single measure of risk that integrates all information from a DNA sample. To address this gap, we develop an integrated genomic model that includes six germline and somatic genetic drivers for coronary artery disease, including polygenic risk score, genetically-proxied proteomic/metabolomic risk scores, and clonal hematopoiesis of indeterminate potential. We evaluated its predictive power in the UK Biobank (N = 391,536), and validate it using data from the TOPMed program (N = 34,177). The 10-year coronary artery disease risk based on the integrated genomic model profile ranges from 1.1% to 15.5% in the UK Biobank and from 3.8% to 33.0% in TOPMed, with a more pronounced gradient in males than females. The integrated genomic model captures the cumulative effect of multiple genetic drivers, identifying individuals at high risk for coronary artery disease despite lacking any single high-risk genetic factor, as well as individuals at low risk despite carrying known high-risk factors. In middle age, the integrated genomic model augments the performance of the Pooled Cohort Equations, a clinical risk calculator for coronary artery disease. While the integrated genomic model yields only modest incremental predictive value over polygenic risk score at the population level, it identifies approximately 13% of high-risk individuals not detected by polygenic risk score alone.

The coronary artery disease (CAD) remains the number one cause of global mortality and morbidity[1]. The early identification of individuals at high risk for CAD is a fundamental strategy in preventing disease. Utilizing DNA information for CAD risk prediction has gained traction due to its ability to identify early-onset cases, predict risk earlier in life, and augment the performance of existing clinical risk measures[2,3]. Significant progress has been made in understanding germline genomic risk drivers of CAD. Both monogenic drivers of risk such as pathogenic variants in familial hypercholesterolemia (FH)-related genes (*LDLR*, *APOB*, and *PCSK9*) and polygenic risk scores (PRS) have shown promise in CAD risk stratification, individually and in combination[4,5]. Moreover, age-related somatic mutations have been associated with increased CAD risk, attributed to clonal hematopoiesis of indeterminate potential (CHIP) and shortened leukocyte telomere length (LTL)[6,7]. Even though germline and somatic genomic variations shape CAD risk, they remain to be studied in aggregate. A single model leveraging all available information from a single DNA biopsy could have the potential to improve risk prediction of CAD.

A comprehensive genomic risk model for CAD would integrate risk from early-life germline mutations with risk from somatic mutations occurring later in life. We previously demonstrated the interplay between monogenic and polygenic risk, highlighting that polygenic background alters the penetrance of FH variants[4]. Thereafter, Zhao et al. reported that a combination of germline and somatic mutations augments the risk of CAD, as evidenced by the interaction between PRS and CHIP[8]. This mounting evidence underscores that an ensemble model that integrates all known germline and somatic risk drivers and their interactions might improve genomic risk prediction of CAD.

As genomic medicine moves towards clinical adoption, it might be beneficial that a single measure of risk is communicated using the

e-mail: pnatarajan@mgh.harvard.edu; wangmx@big.ac.cn; afahed@mgh.harvard.edu

entirety of the data points available from an individual's genome. Mixed information about risk is poised to confuse people unless integrated into a single number. This concept has long been established in clinical risk prediction. For example, a single 10-year cardiovascular risk is provided by integrating information from factors such as blood pressure, cholesterol levels, smoking, and diabetes, each of which might indicate low or high risk for an individual[9]. Similarly, in genomic risk prediction, an individual might have a monogenic FH variant, a low polygenic risk score, no CHIP variant, and a short LTL, challenged by how to interpret this complex combination of genetic risk factors. It is only helpful for the individual in this context to understand the summed effect and risk estimate.

We developed an integrated genomic model (IGM) that combines multiple genetic drivers into a single score to maximize the precision of CAD risk prediction. We demonstrated the accuracy, calibration, and added value of this all-at-once model using half a million people from the UK Biobank, and validated its performance in studies contributing to the Trans-Omics for Precision Medicine (TOPMed) program.

Results

Developing an Integrated Genomic Model for CAD

We used the UK Biobank as a discovery cohort, including 391,536 individuals (mean [SD] age, 56.5 [8.1] years; 53.8% women) including 28,346 (7.2%) participants who developed CAD over a median follow-up of 12.3 (interquartile range [IQR], 1.6) years. In the UK Biobank, 94.1% of participants were White (n = 368,296), with smaller proportions identifying as Asian (2.3%, n = 9106), Black (1.6%, n = 6237), and Other ancestry (2.0%, n = 7897). In contrast, TOPMed showed greater diversity, with 68.3% White (n = 23,333), 25.9% Black (n = 8858), 2.2% Asian (n = 737), and 3.7% Other ancestry (n = 1250). Whole-genome sequencing data was used to curate and compute features of risk previously shown to be associated with CAD. Specifically, we included two somatic features—CHIP and LTL—and four germline features—FH variants, CAD PRS, a proteome PRS (ProPRS) and a metabolome PRS (MetPRS). Of note, ProPRS and MetPRS were constructed using a series of genetic proxies for protein and metabolite levels derived from the atlas of genetic scores that predict multi-omic traits[10], rather than relying on actual measurement of protein and metabolite levels. Using this feature matrix, we developed a somatic risk score, a germline risk score, and an integrated genomic model (IGM) (Fig. 1).

We then validated the IGM in 34,177 individuals from the TOPMed program (mean [SD] age, 62.6 [10.6] years; 66.0% women) (Supplementary Table 3). Incident CAD events occurred in 3972 (14.3%) participants who developed CAD over a median follow-up of 10.5 (IQR, 8.6) years. The IGM model provided individual 10-year risk estimates across percentiles of somatic and germline risk (Fig. 1). Further details on baseline characteristics by genetic drivers are presented in Supplementary Tables 4-8.

Germline and somatic genomic drivers of CAD risk

We first estimated the individual risk of prevalent and incident CAD conferred by each of the two somatic and four germline drivers. When evaluating the association of the germline drivers with prevalent CAD in the UK Biobank, FH variant carriers had a three-fold increase in risk—odds ratio (OR) of 3.08 (95% confidence interval [CI], 2.46–3.85;

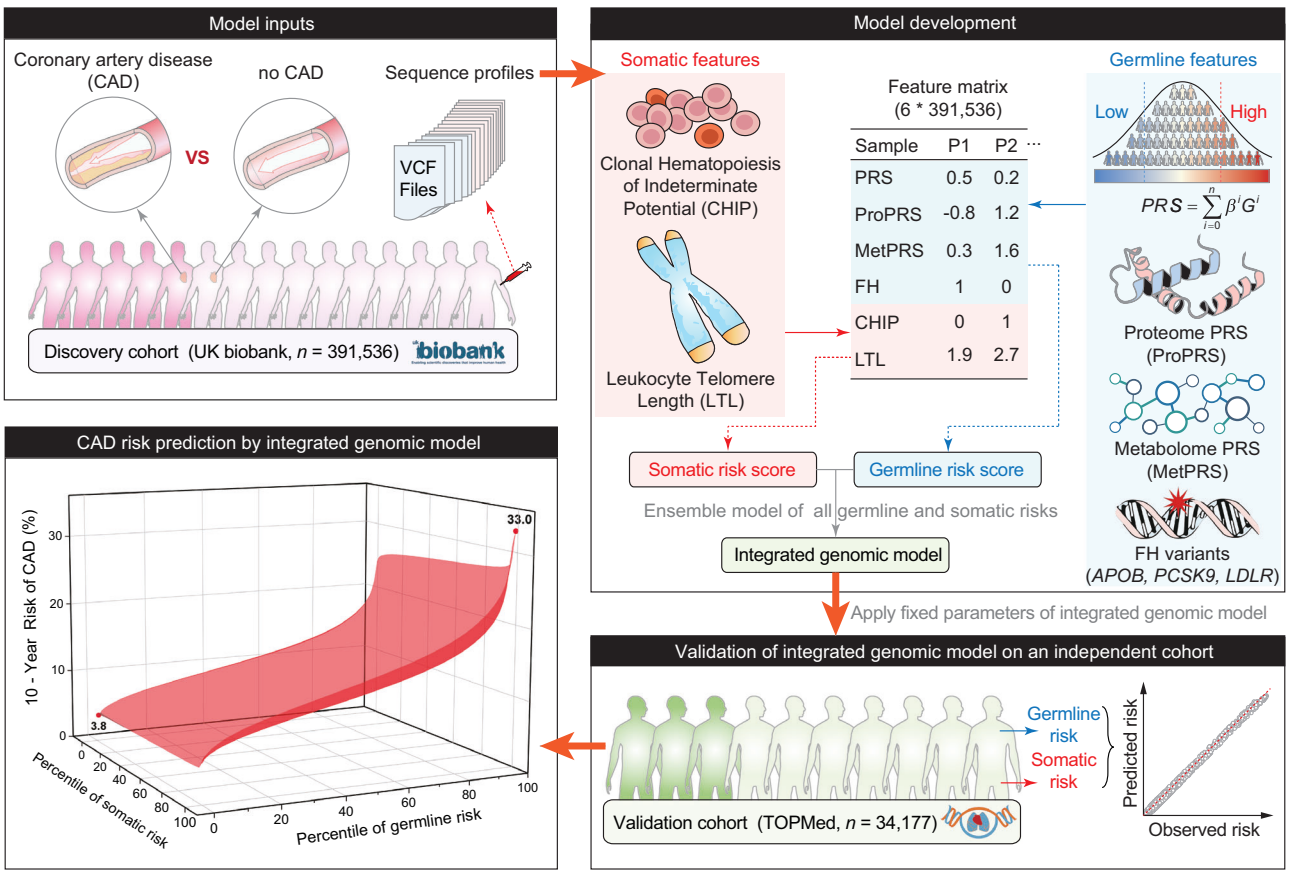


Fig. 1 | Overview of development and validation of integrated genomic model (IGM). Based on sequencing data from UK Biobank (N = 391,536), we curated 6 genomic features that are associated with the risk of CAD. Scores of somatic and germline risks were ensembled to construct IGM, which was then validated in the TOPMed cohort (N = 34,177).

