An atlas of genetic scores to predict multi-omic traits

Citation for published version:
Xu, Y, Ritchie, SC, Liang, Y, Timmers, PRHJ, Pietzner, M, Lannelongue, L, Lambert, SA, Tahir, UA, May-Wilson, S, Foguet, C, Johansson, Å, Surendran, P, Nath, AP, Persyn, E, Peters, JE, Oliver-Williams, C, Deng, S, Prins, B, Luan, J, Bomba, L, Soranzo, N, Di Angelantonio, E, Pirastu, N, Tai, ES, van Dam, RM, Parkinson, H, Davenport, EE, Paul, DS, Yau, C, Gerszten, RE, Mälarstig, A, Danesh, J, Sim, X, Langenberg, C, Wilson, JF, Butterworth, AS & Inouye, M 2023, 'An atlas of genetic scores to predict multi-omic traits', *Nature*, vol. 616, no. 7955, pp. 123-131. https://doi.org/10.1038/s41586-023-05844-9

Digital Object Identifier (DOI):
10.1038/s41586-023-05844-9

Link:
Link to publication record in Edinburgh Research Explorer

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

Published In:
Nature

An atlas of genetic scores to predict multi-omic traits

https://doi.org/10.1038/s41586-023-05844-9

Received: 29 April 2022

Accepted: 15 February 2023

Published online: 29 March 2023

Check for updates

Yu Xu[1,2,3] ✉, Scott C. Ritchie[1,2,3,4], Yujian Liang[5], Paul R. H. J. Timmers[6], Maik Pietzner[7,8,9], Loïc Lannelongue[1,2,3,10], Samuel A. Lambert[1,2,3,4,10,11], Usman A. Tahir[12], Sebastian May-Wilson[6], Carles Foguet[1,2,3,10], Åsa Johansson[13], Praveen Surendran[2], Artika P. Nath[1,14], Elodie Persyn[1,2,3], James E. Peters[15], Clare Oliver-Williams[2], Shuliang Deng[12], Bram Prins[2], Jian'an Luan[7], Lorenzo Bomba[16,17], Nicole Soranzo[4,16,18,19,20], Emanuele Di Angelantonio[2,3,4,10,19,21], Nicola Pirastu[6,20], E. Shyong Tai[5,22], Rob M. van Dam[5,23], Helen Parkinson[11], Emma E. Davenport[16], Dirk S. Paul[2,4], Christopher Yau[24,25,26], Robert E. Gerszten[12,27], Anders Mälarstig[28,29], John Danesh[2,3,4,10,16,19], Xueling Sim[5], Claudia Langenberg[7,8,9], James F. Wilson[6,30], Adam S. Butterworth[2,3,4,10,19] & Michael Inouye[1,2,3,4,10,14,31] ✉

The use of omic modalities to dissect the molecular underpinnings of common diseases and traits is becoming increasingly common. But multi-omic traits can be genetically predicted, which enables highly cost-effective and powerful analyses for studies that do not have multi-omics[1]. Here we examine a large cohort (the INTERVAL study[2]; $n = 50{,}000$ participants) with extensive multi-omic data for plasma proteomics (SomaScan, $n = 3{,}175$; Olink, $n = 4{,}822$), plasma metabolomics (Metabolon HD4, $n = 8{,}153$), serum metabolomics (Nightingale, $n = 37{,}359$) and whole-blood Illumina RNA sequencing ($n = 4{,}136$), and use machine learning to train genetic scores for 17,227 molecular traits, including 10,521 that reach Bonferroni-adjusted significance. We evaluate the performance of genetic scores through external validation across cohorts of individuals of European, Asian and African American ancestries. In addition, we show the utility of these multi-omic genetic scores by quantifying the genetic control of biological pathways and by generating a synthetic multi-omic dataset of the UK Biobank[3] to identify disease associations using a phenome-wide scan. We highlight a series of biological insights with regard to genetic mechanisms in metabolism and canonical pathway associations with disease; for example, JAK–STAT signalling and coronary atherosclerosis. Finally, we develop a portal (https://www.omicspred.org/) to facilitate public access to all genetic scores and validation results, as well as to serve as a platform for future extensions and enhancements of multi-omic genetic scores.

Multi-omic analysis has become a powerful approach to predict disease and analyse its underlying biology. However, collecting transcriptomic, proteomic, metabolomic and other modalities is an extremely expensive and time-consuming process. Because of these barriers, in large-scale population cohorts multi-omic data are typically generated for only a subset of participants (or not at all), which reduces statistical power and creates inequities for studies without ample resources, particularly in underrepresented demographics.

Genetic prediction of complex human traits can have both analytic validity and potential clinical utility[4–7]. Genetic prediction has been extended to omics data; for example, whole-blood[8] and multi-tissue transcriptomics[9], as well as plasma proteomics[10]. Genetically predicted traits can elucidate the molecular aetiology of common diseases, incorporating both directionality (the germline genome is fixed over the life course) and the power of large-scale genotyped biobanks to overcome prediction noise[11,12].

Genetic scores that predict, expand and thereby democratize multi-omics data are of intense interest. Genetic prediction in this area has historically focused on gene expression, drawing on heterogeneous sources for training data with limited sample sizes. With many cohorts now performing multi-omics at scale, there is a unique opportunity to greatly expand and enhance these genetic scores. Given robust external validation, the reliability of multi-omic genetic scores can be quantified and extended to analyses that assess transferability across ancestries, thus facilitating equitable tools for molecular investigation in diverse populations. This approach both facilitates integrative cross-cohort, multi-omic analyses and enables the efficient generation of synthetic multi-omic data for studies with only genetic data.