https://doi.org/10.1038/s41467-026-70379-2

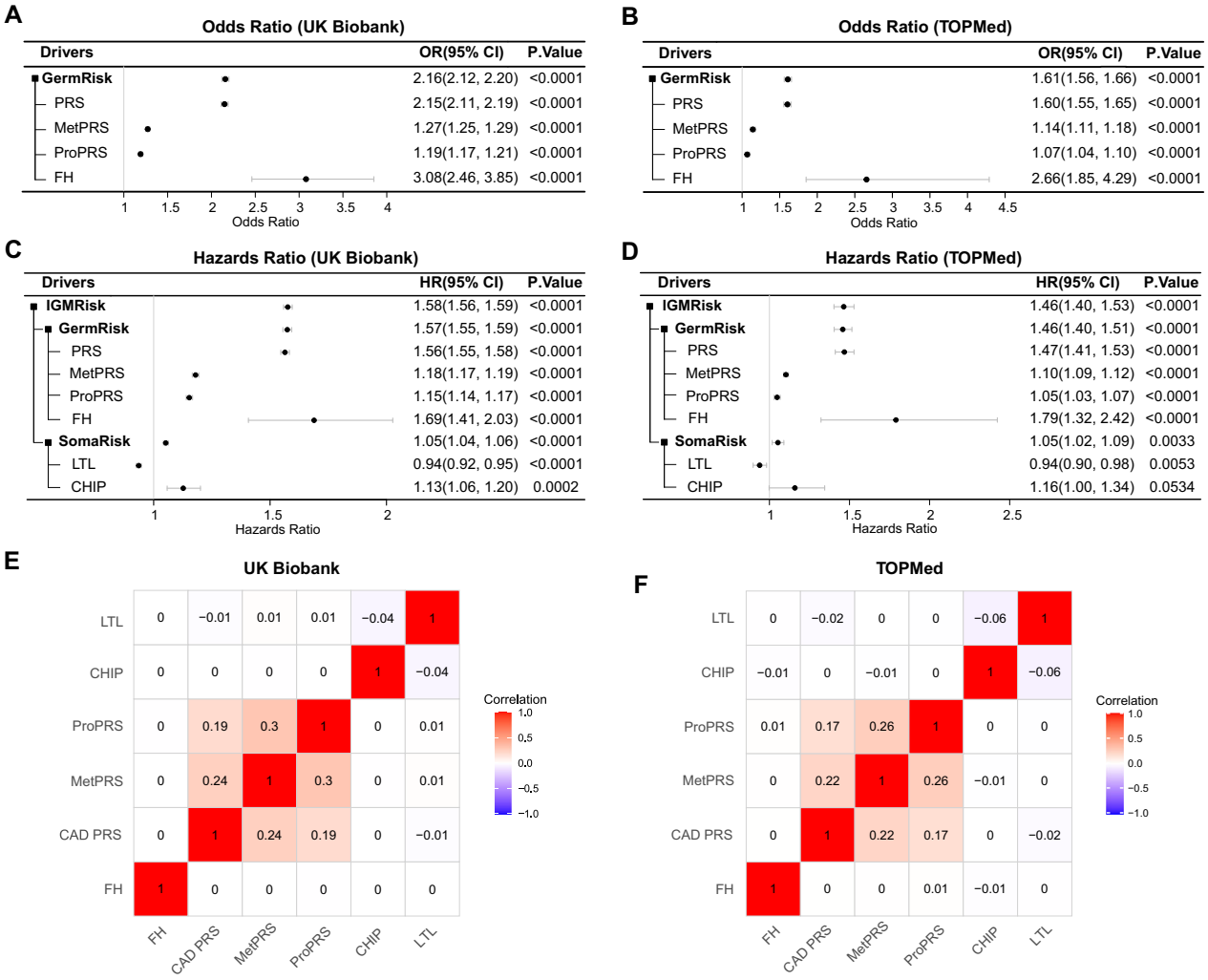


Fig. 2 | Effect size of genomic drivers in UK Biobank and TOPMed. Effect sizes based on prevalent coronary artery disease (CAD) in UK Biobank ($N = 391,536$) (**A**) and TOPMed ($N = 34,177$) (**B**), and incident CAD in UK Biobank ($N = 377,280$) (**C**) and TOPMed ($N = 27,520$) (**D**) are presented. Odds ratio and hazard ratio per standard deviation are shown for all continuous measures (CAD PRS, MetPRS, ProPRS, LTL, GermRisk, SomaRisk, IGMRisk). Odds ratio and hazard ratio per carrier status are shown for FH and CHIP. Estimates are derived from logistic regression (**A**, **B**) or Cox proportional hazards model (**C**, **D**) with sex, recruitment age, and the first 10 principal components of genetic ancestry as covariates, *p*-values are calculated using two-sided Wald tests. GermRisk is a weighted combination of four genetic drivers (CAD PRS, MetPRS, ProPRS, and FH) as a single predictor, with weights estimated by a logistic regression model. SomaRisk is a weighted combination of two somatic drivers (CHIP and LTL) as a single predictor, with weights estimated by a Cox proportional hazards model. IGMRisk is a combination of GermRisk and SomaRisk, weighted from a Cox proportional hazards regression model estimation, *p*-values are calculated using two-sided Wald tests. Correlations among the six genetic drivers are shown for the UK Biobank (**E**) and TOPMed (**F**). CI confidence interval, PRS polygenic risk score for CAD, MetPRS metabolome PRS, ProPRS proteome PRS, LTL leukocyte telomere length, FH familial hypercholesterolemia variants, CHIP clonal hematopoiesis of indeterminate potential. Error bars: 95% CI. Source data are provided as a Source Data file.

$p < 0.001$). The CAD PRS (OR per standard deviation (SD), 2.15; 95% CI, 2.11–2.19; $p < 0.001$), MetPRS (OR per SD, 1.27; 95% CI, 1.25–1.29; $p < 0.001$), and ProPRS (OR per SD, 1.19; 95% CI, 1.17–1.21; $p < 0.001$) were also significantly associated with CAD (Fig. 2). We combined these four germline risk factors into a single predictor called GermRisk. The OR per SD for GermRisk was 2.16 (95% CI, 2.12–2.20; $p < 0.001$), and individuals in the top quintile of GermRisk had 3.6-fold increase in risk compared to everyone else (95% CI, 3.47–3.73; $p < 0.001$) (Supplementary Table 8). The effect sizes of genetic drivers for prevalent CAD in TOPMed were overall consistent with those of UK Biobank (Fig. 2B; Supplementary Table 9).

We then evaluated the association of genomic drivers with incident CAD in the UK Biobank, and demonstrated strong associations of each of the germline and somatic drivers with incident CAD (Fig. 2C; Supplementary Table 10). For germline drivers, the HR for FH carriers was 1.69 (95% CI 1.41–2.03, $p < 0.001$) and HRs per SD for CAD PRS, MetPRS, and ProPRS were 1.56 (95% CI 1.55–1.58, $p < 0.001$), 1.18 (95% CI 1.17–1.19, $p < 0.001$), and 1.15 (95% CI 1.14–1.17, $p < 0.001$), respectively. For somatic drivers, CHIP and LTL were associated with CAD with HR of 1.13 (95% CI 1.06–1.20, $p < 0.001$) and 0.94 (95% CI 0.92–0.95, $p < 0.001$), respectively. We combined these two somatic risk factors into a single predictor called SomaRisk, similar to GermRisk. Both GermRisk and SomaRisk demonstrated a strong association with incident CAD—HR per SD of 1.57 (95% CI 1.55–1.59, $p < 0.001$) and 1.05 (95% CI 1.04–1.06, $p < 0.001$), respectively (Fig. 2C; Supplementary Table 10). The effect sizes of genetic drivers for incident CAD in TOPMed were mostly consistent with those of UK Biobank (Fig. 2D; Supplementary Table 11).

Integrated genomic model to predict CAD risk

We assessed pairwise correlations between six genetic variables from the UK Biobank and TOPMed studies using Pearson correlation

https://doi.org/10.1038/s41467-026-70379-2

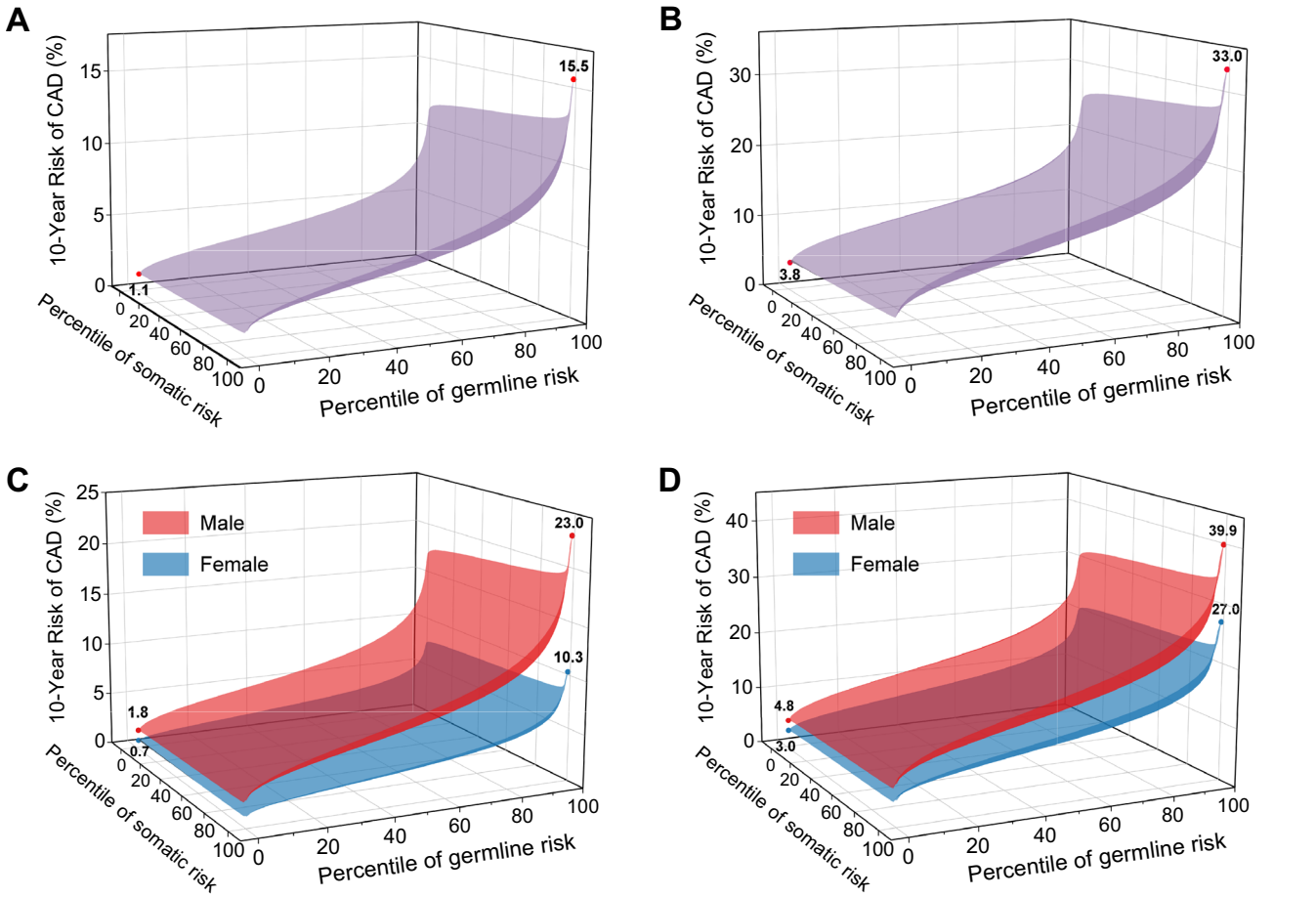


Fig. 3 | Ten-year risk of CAD as a function of somatic and germline risk from the integrated model. Ten-year risk of CAD among participants from UK Biobank ($N = 377,280$) (**A**) and TOPMed ($N = 27,520$) (**B**), and sex-stratified 10-year risks of CAD in UK Biobank (**C**) and TOPMed (**D**) are presented. Source data are provided as a Source Data file.

coefficients to evaluate multicollinearity before combining them in a Cox proportional hazards model. This allowed us to gauge overlapping signals, particularly between CAD PRS, MetPRS, and ProPRS. The correlations among six drivers were weak (≤ 0.3) (Fig. 2E, F), reassuring that each driver contributes distinct signals.