Here, we use the INTERVAL study[2], a cohort of UK blood donors with extensive multi-omic profiling, to train genetic-prediction models. We externally validate these genetic scores in seven external studies, comprising individuals of European, East Asian, South Asian and African American ancestries. We then demonstrate the use of genetically predicted molecular data, including coverage of biological pathways and the identification of multi-omic predictors of diseases and

A list of affiliations appears at the end of the paper.

traits in the UK Biobank (UKB). Finally, we construct an open resource (https://www.omicspred.org/) that makes all genetic scores, validations and biomarker analyses freely available to the wider community.

Development of genetic scores

We developed genetic scores for blood RNA transcripts, proteins and metabolites (Extended Data Fig. 1). We used the INTERVAL study, which collected serum or plasma from participants and performed assays from five omics platforms: SomaScan v.3 (SomaLogic, USA), an aptamer-based multiplex protein assay; Olink Target (Olink Proteomics, Sweden), an antibody-based proximity extension assay for proteins; Metabolon HD4 (Metabolon, USA), an untargeted mass spectrometry metabolomics platform; Nightingale (Nightingale Health, Finland), a proton nuclear magnetic resonance (NMR) spectroscopy platform; and whole-blood RNA sequencing (RNA-seq) with the Illumina NovaSeq 6000 (Illumina, USA) (Methods). INTERVAL participants were genotyped on the Affymetrix Biobank Axiom array and imputed using a combined 1000 Genomes Phase 3-UK10K reference panel (Methods). After quality control, 10,572,788 genetic variants were available.

To train genetic scores, we used Bayesian ridge regression (BR), as it has been shown to be a powerful and robust approach for genetic prediction[7] that is also computationally scalable to the number of traits analysed here (Methods), thus controlling carbon footprint[13]. We confirmed the generalizability of this method across several platforms, and assessed the effect of different variant-filtering strategies (Methods, Supplementary Figs. 1–4 and Extended Data Fig. 2). Overall, we found that the best-performing approach was BR with genome-wide variant selection using $P < 5 \times 10^{-8}$ (Supplementary Figs. 1–4 and Extended Data Fig. 2).

We developed genetic scores for 17,227 biomolecular traits from the 5 platforms, including 726 metabolites (Metabolon HD4), 141 metabolic traits (Nightingale), 308 proteins measured by Olink, 2,384 proteins measured by SomaScan and 13,668 genes from Illumina RNA-seq (Ensembl gene-level counts) (Methods). Across all platforms, we found wide variation in the predictive value (R^2 between genetically predicted and directly measured biomolecular trait) and the number of variants in the genetic scores in internal validation (Extended Data Fig. 3 and Supplementary Fig. 5).

We found that 10,522 biomolecular traits could be genetically predicted at Bonferroni-adjusted significance (correcting for all genetic scores tested), including those for SomaScan (1,052 traits), Olink (206), Metabolon (379), Nightingale (137) and RNA-seq (8,748). Of these, 5,816 and 409 genetic scores had $R^2 > 0.1$ and $R^2 > 0.5$, respectively (Fig. 1 and Supplementary Tables 1–5).

Genetic scores comprised one to 1,862 genetic variants, with 58% including variants from a single linkage disequilibrium (LD) block, 40% spanning 2–5 LD blocks and 2% spanning 5 or more LD blocks[14]. As expected for gene and protein scores, the contribution from genetic variants in *cis* exceeded that in *trans*. For 89% of these omics traits, *cis* signals (within 1 Mb of the transcription start site) contributed most to the genetic score R^2, with the remaining dominated by *trans* signals. We also compared the gain in R^2 of a genetic score to the top single variant (the one with the greatest weight) for omic traits in internal validation, and found that genetic scores had a median R^2 that was 3.1-fold higher than the top variant. As expected, R^2 gain (1.7-fold) was smaller for scores with five variants or fewer.

Validation in cohorts of European ancestry

We performed external validation of SomaScan proteins in the FENLAND study[15]; Olink proteins in the Northern Swedish Population Health Study (NSPHS)[16] and the Orkney Complex Disease Study (ORCADES)[17]; Metabolon metabolites in ORCADES; and Nightingale metabolic traits in the UKB[3], Viking Health Study Shetland (VIKING)[18] and ORCADES studies (Extended Data Fig. 1 and Extended Data Table 1). For Metabolon and RNA-seq, we performed further validation in withheld sets of INTERVAL (Methods). Overall, we found that the performance of most genetic scores was consistent between internal and external validation in cohorts of European ancestry (Fig. 2, Extended Data Fig. 4 and Supplementary Figs. 6–10). As expected, we found that genetic scores with high variant missingness rates had attenuated power (Extended Data Fig. 5).

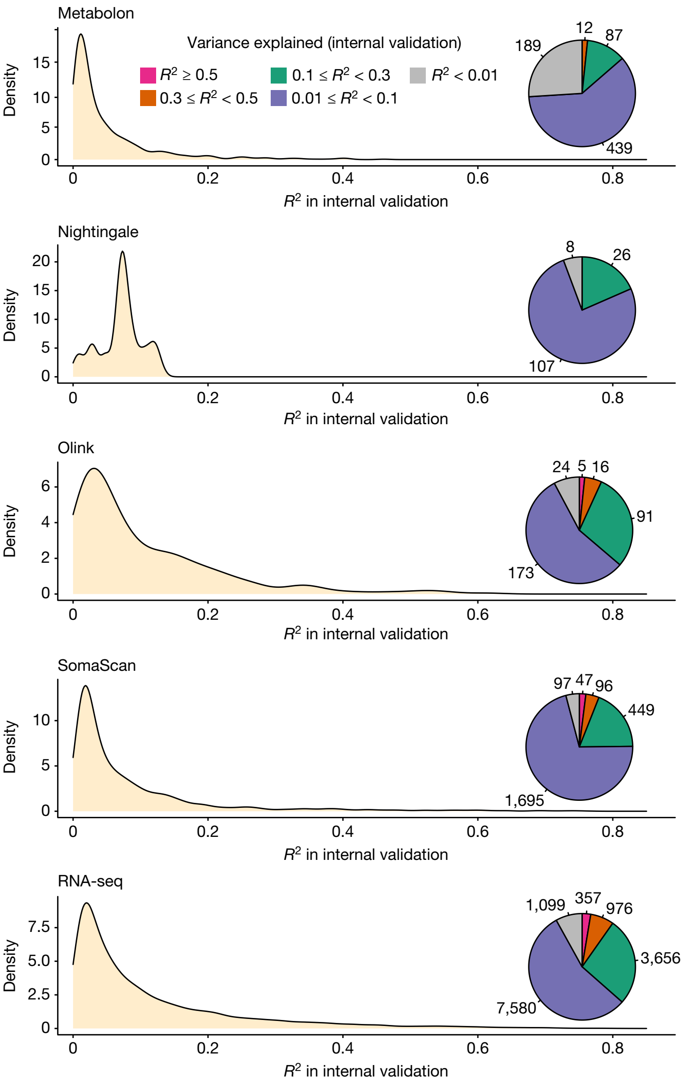


Fig. 1 | Performance of multi-omic genetic scores in internal validation. The variance explained in the measured biomolecular trait (R^2) by the genetic score is assessed in the internal validation set of INTERVAL (Methods). Pie charts reflect the number of genetic scores in a particular R^2 range.

SomaScan quantified 3,622 plasma proteins in INTERVAL, of which 2,384 proteins had at least one significant genetic variant that could be used for genetic-score development (Methods and Extended Data Fig. 3). Internal validation found that SomaScan genetic scores had a median R^2 value of 0.04 (interquartile range (IQR) = 0.08). Most SomaScan genetic scores (89%; n = 2,129) could be tested for external validation in the FENLAND study[15]. Overall, there was high consistency between internal and external R^2 performance (Fig. 2). We metricized validation performance using the slope (λ) of the line of best fit between internal and external R^2. For FENLAND, λ was 0.99. Of the 2,129 externally tested SomaScan genetic scores, we found 45 proteins (2%) with a majority of their variance explained ($R^2 > 0.50$) by the genetic score,