To obtain a comprehensive assessment of a person's CAD risk, we used an integrated genomic model (IGM) to quantify the risk from both germline and somatic drivers—a combined predictor of GermRisk and SomaRisk (Supplementary Fig. 4). The IGM risk was significantly associated with the risk of incident CAD (HR per SD, 1.58; 95% CI, 1.56–1.59; $p < 0.001$), and the effect size was consistent when validated in the TOPMed external data set (HR per SD, 1.46; 95% CI, 1.40–1.53; $p < 0.001$) (Fig. 2C, D; Supplementary Tables 11 and 12).

Joint modeling of germline and somatic drivers indicated substantial gradients in risk of CAD, driven by inherited DNA variants, variation in the rate of LTL shortening, and accumulation of somatic variants leading to CHIP. For the sex-combined estimation of the 10-year risk of CAD in UK Biobank, individuals in the lowest germline and somatic risk percentile have a 10-year risk as low as 1.1%, while those in the highest germline and somatic risk percentile have a 10-year risk as high as 15.5% (Fig. 3A). A similar gradient in risk across germline and somatic variation was observed in TOPMed, ranging from 3.8% to 33.0% (Fig. 3B). For a sex-stratified analysis in the UK Biobank, male individuals had a 10-year risk that ranged from 1.8% to 23.0% across the germline and somatic risk spectrum. This was about 2.3 times higher than the risk spectrum for females, which spanned from 0.7% to 10.3% (Fig. 3C). Large gradients in 10-year risk were consistently observed in the TOPMed for the sex-stratified analysis, with males ranging from 4.8% to 39.9% and females ranging from 3.0% and 27.0% (Fig. 3D).

Heterogeneity of genomic risk profiles captured by the integrated genomic model

The IGM effectively captured a range of genetic risk combinations for CAD, identifying high-risk groups (top 20% overall risk) with diverse genetic profiles in the UK Biobank (Fig. 4A). High-risk IGM group included individuals at high risk in both germline and somatic factors (20.6%), those at high risk for one of the factors (78.6%), and a small proportion with moderately elevated, yet sub-threshold, risks for both factors (0.8%) (Fig. 4A). Individuals at high risk for both germline and somatic factors had the highest 10-year risk (8.8%, 95% CI 8.43–9.19%) within the high-risk IGM group (Supplementary Table 12). As expected, the high-risk group identified by the IGM had a larger number of genetic risk drivers compared to the low-risk group. For example, 63.9% had two or more genetic drivers in the high-risk group, compared to only 3.4 % in the low-risk group in the UK Biobank (Fig. 4B). Notably, people at low IGM risk were not without genetic risk drivers, and a non-negligible proportion of 29.3% had one or more genetic risk drivers (Fig. 4B and Supplementary Table 13); however, mitigating effects from other genetic variants (e.g., low CAD PRS) seem to offset the overall risk, thus classifying these individuals as low risk. Similar proportions of breakdown and distribution patterns were shown in TOPMed (Fig. 4C, D and Supplementary Table 13).

Table 1 presents the detailed genetic profiles of ten representative individuals from the high-risk IGM group. Notably, participants

https://doi.org/10.1038/s41467-026-70379-2

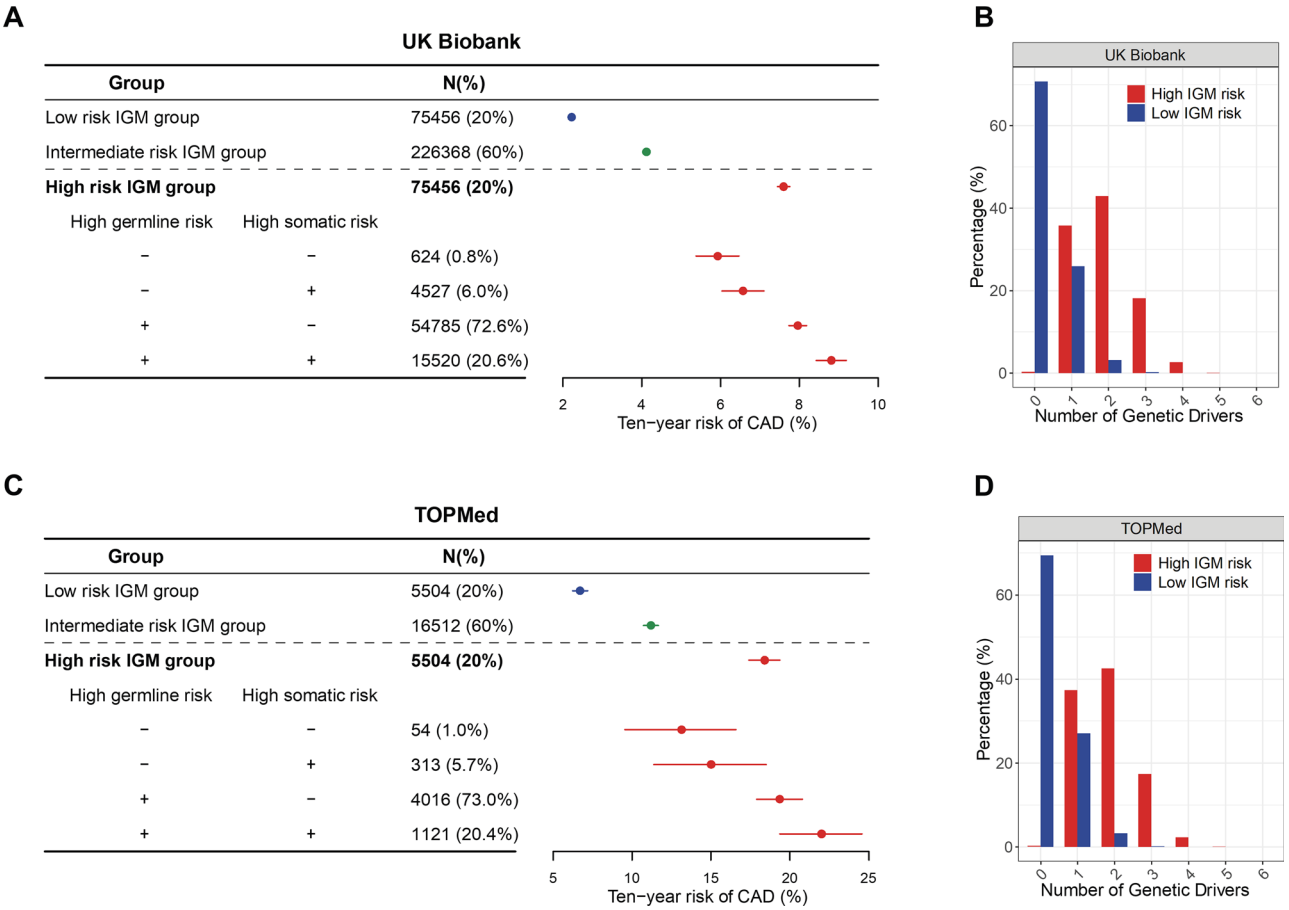


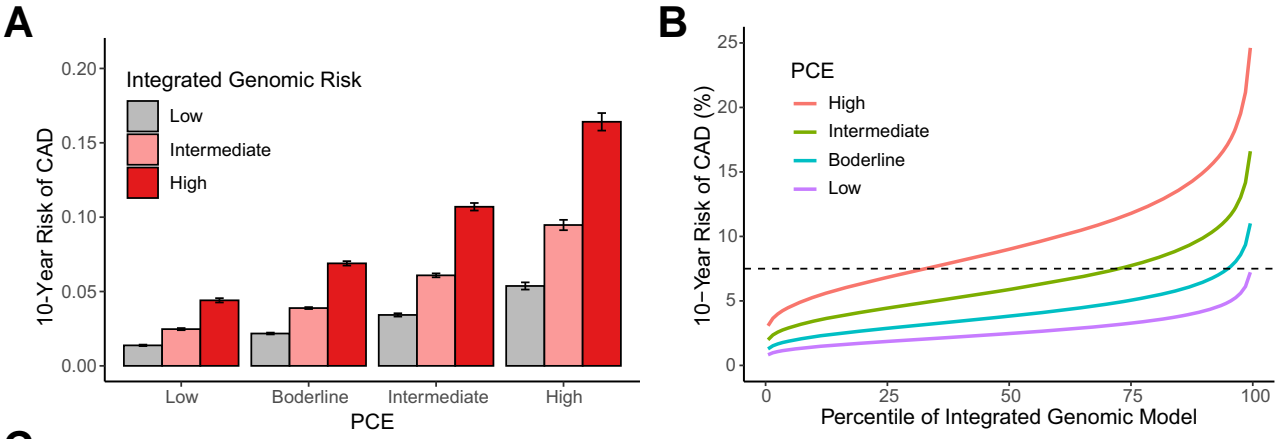
Fig. 4 | Decomposing integrated genomic risk reveals varied genetic paths to elevated CAD susceptibility. The integrated genomic model (IGM) captures participants with diverse genetic risk profiles contributing to risk in the UK Biobank (**A**, **B**) and TOPMed (**C**, **D**). (**A**) and (**C**), The ten-year risk of CAD was estimated for each IGM risk group in the UK Biobank and TOPMed dataset. Categories were defined as Low-risk (bottom 20%), Intermediate risk (middle 60%), and high-risk group (top 20%). In the high-risk group, the genetic profile was further partitioned by the status of carrying a high germline risk or high somatic risk. The high germline risk was defined as the top 20% with a composite risk estimated from four germline genetic risk drivers (FH, CAD PRS, MetPRS, and ProPRS). The high somatic risk was defined as the top 20% with a composite risk estimated from two somatic risk drivers (CHIP and LTL). (**B**) and (**D**), the genetic risk profiles for the IGM Low-risk (bottom 20%) and high-risk (top 20%) groups, respectively. The six drivers are FH, CAD PRS, MetPRS, ProPRS, CHIP, and LTL. The continuous variables were binarized, with individuals in the top 20% treated as carriers, except for LTL (the bottom 20% were treated as carriers). PRS polygenic risk score (PRS) for CAD, MetPRS metabolome PRS, ProPRS proteome PRS, LTL leukocyte telomere length, FH familial hypercholesterolemia variants, CHIP clonal hematopoiesis of indeterminate potential. Error bars: 95% CI. Source data are provided as a Source Data file.

Table 1 | Genetic profiles for individuals in high-risk IGM group

	PRS (%)[a]	MetPRS (%)[a]	ProPRS (%)[a]	FH carrier[b]	CHIP variant carrier[b]	LTL (%)[a]	Overall CAD risk (IGM %)
Participant 1	34.4	75.6	73.0	1	0	14.5	83.4
Participant 2	44.4	91.3	79.3	1	0	54.8	89.6
Participant 3	51.6	93.8	91.7	0	1	28.6	84.9
Participant 4	75.3	23.7	25.8	0	1	96.7	81.5
Participant 5	77.9	30.5	62.7	1	0	93.8	95.1
Participant 6	65.6	94.1	93.6	0	0	0.0	86.2
Participant 7	73.3	3.9	87.4	0	1	71.9	85.8
Participant 8	95.9	26.4	85.4	1	0	83.8	99.7
Participant 9	46.4	21.0	64.4	1	0	56.1	83.8
Participant 10	88.5	99.3	97.9	0	0	63.4	94.1

[a]A higher percentile corresponds to a greater risk level of the genetic driver.
[b]A value of 1 indicates a carrier, and 0 indicates a non-carrier; High-risk IGM group indicate top 20% in IGM score.
PRS polygenic risk score (PRS) for CAD, *MetPRS* metabolome PRS, *ProPRS* proteome PRS, *LTL* leukocyte telomere length, *FH* familial hypercholesterolemia variants, *CHIP* clonal hematopoiesis of indeterminate potential.

https://doi.org/10.1038/s41467-026-70379-2



C

		PCE + IGM		
	PCE	**≤7.5 %**	**>7.5 %**	**Total No. (%) of Participants**
CAD Events	**≤7.5 %**	6965	1858	8823(42.8)
	>7.5 %	1452	10349	11801(57.2)
	Total No. (%) of Participants	8417(40.8)	12207(59.2)	20624 (100)
CAD Nonevents	**≤7.5 %**	252144	16708	268852 (75.4)
	>7.5 %	22954	64850	87804 (24.6)
	Total No. (%) of Participants	275098 (77.1)	81558 (22.9)	356656 (100)

Fig. 5 | The integrated genomic model provides additional risk discrimination within clinically defined risk groups. Stratification (**A**, **B**) and reclassification (C) of 10-year predicted CAD risk based on IGM in UK Biobank ($N = 377,280$). **A** Stratification by IGM risk within PCE risk stratum, error bars: 95% CI. **B** Predicted 10-year CAD risk gradient by genetic risk percentile. **C** Reclassification of 10-year predicted CAD risk. Columns and rows indicate categories of 10-year predicted risk, with the number of individuals in each risk category shown. The number of correctly and incorrectly reclassified samples are highlighted in dark and light blue, respectively. A continuous net reclassification index was 33.45% (95% CI, 32.11–34.76%). IGM categories were defined as Low-risk (bottom 20%), Intermediate risk (middle 60%), and high-risk group (top 20%), respectively. PCE categories were defined as low (estimated risk less than 5%), borderline (risk between 5% to 7.5%), intermediate (risk between 7.5% and 20%), and high (risk greater than 20%), respectively. IGM integrated genomic model, PCE pooled cohort equations. Source data are provided as a Source Data file.

demonstrate diverse combinations of genetic risk factors contributing to their elevated CAD risk. For instance, Participant 1 exhibited below average PRS levels (34.4%) but had substantially elevated MetPRS (75.6%), ProPRS (73.0%), and carried an FH variant, culminating in a high overall CAD risk (83.4%). Conversely, Participant 9 had relatively modest or average genetic risk profiles in PRS (46.4%), MetPRS (21.0%), and LTL (56.1%), and this individual's high CAD risk was primarily influenced by positive FH carrier status. These individual-level examples underscore how the cumulative and multidimensional nature of the IGM can capture high-risk individuals who might otherwise remain unrecognized when considering single genetic factors.

Integrated genomic model and clinical risk

The American College of Cardiology/American Heart Association, Pooled Cohort Equations (PCE) are a guideline-recommended clinical-risk calculator that uses clinical risk factors to identify high-risk people for initiation of preventive treatments (i.e., statin)[11]. Within each of the guideline-defined strata of the PCE risk, the IGM score was a strong predictor of coronary artery disease events and showed a consistent risk gradient by IGM score category (Fig. 5A, B). Among participants at high PCE risk, the 10-year CAD event rates were 5.4%, 9.5%, and 16.4%, for low, intermediate, and high IGM risk groups (Fig. 5A). Remarkably, a 10-year CAD risk threshold of 7.5% for initiating statin therapy as guideline recommended was reached even among individuals in the borderline (5.0–7.4% 10-year risk) PCE category when stratified by IGM percentiles (Fig. 5B). A close match of the model predicted and actual observed 10-year disease risk shows the models were well calibrated (Supplementary Fig. 5). When disaggregated to individual genetic risk drivers, the FH variants and CAD PRS most prominently re-stratified the CAD risk across PCE categories, but other risk factors also have significant stratification ability (Supplementary Fig. 6).

By combining the conventional clinical risk of PCE with the genetic risk of IGM, the model showed the most potent risk stratification ability. When 20,624 individuals who experienced CAD events in the UK Biobank were used to determine reclassification by adding IGM to PCE compared to PCE alone, 1858 were correctly classified at a higher risk, while 1452 were incorrectly placed at a lower risk (Fig. 5C), leading to a net proportion of accurate reclassifications for events is 1.97% (406/20,624). For nonevents, 22,954 individuals were correctly down-classified, and 16,708 were incorrectly up-classified, leading to a net reclassification proportion of 1.75% (6246/356,656) for nonevents. The overall NRI combines the events and nonevents, resulting in 3.72%

(95% CI, 3.15–4.26%). Continuous NRI was 33.45% (95% CI, 32.11–34.76%). As shown in the full stratification table, there was a lower event rate in the Low-risk category for the combined model than in the PCE-alone model (1.9% versus 2.3%), however, a higher event rate was observed for the High-risk category (27.1% versus 23.5%), demonstrating the prominent risk stratification ability of the IGM model with guideline-recommended PCE risk estimator (Supplementary Table 14).

We evaluated the discrimination of PCE, IGM, and combined model compared to a baseline model with age, sex and genetic ancestry. The C-statistic for the base model was 0.701 (95% CI, 0.698–0.704), PCE was 0.725 (95% CI 0.722–0.728), and the combination of base and IGM model was 0.734 (95% CI 0.731–0.737), respectively, in UK Biobank (Supplementary Fig. 7; Supplementary Table 15). The performance was highest when combining PCE and IGM (C-statistic, 0.750; 95% CI, 0.747–0.753). In the TOPMed data, the discrimination with IGM was further improved when limiting prediction to younger individuals (aged ≤ 45 years) (C-statistic 0.805, 95% CI, 0.699–0.913) (Supplementary Fig. 7 and Supplementary Table 15).

Finally, we compared IGM directly against a clinical biomarker model, which included clinical biomarkers—Lp(a), CRP, and creatinine. We deliberately excluded LDL cholesterol from the clinical biomarker model, as total cholesterol and HDL cholesterol are already integral components of the PCE, which served as our primary benchmark comparator for the IGM. The result demonstrated that the clinical biomarker model (Lp(a), CRP, and creatinine; C-statistic 0.708) did not outperform the IGM (C-statistic 0.734), underscoring the incremental predictive value provided by genomic information (Supplementary Fig. 8).

Comparison of genetically proxied and measurement based risk scores

To compare CAD risk score based on genetically proxied and measured proteomic/metabolomic scores, we further constructed proteome risk score (ProRS) and metabolome risk score (MetRS) using large-scale proteomic data from UK Biobank. The ProRs and MetRS were strongly associated with CAD risk (ProRS: OR per SD, 2.77; 95% CI, 2.64–2.91; MetRS: OR per SD, 2.05; 95% CI, 2.01–2.10), recapitulated finding from Helgason et al.[12] that demonstrated the effectiveness of proteomic risk scores in predicting CAD risk. We next compared the correlation between measured and genetically proxied levels for the 124 proteins that were used to develop ProPRS. The correlations were generally variable (Supplementary Fig. 9), likely influenced by several factors. Proteomic and metabolomic measurements rely on a single time point and can fluctuate according to current disease status or other temporary conditions at the time of collection. In contrast, genetically proxied scores, derived from personal genetic data, more closely reflect baseline levels that are less confounded by disease or lifestyle factors (ie, diet). Moreover, as acknowledged elsewhere[10], technical elements (serum versus plasma, batch effects, fasting state), demographic variables, and environmental factors (for example, diet) all played a role in the observed variability in proteomic measurement. While comparing measured and genetically proxied molecular levels represents an interesting approach, it is essential to recognize that these measurements derive from different biological contexts, which introduces complexity to direct comparison. The genetically proxied scores may offer unique insights into CAD risk compared to measurement based scores, as they reflect an individual's baseline levels of key molecules linked to CAD risk. Notably, the genetically proxied scores exhibited effect sizes comparable to CHIP—a well-known CAD risk factor—suggesting their potential utility as strong risk factors of CAD risk.

Discussion

We developed an integrated genomic model for CAD prediction that combines multiple known germline and somatic risk drivers that can be measured using a single DNA biopsy. The model demonstrated value in improving the precision of risk estimation and capturing people at risk due to diverse genetic risk profiles. Based on the IGM, the 10-year CAD risk varied from 1.1% to 15.5% among UK Biobank participants and 3.8% to 33.0% in TOPMed study participants, with a more pronounced gradient in males than females for both cohorts. The integrated genomic model showed a high discrimination when combined with clinical risk score or used in younger age groups. The integrated genomic model captured cumulative and comprehensive effects of multiple genetic drivers, identifying high-risk individuals who may otherwise be overlooked when using conventional risk models relying on a single genomic driver (i.e., PRS or FH).