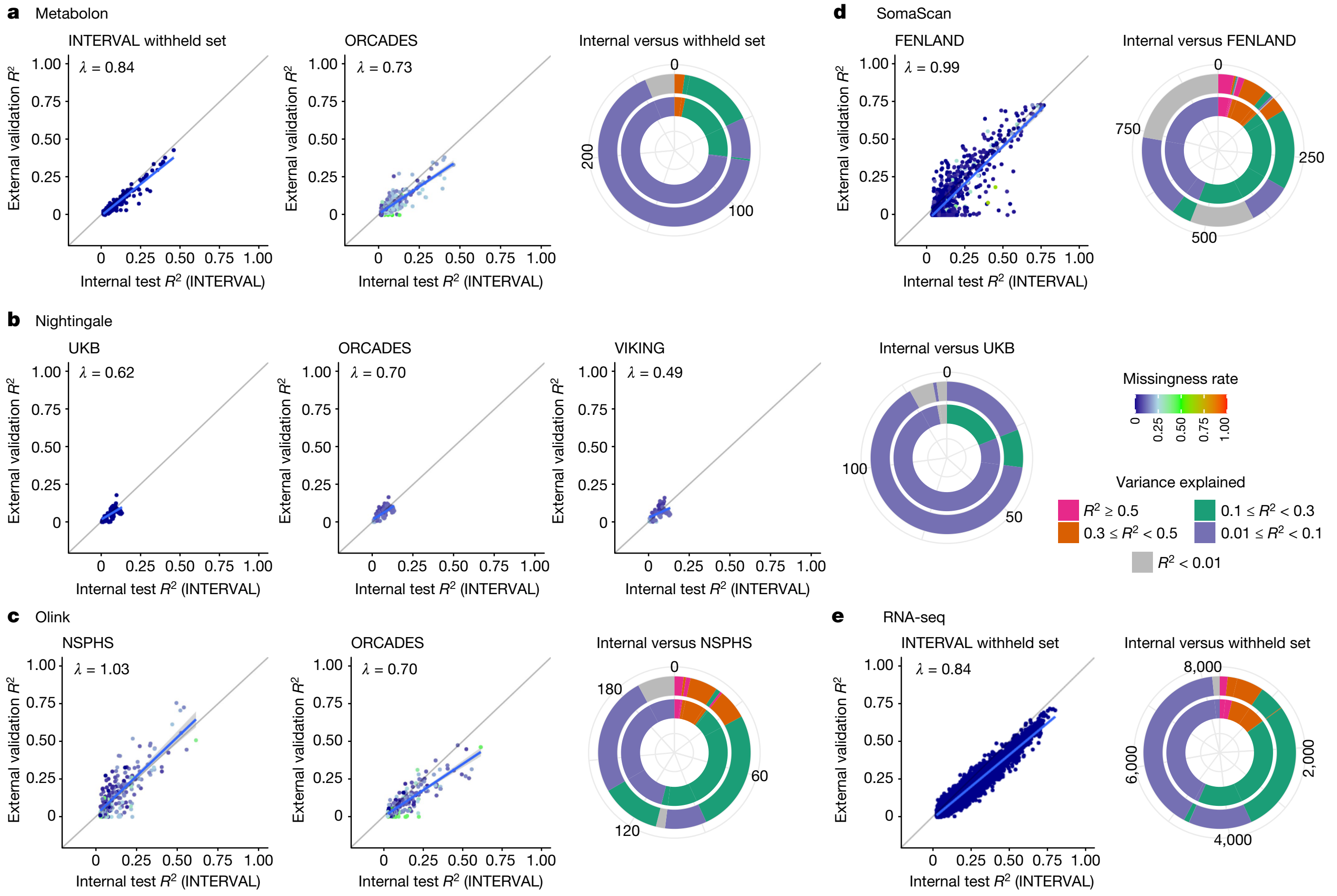


Fig. 2 | External validation of genetic scores in cohorts of European ancestry. **a**–**e**, Comparisons of R^2 in internal validation and external validation for each omic platform (Metabolon (**a**), Nightingale (**b**), Olink (**c**), SomaScan (**d**) and RNA-seq (**e**)) for genetic scores with Bonferroni-adjusted $P < 0.05$ in internal validation (two-sided t-test; correcting for 17,227 omic traits). Data points are coloured by variant missingness rate in the external cohort. Blue lines show fitted linear models and λ are model slopes. Concentric circles show the number of genetic scores in different ranges of R^2 in internal validation (inner ring) and external validation (outer ring).

including several with $R^2 > 0.70$ that are involved in innate and adaptive immune responses (CLEC12A, SIGLEC9, FCGR2A, FCGR2B and LILRB5). A total of 369 SomaScan proteins (17%) could be genetically predicted with $R^2 > 0.10$ in external validation.

Olink proteomics in INTERVAL quantified the levels of 368 plasma proteins from four panels (Inflammation, Cardiovascular 2, Cardiovascular 3 and Neurology), of which 308 unique proteins qualified for genetic-score development (Methods). Internal validation found that Olink genetic scores had a median R^2 value of 0.06 (IQR = 0.12). We were able to test 302 and 301 genetic scores in external cohorts of European ancestry, NSPHS (λ = 1.03) and ORCADES (λ = 0.70), respectively (Methods and Fig. 2). In both external validation cohorts, we found four proteins (FCGR2B, IL-6R, MDGA1 and SIRPA) that had a majority of their variance explained ($R^2 > 0.50$) by the genetic score (Fig. 2). As compared to SomaScan, a larger proportion of Olink proteins in NSPHS ($n = 117$; 39%) and ORCADES ($n = 87$; 29%) could be genetically predicted with $R^2 > 0.10$. Overall, we found consistency between validations in NSPHS and ORCADES (Supplementary Fig. 11).

Metabolon HD4 quantifies more than 900 plasma metabolites and was used here in 2 phases of the INTERVAL study (Methods). Phase 1 ($n = 8,153$) was used for the development and internal validation of Metabolon genetic scores, and phase 2 ($n = 8,114$) was used for external validation (no individuals overlapping between phases). We conducted further external validation in ORCADES. Internal validation found that Metabolon genetic scores had a median R^2 value of 0.02 (IQR = 0.05). A total of 726 Metabolon metabolites had significant genetic variants with which to construct genetic scores in INTERVAL, of which 527 and 455 metabolites (399 overlapping) could be externally validated in the phase 2 set (λ = 0.84) and ORCADES (λ = 0.73), respectively (Fig. 2). We again found broad consistency between the two external validation sets (Supplementary Fig. 11). There were no Metabolon genetic scores with $R^2 > 0.50$ in either the phase 2 set or ORCADES; however, six metabolites had $R^2 > 0.3$ in both the phase 2 set and ORCADES (four metabolites overlapping). Of metabolites that could be externally validated, 10% and 13% ($n = 50$ and $n = 59$) had an $R^2 > 0.10$ in the phase 2 set and ORCADES, respectively. The top-performing genetic scores included ethylmalonate (phase 2 set $R^2 = 0.43$; ORCADES $R^2 = 0.33$), N-acetylcitrulline (both phase 2 set and ORCADES $R^2 = 0.38$) and androsterone sulfate (phase 2 set $R^2 = 0.35$; ORCADES $R^2 = 0.17$).

Nightingale NMR was used to quantify 230 serum metabolic biomarkers from 45,928 INTERVAL participants. Our analyses focused on directly measured (non-derived) metabolic biomarkers, and genetic scores for 141 Nightingale biomarkers were developed using INTERVAL (Methods). Internal validation found that Nightingale genetic scores had a median R^2 value of 0.07 (IQR = 0.03). Genetic scores were externally validated in UKB, ORCADES and VIKING, with λ values of 0.62, 0.70 and 0.49, respectively (Fig. 2). Overall, genetic scores for Nightingale explained somewhat less variation in directly measured traits compared with other platforms (Fig. 2 and Extended Data Fig. 4). Across UKB, ORCADES and VIKING, 27 Nightingale metabolic biomarkers had an

$R^2 > 0.10$ in at least one external validation cohort, with no biomarkers having $R^2 > 0.30$. However, Nightingale genetic scores performed consistently across cohorts, with the same mean R^2 for all 141 Nightingale biomarkers of 0.06 across the 3 external cohorts. The most predictive genetic scores were related to low-density lipoprotein (LDL); for example, concentrations of cholesteryl esters in small LDL, cholesterol in small LDL, cholesteryl esters in medium LDL, cholesterol in medium LDL and LDL cholesterol (Supplementary Table 2).