Conventional genomic risk models for CAD that focused on monogenic (FH) or polygenic (PRS) drivers were limited in their uniaxial approach, falling short of encompassing the wide range of genetic combinations present in real-world populations[13]. For instance, individuals carrying FH or CHIP variants might still have a low overall risk of CAD if their effects are offset by protective PRS and low MetPRS. Other individuals may present with different combinations of risk and protective genetic factors that collectively indicate a high CAD risk. Such dynamics complicate CAD risk assessment using standard genomic approaches that rely on stratification by a single genetic driver. Our IGM captured such dynamics by integrating all known genomic risk drivers for CAD, including germline, somatic, and genetically predicted proteomic/metabolomic drivers in a single model, without compromising the population-level performance represented by C-statistic and OR/SD. We successfully demonstrated that IGM captures cumulative and comprehensive effects of multiple genetic drivers, identifying high-risk individuals who do not have obvious germline and somatic risk but whose aggregate genetic risk escalates to a high overall CAD risk (Fig. 4; Table 1). Conversely, a subset of individuals classified in the low risk group by IGM possessed one or more high-risk genetic drivers (Fig. 4), indicating other drivers may confer protective benefits and thus reduce the overall risk. Such a dynamic profile of personal genomics should be considered to fully achieve the goals of precision prevention and personalized care.

The PRS emerged clearly as the primary driver of IGM performance, with additional genetic factors such as FH variants, MetPRS, and ProPRS contributing relatively modest incremental benefits in population-level metrics (i.e., C statistics, OR/SD). Yet, integrating these supplementary genetic elements notably enriched the diversity and inclusivity of CAD risk prediction. Specifically, about 13% of individuals identified by the IGM as high-risk did not exhibit elevated PRS levels (Supplementary Table 16). This indicates that 13% of high-risk individuals would have gone unnoticed if only PRS was used, as their risk is attributed to a buildup of various other genetic risk components. A detailed individual-level inspection (Table 1) underscores how combined genetic influences—beyond PRS alone—drive CAD risk. Notably, the IGM pinpointed individuals with elevated CAD risk through FH variants or increased MetPRS/ProPRS scores, even in the absence of high PRS values. Considering individual-level perspectives alongside population-level metrics provides a clearer understanding of IGM's ability to represent diverse genetic profiles including non-PRS genetic factors. We envision IGM can facilitate a nuanced approach to CAD risk stratification and identify at-risk individuals who would have otherwise been overlooked by conventional genomic models (ie, PRS) that are not designed to capture totality of the genetic combinations.

The integrated genomic model based on comprehensive DNA information is a strong predictor of CAD in young adults, enabling primordial prevention prior to the onset of clinical risk factors. While IGM had modestly higher discriminative capacity for incident CAD compared with the clinical risk score, its predictive accuracy was substantially higher in younger individuals (aged ≤45 years) in the TOPMed program studies (C-statistic 0.805) (Supplementary Table 15). Our findings imply that the use of genetic information to predict future

CAD risk might be more profound for young adults, consistent with previous findings[14,15]. In contrast to clinical risk models, the IGM is available even before clinical risk factors manifest, providing an additional benefit to young adults who often remain undetected on the radar of traditional assessments.

The highest prediction was achieved by combining IGM with clinical risk score, highlighting the value of genetics in complementing clinical risk prediction. The IGM showed ability to further stratify risk within each clinical risk stratum, suggesting it could provide additional information for risk assessment beyond clinical factors alone. Low genetic risk individuals in the high PCE group demonstrated equivalent 10-year CAD risk with average genetic risk individuals in the intermediate PCE group, and a consistent downward trend was observed across all clinical risk strata (Fig. 5A). Conversely, an upward trend was observed for individuals with high genetic risk identified by IGM. Current guidelines recommend initiating statin therapy for individuals predicted to have a 10-year CAD risk of 7.5% or higher[16]. However, our findings indicate that individuals in the borderline PCE category who have a high genetic risk may also warrant targeted interventions, as their 10-year CAD risk is comparable (Fig. 5A). It is noteworthy that individuals at the 33rd, 72nd, and 95th percentiles of integrated genomic model all exhibited an equivalent 10-year CAD risk of 7.5% (Fig. 5B), despite being categorized in high, intermediate, and borderline PCE groups, respectively. This further indicated that existing clinical-focused models might not adequately encompass the multifaceted nature of CAD risk, warranting the consideration of the interplay between genetic and clinical factors in risk evaluations.

While monogenic and polygenic drivers of risk have become well-established, emerging models based on proteomic data are now being developed to predict cardiovascular risk, expanding beyond traditional clinical and genomic models[12,17]. Helgason et al. recently developed a protein risk score based on 4963 plasma proteins from 13,540 Icelanders and demonstrated its reliable predictability for major cardiovascular risk[12]. Nevertheless, implementation of proteome measurement in clinical practice remains challenging due to the cost and feasibility constraints. To make the most of proteomic and metabolomic insights in settings without their direct measurements, we developed proteome and metabolome PRSs, leveraging genetically-predicted protein and metabolite levels instead of actual serum protein levels based on the genetic score atlas for multi-omic traits[10]. We calculated genetic score for 2692 proteins and 876 metabolites levels, with 124 proteins and 142 metabolites comprising the final ProPRS and MetPRS models after Lasso penalty was applied, respectively (Supplementary Fig. 2; Supplementary Data 2-3). Although some correlation was present among the MetPRS, ProPRS and CAD PRS, the magnitude was weak (≤ 0.3) (Fig. 2), implicating that each driver contributes a distinct, non-overlapping signals. We have developed genetically-predicted protein and metabolite risk scores for CAD that leverage readily available genetic data from standard DNA microarray or sequencing platforms, offering a potentially more cost-effective approach than direct serum protein measurements. Our purpose was to make the most out of a single DNA biopsy, which is becoming increasingly feasible through adoption of genomic medicine and large biobanking efforts.

This study has several limitations. First, we have not provided ancestry-specific results given the majority composition with European and smaller sample size for non-European ancestries. However, to promote inclusion and equity, we used multi-ancestry cohorts from UK Biobank and TOPMed in the primary analysis, as well as multi-ancestry PRS that has been demonstrated to perform well for both European and non-European ancestries[13]. Second, this study evaluated CAD as the primary outcome whereas PCE was developed to predict cardiovascular disease which includes CAD and stroke. Nevertheless, previous studies have shown that the PCE is effective in predicting CAD[18,19]. Third, the CAD PRS used in this study included low frequency and common variants, but did not incorporate rare high-impact genetic variants associated with CAD risk. However, based on whole genome sequencing data from half a million populations, we separately identified somatic mutations and curated significant rare variants such as those linked to FH to develop a comprehensive genomic model. This approach allowed us to comprehensively capture the diverse spectrum of genetic variant frequencies linked to CAD. Fourth, baseline CAD risk is different in the UK Biobank and TOPMed because UK Biobank consists of healthier individuals compared to TOPMed. The incidence of CAD was 7.2% and 12.7% respectively for UK Biobank and TOPMed (Supplementary Table 3), and this was reflected in the risk gradient captured by IGM the 10-year CAD risk ranged from 1.1% to 15.5% among UK Biobank participants, and 3.8% to 33.0% in TOPMed study participants (Fig. 3). Fifth, the evolving CHIP status over time might affect the model's predictive accuracy during prolonged follow-up. Nevertheless, the low incidence of CHIP in younger and middle-aged individuals[20] indicates that this limitation is unlikely to influence the performance of the model in this population for whom IGM offers the greatest potential benefit. Lastly, while our integrated genomic model was developed using data from UK Biobank and externally validated in TOPMed, its clinical implementation and actionability have not yet been explored. Future prospective validation studies and trialsshould focus on understanding how the integrated genomic model can be effectively incorporated into clinical practice and assessing its practical benefits.

As clinical genomics increasingly informs precision medicine, the emphasis is shifting from population-level performance metrics to individual-level interpretability. Although its incremental predictive gain over PRS was modest at the population level, the IGM identified 13% of high-risk individuals who would otherwise remain unrecognized in a PRS-only model, illustrating the potential utility of individualized genomic profiling in the era of precision medicine. Importantly, this work represents an early, proof-of-concept framework that integrates multiple genomic risk drivers into a single risk estimate. Future studies are warranted to establish analytic validity and clinical utility. Beyond CAD, this approach may be extended to other complex disorders where multiple genetic pathways converge on shared phenotypic outcomes.

Methods

Ethics

The UK Biobank was approved by the UK Biobank Research Ethics Committee (ref. 11/NW/0382). All UK Biobank participants provided written informed consent. Use of UK Biobank (application number 89885) data was approved by the Beijing Institute of Genomics review board. All contributing studies in TOPMed consortium obtained ethics approval from their respective local institutional review boards or ethics committees, with written informed consent from participants. Use of the TOPMed data was approved by the Massachusetts General Hospital institutional review board and TOPMed Atherosclerosis Working Group.

Dataset and quality control

Access to the UK Biobank data was approved with application ID 89885. Samples with discordance between self-reported sex (Field 31) and genetically inferred sex (Field 22001) were removed. Additionally, samples with individual-level genotype missing rates (Field 22005) greater than 5%, outliers for heterozygosity or missing rate (Field 22027), or sex chromosome aneuploidy (Field 22019) were excluded. To remove close relatives in the samples, we excluded one of the samples whose pairwise kinship value is greater than or equal to 0.0884 (threshold of the second-degree close relatives[21]) but also tried to keep as many samples as possible. Finally, 391,536 individuals were included in the final analysis (Supplementary Fig. 1A).

A total of 80,588 participants from the NHLBI's TOPMed program with available whole genome sequencing data were considered. These

studies primarily consist of observational cohorts that have been described in detail previously[6]. Among 80,588 participants, we excluded 49 samples with conflicting sex information, 1 sample without principal components, and 17,237 samples with excess kinship, defined as a second-degree relationship or closer, indicated by a KING coefficient greater than 0.0884. Finally, 34,177 participants remained for analysis after excluding an additional 29,124 participants without CAD phenotype (Supplementary Fig. 1B). Cohorts contributed to this population are Amish, Atherosclerosis Risk in Communities Study[ARIC], Cardiovascular Health Study[CHS], Genetic epidemiology of COPD[COPDGene], Diabetes Heart Study[DHS], Framingham Heart Study[FHS], Genetic Study of Atherosclerosis Risk[GeneSTAR], Genetic Epidemiology Network of Arteriopathy[GENOA], Jackson Heart Study[JHS], Multi-Ethnic Study of Atherosclerosis[MESA], and Women's Health Initiative[WHI].