Whole-blood RNA-seq from 4,778 individuals in INTERVAL was performed using Illumina NovaSeq (Methods); 4,136 individuals were used to develop and test genetic scores, and 598 individuals were kept as a withheld set for validation. INTERVAL RNA-seq data allowed for the construction of genetic scores using both *cis* and *trans* expression quantitative trait loci (eQTLs) for 13,668 genes, of which 12,958 (95%) could be assessed in the withheld validation set (Fig. 2). Internal validation found that RNA-seq genetic scores had a median R^2 value of 0.06 (IQR = 0.13). Overall, we found a strong correlation of R^2 between the internal and the withheld validation sets ($\lambda = 0.84$). There were 141 genes with $R^2 > 0.50$ in the withheld validation set, and 798 genes with $R^2 > 0.30$. The most predictive genes were those involved in proteolysis (*RNPEP*; $R^2 = 0.71$*)*, solute cotransport (*SLC12A7*; $R^2 = 0.72$), RNA helicase activity (*DDX11*; $R^2 = 0.71$) and spliceosome function (*U2AF1*; $R^2 = 0.72$).

Transferability of genetic scores

To assess the performance of the genetic scores developed in the predominantly European INTERVAL cohort in individuals of non-European ancestry, we used the Singapore Multi-Ethnic Cohort (MEC)[19] and the Jackson Heart Study (JHS)[20]. The MEC data comprised individuals from Chinese, Indian and Malay populations with matched genotypes, plasma Nightingale NMR and plasma SomaScan; and the JHS comprised African American individuals with matched genotypes and plasma SomaScan (Extended Data Table 1 and Methods).

Overall, we found that genetic scores developed in INTERVAL could predict the levels of Nightingale and SomaScan traits in individuals of Asian or African American ancestry, but, as expected, the performances of these scores were significantly reduced relative to European-ancestry cohorts (Fig. 3 and Extended Data Fig. 6). For Nightingale, genetic-score performance in external validation generally decreased from European ($\lambda = 0.62$ in UKB) to MEC Chinese ($\lambda = 0.41$) to MEC Indian ($\lambda = 0.35$) to MEC Malay ($\lambda = 0.15$) ancestries (Figs. 2 and 3a and Supplementary Fig. 12). However, of the 138 genetic scores that were statistically significant (nominal $P < 0.05$) in the UKB validation, nearly all were significantly predictive in Chinese (133), Indian (132) and Malay (134) ancestries (Supplementary Table 2). Genetic scores for LDL subclasses showed some of the most variable cross-ancestry R^2 values (Fig. 3b). The most consistently transferrable Nightingale genetic scores were the levels of triglycerides—either in total or the triglycerides in LDL, large LDL or medium HDL—and the level of phosphatidylcholines (Fig. 3b).

The transferability of SomaScan genetic scores was substantially greater than that of Nightingale (Fig. 3c). The λ for SomaScan in cohorts of European ancestry (FENLAND) was 0.99, as compared to 0.75, 0.68, 0.66 and 0.51 in the MEC Indian, MEC Malay, MEC Chinese and JHS African American groups, respectively (Figs. 2 and 3c and Supplementary Fig. 13). There were 1,309 genetic scores that were statistically significant in FENLAND external validation, which decreased to 935, 893, 806 and 451 in the MEC Indian, MEC Malay, MEC Chinese and JHS African American groups, respectively (Supplementary Table 4). The SomaScan genetic scores that attenuated most in cohorts of non-European ancestry were those for CD177 (a cell-surface protein on neutrophils and regulatory T cells) and LEPR (leptin receptor) (Fig. 3d). The most transferable SomaScan genetic scores included SIGLEC9 (which mediates the binding of sialic acid to cells), SIRPA (a cell-surface receptor for CD47 that is involved in signal transduction) and ACP1 (an acid and protein tyrosine phosphatase), with all internal- and external-validation R^2 values being greater than 0.50 (Fig. 3d). Given that the MEC used longitudinal sampling, we further assessed the longitudinal stability of Nightingale and SomaScan genetic scores across ancestries, finding strong consistency of genetic-score performance over a mean of 6.3 years (Methods and Extended Data Fig. 7).

Genetic control of biological pathways

Multi-omic genetic scores can be used to probe the relevance of a biological pathway to a particular trait or disease. To assess coverage of biological pathways by the proteomic genetic scores we present here, we applied genetic scores for SomaScan and Olink to assess the extent to which pathways are genetically controlled (Methods). Here, we considered all genetic scores with $R^2 > 0.01$ in internal validation (2,205 unique proteins) and jointly mapped the SomaScan and Olink scores onto data curated from Reactome[21] (Fig. 4a and Extended Data Fig. 8).

We found wide variation among the 27 super-pathways (the high-level grouping of related pathways that share a common biological theme or process in Reactome); some super-pathways were under relatively little genetic control (for example, chromatic organisation, or transport of small molecules), and others were under substantially greater genetic control (for example, digestion and absorption, or extracellular matrix organization) (Fig. 4a). Approximately 18% of proteins in the digestion and absorption super-pathway had an internal-validation $R^2 > 0.10$, and around 4% had $R^2 > 0.30$. For the lowest-level pathway annotation ($n = 1{,}717$) of the 27 super-pathways, we found that a majority ($n = 1{,}169$, 68%) were covered by at least one SomaScan or Olink genetic score, with internal-validation $R^2 > 0.01$ (Extended Data Fig. 8). For both the digestion and absorption and the extracellular matrix organization super-pathways, 25% and 42%, respectively, of lowest-level pathway annotations were covered by at least one SomaScan or Olink genetic score with internal $R^2 > 0.30$.

Phenome-wide association analysis

We next generated genetically predicted Metabolon, Nightingale, Olink, SomaScan and whole-blood RNA-seq data for UKB (Methods). Using these predicted multi-omics data of UKB, we performed a phenome-wide association study (PheWAS) using PheCodes[22] (ICD-9- and ICD-10-based diagnosis codes collapsed into hierarchical clinical disease groups; Methods). For simplicity and to maximize the number of qualified PheCodes, we focused the analysis on UKB individuals of white British ancestry. Multiple testing was controlled using a Benjamini–Hochberg false discovery rate (FDR) of 5% (Methods).

At an FDR of 5%, we identified 18,404 associations between genetic scores for the multi-omic traits and 18 categories of PheCodes (Fig. 4b). These associations comprised 1,668 for Metabolon HD4, 2,854 for Nightingale NMR, 740 for Olink, 5,501 for SomaScan and 7,641 for RNA-seq (Supplementary Tables 6 and 7). Circulatory, endocrine, metabolic and digestive diseases yielded the largest number of associations across platforms (Fig. 4b).

The PheWAS detected many known blood biomarkers as well as other notable associations. For example, total cholesterol was significantly associated with myocardial infarction (hazard ratio (HR) = 1.13 per s.d., FDR-corrected $P = 1 \times 10^{-61}$). Interleukin-6 (IL-6) pathways have been shown to have a causal association with coronary artery disease (CAD)[23], and the genetic scores for the IL-6 receptor (IL-6R) in SomaScan and Olink had $R^2 > 0.50$ in both internal and external validation, showing a high genetic predictability. Genetically predicted levels of IL-6R in both Olink and SomaScan were significantly associated with myocardial infarction (HR = 0.97 per s.d.;, FDR-corrected $P = 2 \times 10^{-4}$; HR = 0.97 per s.d., FDR-corrected $P = 4 \times 10^{-4}$, respectively). β-Microseminoprotein has been identified as a biomarker for prostate cancer[24] and PheWAS findings support this association (HR = 0.87 per s.d., FDR-corrected