Curation of FH, CHIP, and LTL

To determine the carrier status of familial hypercholesterolemia (FH) for samples with available whole-exome sequencing data, a combination of variant selection criteria was applied: 1) Variants from previous publications which were manually curated by clinical geneticists[4,22–24]; 2) Variants in *LDLR, APOB* and *PCSK9* from the ClinVar database (downloaded February 27th, 2023, GRCh38) annotated as pathogenic, likely pathogenic, or pathogenic/likely pathogenic without conflicts. For *APOB* and *PCSK9* gene, only variants associated with hypercholesterolemia but not hypobetalipoproteinemia were included; 3) Variants in *LDLR* annotated as high-confidence loss-of-function by the VEP Loss-Of-Function Transcript Effect Estimator (LOFTEE) plugin were also included[25,26]. Only variants with a net positive association with LDL cholesterol level accessed by an iterative conditional regression analysis were included in the final variant list for subsequent analysis (Supplementary Data 1)[27].

The carrier status of clonal hematopoiesis of indeterminate potential (CHIP) was detected following a similar procedure described in Yu Zhi et al. and others[28–30]. Specifically, carrier status was determined by carrying CHIP variants from one of the genes *TET2, ASXL1, JAK2, PPM1D, TP53, SRSF2, and SF3B1*. Leukocyte telomere length (LTL) was log-transformed to obtain a normal distribution and then Z-standardized using the distribution of all individuals with a telomere length measurement (Field 22192). Details of processing were described in V. Codd et al.[31]. Unless otherwise specified, genetic drivers have been curated comparably for UK Biobank and TOPMed. More details on the curation of CHIP and LTL for TOPMed are described elsewhere[6].

Development of CAD polygenic risk score and genetically proxied risk scores for metabolome and proteome

In the UK Biobank, CAD PRS was calculated using the imputed genotype data and variant weights from the multi-ancestry and multi-trait polygenic risk score for CAD described in A. P. Patel et al.[13], implemented with PLINK2[32]. Instead of relying on direct measurement of metabolite and protein levels of the limited number of samples, we utilized previously identified genetic proxies for these traits[10], allowing us to leverage the vast amounts of genomic data available for large cohorts. Variant weights for the genetic proxies were downloaded from OMICSPRED resource which derived from the INTERVAL study cohort (https://www.omicspred.org/Scores/Somalogic/INTERVAL)[10], including 867 metabolites (726 Metabolon, and 141 Nightingale), and 2692 proteins (2384 Somalogic, and 308 Olink), respectively. The 867 predicted metabolite levels and 2692 protein levels were calculated using the imputed genotype data by PLINK2[32]. Then we randomly sampled 200,000 individuals from UK Biobank as the training set, and a lasso penalty was applied to the Cox proportional hazards regression model with age, sex, and the top 10 principal components (PCs) as covariates and incident CAD as an outcome to develop the proxy risk scores for metabolome (MetPRS) and proteome (ProPRS). Five-fold cross-validation was employed to select the optimal penalization strength for hyperparameters, similar to the process described elsewhere[12]. The number of genetically predicted metabolome and proteome scores retained in the MetPRS and ProPRS lasso models were 142 and 124, respectively (Supplementary Data 2-3). Variant weights were further adjusted with the weights of genetic proxies determined by the cross-validation procedures, and finally applied to calculate the MetPRS and ProPRS for all the UK Biobank samples and the TOPMed cohort samples accordingly for subsequent analysis.

CAD definition

In the UK Biobank, CAD was defined based on self-report at enrollment, hospitalization records, or death registry records as previously described (Supplementary Table 1)[4]. In the TOPMed studies, CAD was defined as ischemic heart disease events, including myocardial infarction and coronary revascularization[6]. Incident CAD cases were defined as those diagnosed after recruitment. The survival year was defined as the years between recruitment and diagnosis for incident CAD cases, or between the time of the recruitment and the last censoring for controls.

Covariates and adjustment

Untreated blood pressure was estimated by adjusting the raw value for anti-hypertensive medication intake by adding 15 mmHg to the systolic blood pressure and 10 mmHg to the diastolic blood pressure[33,34]. Untreated lipid levels were estimated by adjusting the raw lab-tested value according to lipid-lowering medication intake[13] and detailed in Supplementary Table 2. To eliminate potential confounding effects of covariates, PRS, MetPRS, ProPRS, and LTL were regressed on recruitment age, sex, and the first 10 principal components of genetic ancestry (Supplementary Fig. 2). The scaled residuals with mean zero and standard deviation of one were then used in the subsequent analyses.

Developing the integrated genomic model

We evaluated pairwise correlations between six genetic variables from the UK Biobank and TOPMed studies to assess potential multicollinearity before including these variables in a regression model. The correlation matrix was computed using Pearson correlation coefficients. As germline genetic risk drivers (FH, PRS, MetPRS, and ProPRS) remain constant from birth, while CHIP accumulates and LTL shortens with age (somatic risk drivers), we systematically characterized and investigated the joint effects of germline and somatic risk drivers on CAD. Germline risks were combined into a single value termed GermRisk, and somatic risks were combined into a single value termed SomaRisk, the combination weights were estimated by a joint regression model. Specifically, for prevalent CAD, the germline risk drivers (FH, PRS, MetPRS, and ProPRS) were fitted into a logistic regression model, and the coefficients were used as weights to combine the values of the drivers into GermRisk linearly. For incident CAD, the germline risk drivers were fitted to a Cox proportional hazards model, and GermRisk was calculated as a weighted summation of germline variables and their corresponding Cox coefficients, and similar procedures were performed to obtain SomaRisk.

Next, GermRisk and SomaRisk were regressed on recruitment age, sex, and the top 10 PCs, and the residuals were scaled to have zero mean and unit standard deviation, respectively. The standardized residuals of GermRisk and SomaRisk were then fitted to a Cox proportional hazards model. A final predictor, termed IGM (integrated genomic model), was a linear summation of GermRisk and SomaRisk residual according to the Cox coefficient estimation. The weights were fixed and applied to TOPMed data for independent validation. The overall framework for developing and validating the integrated genomic model is shown in Fig. 1.

Estimating effect sizes

To investigate the genomic factors driving the risk of prevalent and incident CAD, a logistic regression model and Cox proportional hazards regression model were employed respectively to estimate the effect sizes for individual genetic drivers. Additionally, PRS, MetPRS, ProPRS, LTL, and the ensembled ones (IGMRisk, GermRisk and SomaRisk) were binarized, with individuals in the top 20% treated as carriers (hPRS, hMetPRS, hProPRS, hIGMRisk, hGermRisk and hSomaRisk) or bottom 20% treated as a carrier (sLTL) for telomere length.

Interplay between genomic drivers and clinical risk score

To investigate the interplay between genomic risk and PCE, participants in the UK Biobank were divided into four groups based on guideline-defined categories of the PCE—low (estimated risk less than 5%), borderline (risk between 5% to 7.5%), intermediate (risk between 7.5% and 20%), and high (risk greater than 20%)[16]. For the IGM, we divide samples into three risk groups as standard in genomic analyses —high genomic risk (top quintile of the distribution), intermediate genomic risk (middle three quintiles), and low risk (bottom quintile).

Estimating 10-year risk of CAD

The 10-year risk of CAD was estimated by Cox proportional hazards regression with GermRisk and SomaRisk as predictors and the age, sex, and first 10 principal components of genetic ancestry as covariates. To investigate the stratification capacity of different models, the C-statistic from 1) a base model (sex, age, and first 10 PCs), 2) a log-transformed value of PCE, 3) a base model and IGM, and 4) a base model, IGM and log(PCE), was estimated by a Cox proportional hazards model, respectively.

To evaluate the improvement of prediction by adding the IGM to PCE, the net reclassification improvement was calculated based on a 10-year risk threshold of 7.5% for categorical reclassification and threshold 0 for continuous reclassification as demonstrated elsewhere[18,19]. Confidence intervals were estimated by 100 times bootstrap. All statistical analyses were done using R v4.2.2 (R Foundation, Vienna, Austria), including the following packages: survival (v3.5-7), survminer (v0.4.9), tableone (v0.13.2), pROC(v1.18.5), nricens(v1.6), rms(v6.7-1) and glmnet (v4.1-8).

Development of metabolome and proteome risk scores based on measured levels

We developed the risk scores using measured metabolomic and proteomic data from 270k and 50 K UK Biobank participants, respectively. We used 251 NMR metabolic biomarkers from Nightingale Health to create a metabolome risk score for CAD. After excluding participants with missing data, the remaining cohort ($N = 212{,}154$) was randomly split into equal training and test sets. Using the training set, we performed five-fold cross-validated least absolute shrinkage and selection operator (LASSO) regression to select metabolite features associated with CAD, adjusting for age, sex, and the top 10 PCs of genetic ancestry, similar to our development of the genetically proxied proteomic and metabolomic risk scores (ProPRS and MetPRS). The metabolome risk score was then evaluated in the independent test set for its association with CAD risk. We used similar procedures to develop a proteome risk score with 2923 protein measures via Olink ($N = 45{,}574$).

Reporting summary

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

Data availability

De-identified participant data, including genotyping array, sequencing, and phenotype data were obtained from the UK Biobank [https://www.ukbiobank.ac.uk/]. All data are made available from the UK Biobank (https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access) to researchers from universities and other institutions with genuine research inquiries following institutional review board and UK Biobank approval. The weights of MetPRS and ProPRS are available in the Polygenic Score Catalog (IDs: PGS005093 and PGS005094). This paper used the TOPMed whole genome sequencing (WGS) data and cardiovascular disease phenotype data. Genotype and phenotype data are both available in database of Genotypes and Phenotypes (dbGaP [https://dbgap.ncbi.nlm.nih.gov/]). The TOPMed WGS data were from the following eleven study cohorts: Amish, Atherosclerosis Risk in Communities Study (ARIC), Cardiovascular Health Study (CHS), Genetic epidemiology of COPD (COPDGene), Diabetes Heart Study (DHS), Framingham Heart Study (FHS), Genetic Study of Atherosclerosis Risk (GeneSTAR), Genetic Epidemiology Network of Arteriopathy (GENOA), Jackson Heart Study (JHS), Multi-Ethnic Study of Atherosclerosis (MESA), and Women's Health Initiative (WHI). Source data for figures are provided with this paper. All the data used in this study was previously published and we have not produced new experimental data. Source data are provided with this paper.