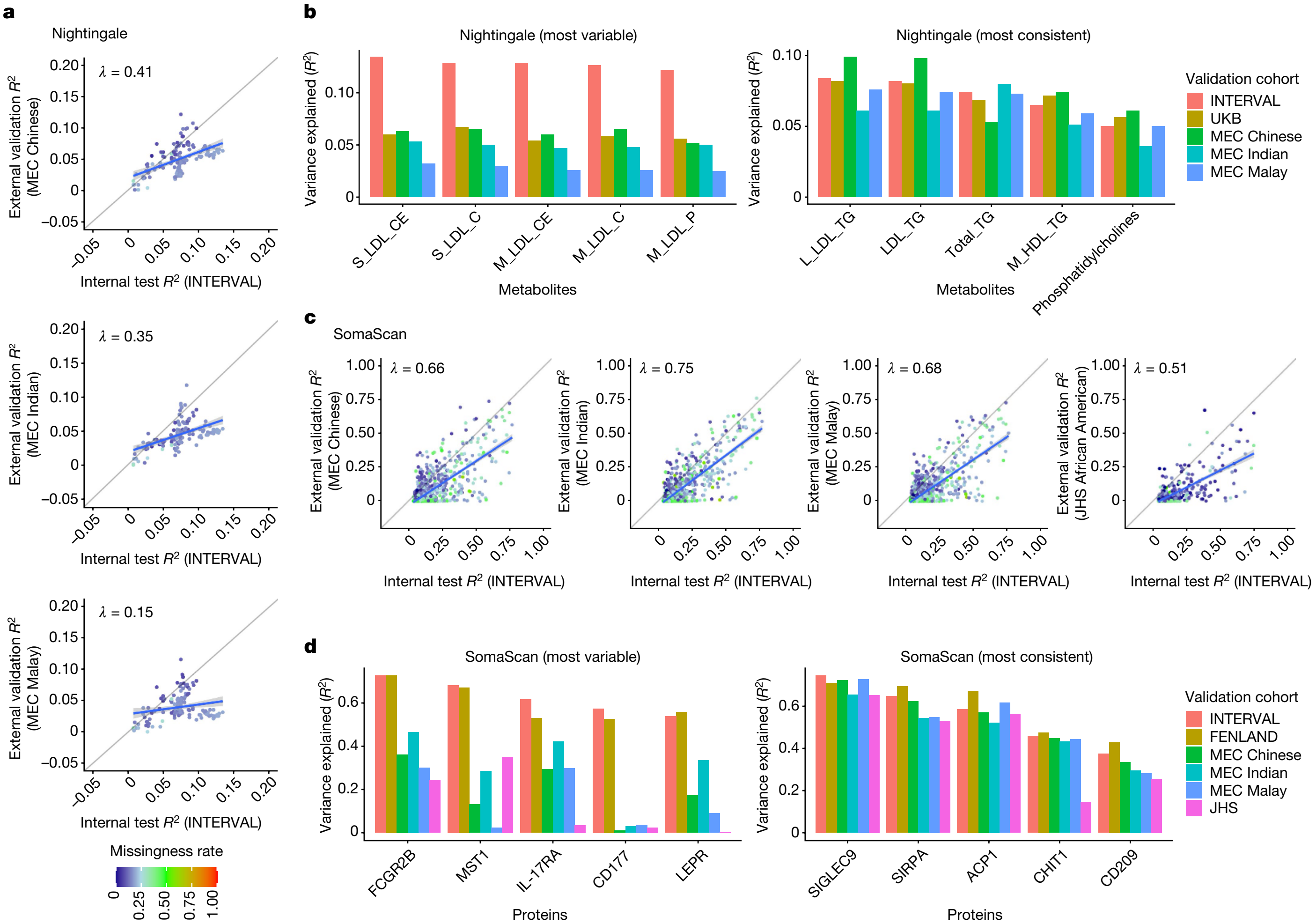


Fig. 3 | Transferability of genetic scores to cohorts of Asian and African American ancestries. **a**,**c**, Scatter plots showing comparisons of performance between internal validation and external validation in cohorts of non-European ancestry for Nightingale (**a**) and SomaScan (**c**) genetic scores. Transferability was tested for genetic scores with Bonferroni-adjusted $P < 0.05$ in internal validation (two-sided *t*-test; correcting for 17,227 omic traits). Data points are coloured by variant missingness rate in the external cohort. **b**,**d**, R^2 of genetic scores for Nightingale (**b**) and SomaScan (**d**) with the five most variable or five most consistent for prediction in multi-ancestry validation, as quantified by the mean absolute difference in R^2 for genetic scores with Nightingale $R^2 > 0.05$, SomaScan $R^2 > 0.30$ in internal validation. The most variable Nightingale genetic scores include cholesteryl esters in small LDL (S_LDL_CE), cholesterol in small LDL (S_LDL_C), cholesteryl esters in medium LDL (M_LDL_CE), cholesterol in medium LDL (M_LDL_C) and concentration of medium LDL particles (M_LDL_P); most transferable scores include triglycerides in large LDL (L_LDL_TG), triglycerides in LDL (LDL_TG), total triglycerides (Total_TG), triglycerides in medium HDL (M_HDL_TG) and phosphatidylcholines (Phosphatidylc).

$P = 3 \times 10^{-49}$). Genetically predicted sex hormone-binding globulin (SHBG) protein was associated with type 2 diabetes (HR = 0.98 per s.d., FDR-corrected $P = 0.03$), consistent with previous observational and genetic analyses[25]. Similarly, we found associations between proteins related to insulin signalling—for example, insulin receptor (INSR) and insulin-like growth factor 1 receptor (IGF1R)—and type 2 diabetes[26]; ABO (ref. 27) and type 2 diabetes; IL-6 and asthma[28]; and *HLA-DQA1* (and/or *HLA-DQB1*) and coeliac disease[29] (Supplementary Table 6).

Our results validate those of a previous study that identified putative causal plasma protein mediators between polygenic risk and incident cardiometabolic disease[4], including six novel putatively causal associations for CAD (Supplementary Table 6). Among the strongest signals, we found notable associations, including that of chronic pericarditis ($n = 266$ cases) with the genetically predicted gene expression of phospholipase *NAPEPLD* (HR = 0.88 per s.d., FDR-corrected $P < 1 \times 10^{-307}$), and rhesus isoimmunization in pregnancy ($n = 302$ cases) with the genetically predicted protein levels of ICAM4 (HR = 0.19 per s.d., FDR-corrected $P = 3 \times 10^{-93}$). ICAM4 is critical to the Landsteiner–Weiner blood system, which is genetically independent of the rhesus-factor blood-group system. Despite the *ICAM4* locus showing no significant association with rhesus isoimmunization in pregnancy (PheWeb[30]), our ICAM4 results show that genetic prediction of plasma proteins can identify biologically plausible candidate associations.

Biological insights

Here, we highlight a series of five findings in which multi-omic genetic scores are used to inform putative genetic mechanisms and pathophysiology. The first three of these investigate the metabolic mechanisms of relatively simple genetic scores for Metabolon traits, and the latter two comprise the integration of genetic scores across multiple omics to uncover pathway insights into disease biology.

The genetic score for histidine (Metabolon) consisted of three variants, two of which (rs61937878, rs117991621) are in the coding region of *HAL*, which encodes the enzymatic catalyst for the first reaction in histidine catabolism. We found that rs61937878 is also the sole variant in the genetic score for γ-glutamylhistidine. γ-Glutamylhistidine can be formed from the condensation of histidine and glutamate; thus,