References

1. Martin, S. S. et al. Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. *Circulation* **149**, 2024 (2024).
2. Fahed, A. C. & Natarajan, P. Clinical applications of polygenic risk score for coronary artery disease through the life course. *Atherosclerosis* **386**, 117356 (2023).
3. Mars, N. et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat. Med.* **26**, 549–557 (2020).
4. Fahed, A. C. et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat. Commun.* **11**, 3635 (2020).
5. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
6. Nakao, T. et al. Mendelian randomization supports bidirectional causality between telomere length and clonal hematopoiesis of indeterminate potential. *Sci. Adv.* **8**, eabl6579 (2022).
7. Haycock, P. C. et al. Leucocyte telomere length and risk of cardiovascular disease: systematic review and meta-analysis. *BMJ* **349**, g4227 (2014).
8. Zhao, K. et al. Somatic and Germline Variants and Coronary Heart Disease in a Chinese Population. *JAMA Cardiol* **9**, 233–242 (2024).
9. Muntner, P. et al. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. *JAMA* **311**, 1406–1415 (2014).
10. Xu, Y. et al. An atlas of genetic scores to predict multi-omic traits. *Nature* **616**, 123–131 (2023).
11. Goff, D. C. et al. ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* **129**, 2014 (2013).
12. Helgason, H. et al. Evaluation of Large-Scale Proteomics for Prediction of Cardiovascular Events. *JAMA* **330**, 725–735 (2023).
13. Patel, A. P. et al. A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. *Nat. Med.* **29**, 1793–1803 (2023).
14. Marston, N. A. et al. Predictive Utility of a Coronary Artery Disease Polygenic Risk Score in Primary Prevention. *JAMA Cardiol* **8**, 130–137 (2023).
15. Urbut, S. M. et al. Dynamic Importance of Genomic and Clinical Risk for Coronary Artery Disease Over the Life Course. *MedRxiv Prepr. Serv. Health Sci.* 2023.11.03.23298055 (2023) https://doi.org/10.1101/2023.11.03.23298055.

16. Arnett, D. K. et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: Executive Summary. *J. Am. Coll. Cardiol.* **74**, 1376–1414 (2019).
17. Carrasco-Zanini, J. et al. Proteomic signatures improve risk prediction for common and rare diseases. *Nat. Med.* (2024) https://doi.org/10.1038/s41591-024-03142-z.
18. Mosley, J. D. et al. Predictive Accuracy of a Polygenic Risk Score Compared With a Clinical Risk Score for Incident Coronary Heart Disease. *JAMA* **323**, 627–635 (2020).
19. Elliott, J. et al. Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. *JAMA* **323**, 636–645 (2020).
20. Uddin, M. M. et al. Long-term longitudinal analysis of 4187 participants reveals insights into determinants of clonal hematopoiesis. *Nat. Commun.* **15**, 7858 (2024).
21. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinforma. Oxf. Engl.* **26**, 2867–2873 (2010).
22. Khera, A. V. et al. Diagnostic Yield and Clinical Utility of Sequencing Familial Hypercholesterolemia Genes in Patients With Severe Hypercholesterolemia. *J. Am. Coll. Cardiol.* **67**, 2578–2589 (2016).
23. Khera, A. V. et al. Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction. *Circulation* **139**, 1593–1602 (2019).
24. Reeskamp, L. F. et al. Polygenic Background Modifies Risk of Coronary Artery Disease Among Individuals With Heterozygous Familial Hypercholesterolemia. *JACC Adv.* **2**, 100662 (2023).
25. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
26. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
27. Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P.-R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat. Genet.* **53**, 1260–1269 (2021).
28. Yu, Z. et al. Genetic modification of inflammation- and clonal hematopoiesis-associated cardiovascular risk. *J. Clin. Invest.* **133**, e168597 (2023).
29. Bick, A. G. et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763–768 (2020).
30. Vlasschaert, C. et al. A practical approach to curate clonal hematopoiesis of indeterminate potential in human genetic data sets. *Blood* **141**, 2214–2223 (2023).
31. Codd, V. et al. Measurement and initial characterization of leukocyte telomere length in 474,074 participants in UK Biobank. *Nat. Aging* **2**, 170–179 (2022).
32. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
33. The Million Veteran Program et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat. Genet.* **50**, 1412–1425 (2018).
34. Tobin, M. D., Sheehan, N. A., Scurrah, K. J. & Burton, P. R. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Stat. Med.* **24**, 2911–2935 (2005).

Acknowledgements

M.W. is supported by the National Natural Science Foundation of China (grant number 82470352), Noncommunicable Chronic Diseases-National Science and Technology Major Project-2023ZD0503201, Pioneering Action Grants of the Chinese Academy of Sciences, Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No.XDA0460400), the Scientific data Plan of Chinese Academy of Sciences (Grant No.CAS-SJ2025-01), and Zhejiang Provincial Clinical Research Center for Pediatric Diseases-ZJEK2302Z. P.T.E. is supported by grants from the National Institutes of Health (R01HL092577, 1R01HL157635, 5R01HL139731), from the American Heart Association (18SFRN34110082, 961045) and from the European Union (MAESTRIA 965286). P.N. is funded by grants R01HL1427, R01HL148565 R01HL148050, and U01HG011719 from the National Institutes of Health. A.C.F. receives funding from the National Heart Lung and Blood Institute under award numbers K08 HL161448 and R01 HL164629. P.S.de.V. is funded by R01HL146860 from the National Heart, Lung and Blood Institute (NHLBI). M.W. is supported by the Pioneering Action Grants of the Chinese Academy of Sciences. Molecular data for the Trans Omics in Precision Medicine (TOPMed) program was supported by the NHLBI. Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC and general program coordination was provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute, the National Institutes of Health, or the U.S. Department of Health and Human Services. We wish to acknowledge the contributions of the consortium working on the development of the NHLBI BioData Catalyst ecosystem. Support for the Genetic Epidemiology Network of Arteriopathy (GENOA) was provided by the National Heart, Lung and Blood Institute (U01 HL054457, U01 HL054464, U01 HL054481, R01 HL119443, and R01 HL087660) of the National Institutes of Health. DNA extraction for "NHLBI TOPMed: Genetic Epidemiology Network of Arteriopathy" (phs001345) was performed at the Mayo Clinic Genotyping Core, and WGS was performed at the DNA Sequencing and Gene Analysis Center at the University of Washington (3R01HL055673-18S1) and the Broad Institute (HHSN268201500014C). We would like to thank the GENOA participants. The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the NHLBI and the National Institute on Minority Health and Health Disparities (NIMHD). Genome sequencing for "NHLBI TOPMed: The Jackson Heart Study" (phs000964.v1.p1) was performed at the Northwest Genomics Center (HHSN268201100037C). The authors also wish to thank the staffs and participants of the JHS. The MESA projects are conducted and supported by NHLBI in collaboration with MESA investigators. Support for the Multi-Ethnic Study of Atherosclerosis (MESA) projects are conducted and supported by the NHLBI in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420, UL1TR001881, DK063491, R01HL105756, and R01HL146860. Genome sequencing for "NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Multi-Ethnic Study of Atherosclerosis Study (MESA)" (phs001416) was performed at Broad Institute of MIT and Harvard Genomics Platform (3U54HG003067-13S1). The authors thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutes can be found at http://www.mesa-nhlbi.org. The Women's Health Initiative (WHI) program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004,

75N92021D00005. Genome sequencing for "NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Women's Health Initiative Study (WHI)" (phs001237) was performed at Broad Institute of MIT and Harvard Genomics Platform (HHSN268201500014C). Support for the Diabetes Heart Study (DHS) by R01 HL92301, R01 HL67348, R01 NS058700, R01 AR48797, R01 DK071891, R01 AG058921, the General Clinical Research Center of the Wake Forest University School of Medicine (M01 RR07122, F32 HL085989), the American Diabetes Association, and a pilot grant from the Claude Pepper Older Americans Independence Center of Wake Forest University Health Sciences (P60 AG10484). Genome sequencing for "NHLBI TOPMed: The Diabetes Heart Study" (phs001412) was performed at the Broad Institute of MIT and Harvard Genomic Platform (HHSN268201500014C). The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services, under Contract nos. (75N92022D00001, 75N92022D00002, 75N92022D00003, 75N92022D00004, 75N92022D00005). The authors thank the staff and participants of the ARIC study for their important contributions. WGS for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Atherosclerosis Risk in Communities (ARIC)" (phs001211) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201500015C and 3U54HG003273-12S2) and the Broad Institute for MIT and Harvard (3R01HL092577- 06S1). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL- 120393- 02S1). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The Genome Sequencing Program (GSP) was funded by the National Human Genome Research Institute (NHGRI), the National Heart, Lung, and Blood Institute (NHLBI), and the National Eye Institute (NEI). The GSP Coordinating Center (U24 HG008956) contributed to cross program scientific initiatives and provided logistical and general study coordination. The Centers for Common Disease Genomics (CCDG) program was supported by NHGRI and NHLBI, and whole genome sequencing was performed at the Baylor College of Medicine Human Genome Sequencing Center (UM1 HG008898). The COPDGene study (NCT00608764) is supported by grants from the NHLBI (U01HL089897 and U01HL089856), by NIH contract 75N92023D00011, and by the COPD Foundation through contributions made to an Industry Advisory Committee that has included AstraZeneca, Bayer Pharmaceuticals, Boehringer-Ingelheim, Genentech, GlaxoSmithKline, Novartis, Pfizer and Sunovion. A full listing of COPDGene investigators can be found at: http://www.copdgene.org/directory. Genome sequencing for "NHLBI TOPMed: Genetic Epidemiology of COPD Study" (phs000951) was performed at Northwest Genomics Center and Broad Genomics (3R01HL089856-08S1, HHSN268201500014C, HHSN268201500014C). This Cardiovascular Health Study research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, 75N92021D00006; and NHLBI grants U01HL080295, R01HL087652, R01HL105756, R01HL103612, R01HL120393, and U01HL130114 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through R01AG023629 from the National Institute on Aging (NIA). Genome sequencing for "NHLBI TOPMed: Cardiovascular Health Study" (phs001368.v2.p1) was performed at the Baylor College of Medicine Human Genome Sequencing Center (3U54HG003273-12S2, HHSN268201500015C, HHSN2682016000033I). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. The TOPMed component of the Amish Research Program was supported by NIH grants R01 HL121007, U01 HL072515, and R01 AG18728. Genome sequencing for NHLBI TOPMed: Amish (phs000956) was performed at the Broad Institute of MIT and Harvard (3R01HL121007-01S1). The Framingham Heart Study (FHS) acknowledges the support of contracts NO1-HC-25195, HHSN2682015000001I and 75N92019D00031 from the National Heart, Lung and Blood Institute and grant supplement R01 HL092577-06S1 for this research. Genome sequencing for "NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study (FHS)" (phs000974) was performed at Broad Institute of MIT and Harvard Genomics Platform (3U54HG003067-12S2). We also acknowledge the dedication of the FHS study participants without whom this research would not be possible. R.S.V. is supported in part by the Evans Medical Foundation and the Jay and Louis Coffman Endowment from the Department of Medicine, Boston University School of Medicine. GeneSTAR was supported by the National Institutes of Health/National Heart, Lung, and Blood Institute (U01 HL72518, HL087698, HL112064, HL49762, HL59684, HL58625, HL071025), by the National Institutes of Health/ National Institute of Nursing Research (NR0224103, NR008153), and by a grant from the National Institutes of Health/National Center for Research Resources (M01-RR000052) to the Johns Hopkins General Clinical Research Center. Genome sequencing for NHLBI TOPMed: GeneSTAR (Genetic Study of Atherosclerosis Risk) (phs001218) was performed at the Broad Institute of MIT and Harvard (HHSN268201500014C), at PsomaGen (formerly Macrogen, HHSN268201500014C), and at Illumina (HL112064). We gratefully acknowledge the studies and participants who provided biological samples and data for UK Biobank.

Author contributions

P.N., M.W., A.C.F., M.S.K., and X.Y designed the study. X.Y., and M.S.K. performed analysis. X.Z., M.M.U., T.N., S.M.J.C., S.K., T.X., L.F.R., R.Z., Z.Ll, and Y.A. contributed to data preparation, preprocess, and analytic planning. P.S.V., R.S.V., E.B., A.C.M., B.M.P., R.P.T., S.R.H., M.H.C., J.H.Y., N.D.P., D.W.B., J.M.M., D.L., N.L.H., G.T.O., L.C.B., B.G.K., L.R.Y., L.M.R., B.H., J.I.R., S.S.R., K.D.T., W.S.P., C.K., A.P.R., B.D.M., S.L.R.K., J.A.S., P.A.P., D.K.A., A.V.S., S.B.G., L.F.B. and NHLBI Trans-Omics for Precision Medicine Consortium provided and managed data from TOPMed. M.S.K., and X.Y. drafted the paper. D.K.Y., H.H.W., P.T.E., P.N., M.W., and A.C.F. revised and contributed to final draft. All authors reviewed the paper and provided critical revisions.

Competing interests

L.F.R. is cofounder of Lipid Tools and reports speaker fees from Ultragenyx, Novartis, and Daiichi Sankyo. P.T.E. receives sponsored research support from Bayer AG, Bristol Myers Squibb, Pfizer and Novo Nordisk; he has also served on advisory boards or consulted for Bayer AG, all unrelated to the present work. P.N. reports research grants from Allelica, Amgen, Apple, Boston Scientific, Genentech/Roche, and Novartis, personal fees from Allelica, Apple, AstraZeneca, Blackstone Life Sciences, Creative Education Concepts, CRISPR Therapeutics, Eli Lilly & Co, Esperion Therapeutics, Foresite Labs, Genentech/Roche, GV, HeartFlow, Magnet Biomedicine, Merck, Novartis, TenSixteen Bio, and Tourmaline Bio, equity in Bolt, Candela, Mercury, MyOme, Parameter Health, Preciseli, and TenSixteen Bio, and spousal employment at Vertex Pharmaceuticals, all unrelated to the present work. A.C.F. reports being cofounder of Goodpath and Avigena, serving as scientific advisor to MyOme, Arboretum Health, Novartis, and Aditum Bio and receiving sponsored research awards from Foresite, Sarepta Therapeutics, and Allelica, all of which are unrelated to the present work. L.M.R. and S.S.R. are consultants for the NHLBI TOPMed Administrative Coordinating Center (through Westat). M.H.C. has received grant support from Bayer and consulting fees from Apogee and BMS. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-70379-2.

Correspondence and requests for materials should be addressed to Pradeep Natarajan, Minxian Wang or Akl C. Fahed.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Xiong Yang[1,2,54,55], Min Seo Kim[3,4,55], Xinyu Zhu[1,2,5], Md Mesbah Uddin[3,4], Tetsushi Nakao[3,4,6,7], So Mi Jemma Cho[3,4], Satoshi Koyama[3,4], Tingfeng Xu[1,2,5], Laurens F. Reeskamp[3,8,9], Rufan Zhang[1,2], Zhaoqi Liu[1,2,5], Yunga A[1,2,5], Paul S. de Vries[10], Ramachandran S. Vasan[11,12], Eric Boerwinkle[10], Alanna C. Morrison[10], Bruce M. Psaty[13,14,15], Russell P. Tracy[16,17], Susan R. Heckbert[18], Michael H. Cho[19,20], Jeong H. Yun[19,20], Nicholette D. Palmer[21], Donald W. Bowden[21], Joanne M. Murabito[22,23], Daniel Levy[24,25], Nancy L. Heard-Costa[25,26], George T. O'Connor[25,27], Lewis C. Becker[28,29], Brian G. Kral[28,29], Lisa R. Yanek[29], Laura M. Raffield[30], Bertha Hidalgo[31], Jerome I. Rotter[32], Stephen S. Rich[33], Kent D. Taylor[32], Wendy S. Post[34], Charles Kooperberg[35], Alexander P. Reiner[18], Braxton D. Mitchell[36], Sharon L. R. Kardia[37], Jennifer A. Smith[37], Patricia A. Peyser[37], Lawrence F. Bielak[37], Dong Keon Yon[38,39,40,41,42], Hong-Hee Won[43,44], Donna K. Arnett[45], Albert V. Smith[46], Stacey B. Gabriel[47], Patrick T. Ellinor[3,4,48,49], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Pradeep Natarajan[3,4,49,50,56] ✉, Minxian Wang[1,2,5,51,52,53,56] ✉ & Akl C. Fahed[3,4,48,49,56] ✉

[1]National Genomics Data Center, China National Center for Bioinformation, Beijing, China. [2]Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China. [3]Program in Medical and Population Genetics and the Cardiovascular Disease Initiative, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [5]College of Future Technology, Sino-Danish College, University of Chinese Academy of Sciences, Beijing, China. [6]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [7]Division of Cardiovascular Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [8]Department of Vascular Medicine, Amsterdam Cardiovascular Sciences, Amsterdam University Medical Centers, Location AMC, University of Amsterdam, Amsterdam, The Netherlands. [9]Department of Internal Medicine, OLVG, Amsterdam, The Netherlands. [10]Human Genetics Center, Department of Epidemiologys, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA. [11]School of Public Health, University of Texas, San Antonio, TX, USA. [12]Department of Medicine, Boston University School of Medicine, Boston, MA, USA. [13]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA. [14]Cardiovascular Health Research Unit, Department of Epidemiology, University of Washington, Seattle, WA, USA. [15]Cardiovascular Health Research Unit, Department of Health Systems and Population Health, University of Washington, Seattle, WA, USA. [16]Department of Pathology and Laboratory Medicine, The Robert Larner M.D. College of Medicine, University of Vermont, Burlington, VT, USA. [17]Department of Biochemistry, The Robert Larner M.D. College of Medicine, University of Vermont, Burlington, VT, USA. [18]Department of Epidemiology, University of Washington, Seattle, WA, USA. [19]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [20]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [21]Department of Biochemistry, Wake Forest School of Medicine, 1 Medical Center Blvd, Winston-Salem, NC, USA. [22]Framingham Heart Study, National Heart, Lung, and Blood Institute and Boston University Chobanian & Avedisian School of Medicine, Framingham, MA, USA. [23]Department of Medicine, Section of General Internal Medicine, Boston University Chobanian & Avedisian School of Medicine and Boston Medical Center, Boston, MA, USA. [24]Population Sciences Branch, Division of Intramural Research National Heart, Lung, and Blood Institute, Framingham, MA, USA. [25]Framingham Heart Study, Framingham, MA, USA. [26]Department of Neurology, Chobanian & Avedisian School of Medicine, Boston University, 72 E Concord St, Boston, MA, USA. [27]Boston University School of Medicine, Pulmonary Center, Boston, MA, USA. [28]Department of Medicine, Division of Cardiology, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [29]Department of Medicine, Division of General Internal Medicine, GeneSTAR Research Program, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [30]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA. [31]Center for Clinical and Translational Science, The University of Alabama at Birmingham, Birmingham, AL, USA. [32]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA. [33]Department of Genome Sciences, University of Virginia, Charlottesville, VA, USA. [34]Division of Cardiology, Department of Medicine, Johns Hopkins University, Baltimore, MD, USA. [35]Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA, USA. [36]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA. [37]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA. [38]Center for Digital Health, Medical Science Research Institute, Kyung Hee University College of Medicine, Seoul, South Korea. [39]Department of Precision Medicine, Kyung Hee University College of Medicine, Seoul, South Korea. [40]Department of Regulatory Science, Kyung Hee

University, Seoul, South Korea. [41]Department of Medicine, Kyung Hee University College of Medicine, Seoul, South Korea. [42]Department of Pediatrics, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, South Korea. [43]Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology (SAIHST), Sungkyunkwan University, Seoul, South Korea. [44]Samsung Genome Institute, Samsung Medical Center, Seoul, South Korea. [45]Office of the Provost, University of South Carolina, Center for Statistical Genetics, Columbia, SC, USA. [46]Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA. [47]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [48]Cardiology Division, Massachusetts General Hospital, Boston, MA, USA. [49]Harvard Medical School, Boston, MA, USA. [50]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [51]Department of Cardiovascular Surgery, Zhongnan Hospital of Wuhan University, Wuhan, China. [52]Hubei Provincial Engineering Research Center of Minimally Invasive Cardiovascular Surgery, Wuhan, China. [53]Wuhan Clinical Research Center for Minimally Invasive Treatment of Structural Heart Disease, Wuhan, China. [54]Present address: Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing, China. [55]These authors contributed equally: Xiong Yang, Min Seo Kim. [56]These authors jointly supervised this work: Pradeep Natarajan, Minxian Wang, Akl C. Fahed. A list of members and their affiliations appears in the Supplementary Information. ✉e-mail: pnatarajan@mgh.harvard.edu; wangmx@big.ac.cn; afahed@mgh.harvard.edu