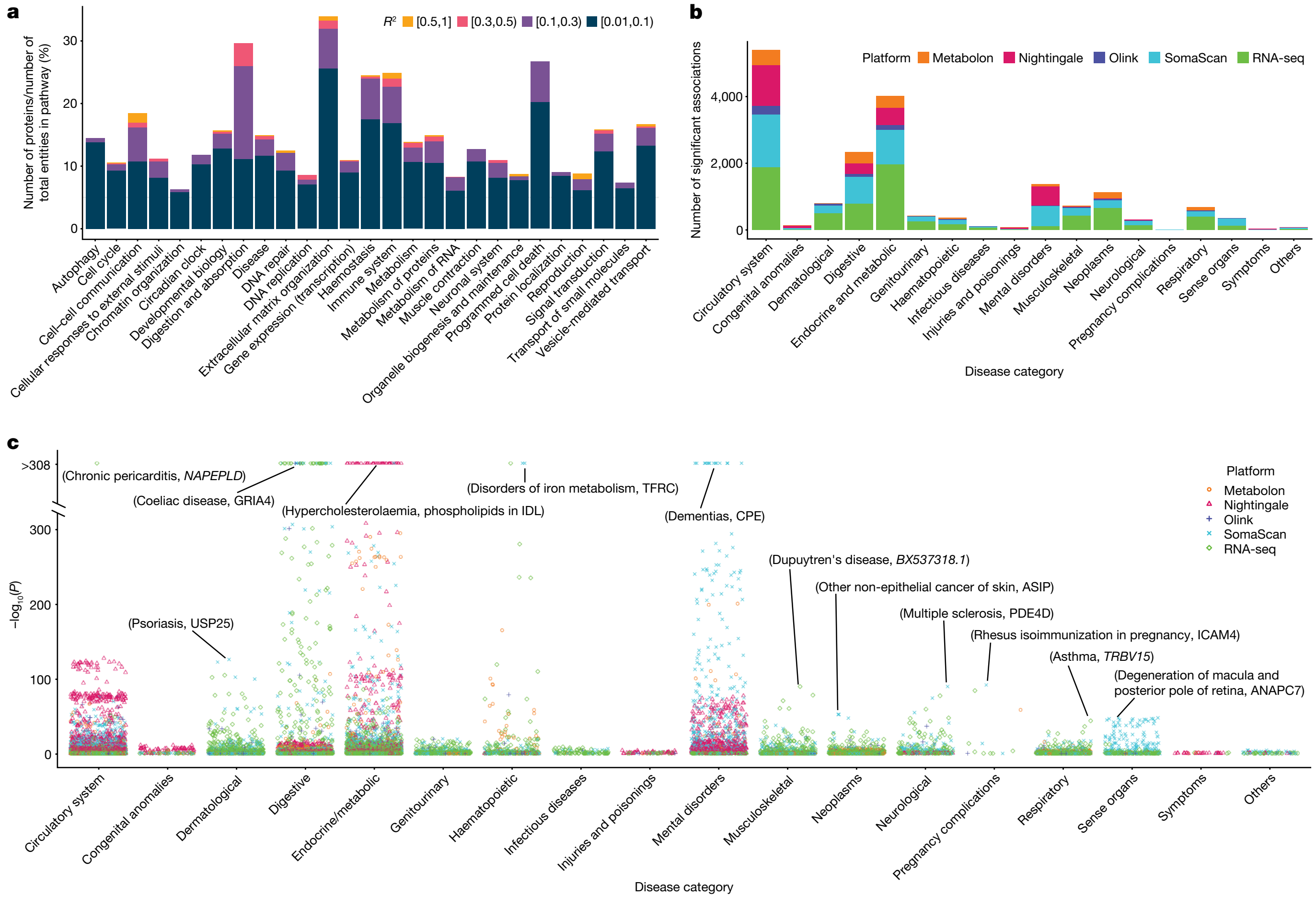


Fig. 4 | Applications of genetic scores of multi-omic traits. a, Genetic control of Reactome super-pathways using SomaScan and Olink genetic scores of varying R^2 in internal validation (Methods). **b**, PheWAS in UKB. Stacked bar plots show the number of detected significant associations by PheCode category of disease and omic platform (two-sided Wald test and FDR-corrected $P < 0.05$ for 11,576 tested traits). **c**, Strength of associations by category of disease and omic platform. Association with the lowest *P* value for each disease category are labelled.

we hypothesize that this genetic variant in *HAL* changes the levels of γ-glutamylhistidine by modulating histidine availability.

The 2-methylbutyrylcarnitine (Metabolon) genetic score contained five variants, including rs11753995, which is located within *SLC22A1*, encoding a transmembrane transporter of 2-methylbutyrylcarnitine and other acyl-carnitines[31]. Notably, two variants (rs200800380 and rs274555) in this genetic score are located in *SLC22A4* and *SLC22A5*, respectively, which are involved in carnitine transport[32]. The 2-methylbutyrylcarnitine genetic score also contains an intronic variant (rs4128783) which maps to the gene encoding Acyl-CoA dehydrogenase short/branched chain (ACADSB). ACADSB catalyses the dehydrogenation of 2-methylbutyryl-CoA. Because 2-methylbutyrylcarnitine is produced by transferring the acyl chain from 2-methylbutyryl-CoA to carnitine, these genetic variants (rs200800380, rs274555 and rs4128783) might influence the levels of 2-methylbutyrylcarnitine by modulating the availability of substrates.

The genetic score for DSGEGDFXAEGGGVR (Metabolon) contained a single variant (rs567455090) intronic to *SLC9A1*. Notably, SLC9A1 is a transmembrane exchanger of Na^+/H^+ that regulates the pH and volume of platelets and has a significant role in their activation[33]. Activated platelets secrete α-granules of thrombin precursor (prothombin) and fibrinogen. DSGEGDFXAEGGGVR is a peptide derived from the cleavage of fibrinogen by thrombin[34]; thus, rs567455090 might modulate the function and activation of platelets that, in turn, change the levels of DSGEGDFXAEGGGVR.

Our PheWAS in UKB identified a series of gene transcripts and proteins in the JAK–STAT signalling pathway as being associated with the risk of CAD (Fig. 5a,b). JAK–STAT regulates cellular proliferation, differentiation and apoptosis and also has a role in modulating inflammation. The SomaScan levels of AKT2 and CTF1 and transcript levels of *STAT1* were associated with an increased risk of CAD, consistent with the anti-atherogenic effects of targeting these genes in mouse models of hypocholesterolemia[35–37]. The transcript levels of *PIM1* and *CISH1* (also known as *SOCS1*), which inhibit the JAK–STAT pathway[38,39], were associated with a decreased risk of CAD. We further found that the levels of IL-6 (Olink) and IL-6R (Olink and SomaScan) were associated with CAD. Consistent with our findings, circulating IL-6 is a well-established biomarker of CAD, and IL-6–IL-6R signalling has been shown to have a putative causal effect on CAD[23]. Our PheWAS supports the investigation of inhibitors of JAK–STAT, which are clinically approved for chronic inflammatory disorders, as candidates against CAD[40].

We also identified transcripts and proteins involved in WNT signalling (Fig. 5c–e) as associated with hypothyroidism. Notably, there is a well-established cross-talk between WNT and thyroid hormone signalling: thyroid hormone nuclear receptors can modulate the expression, stability and localization of proteins of the WNT pathway, whereas the latter modulates thyroid hormone activity by controlling the expression of deiodinases[41], enzymes that regulate thyroid hormones. Furthermore, WNT signalling is active in thyroid cells[42] and is thought to contribute to thyroid homeostasis[43]. In this regard, pharmacological

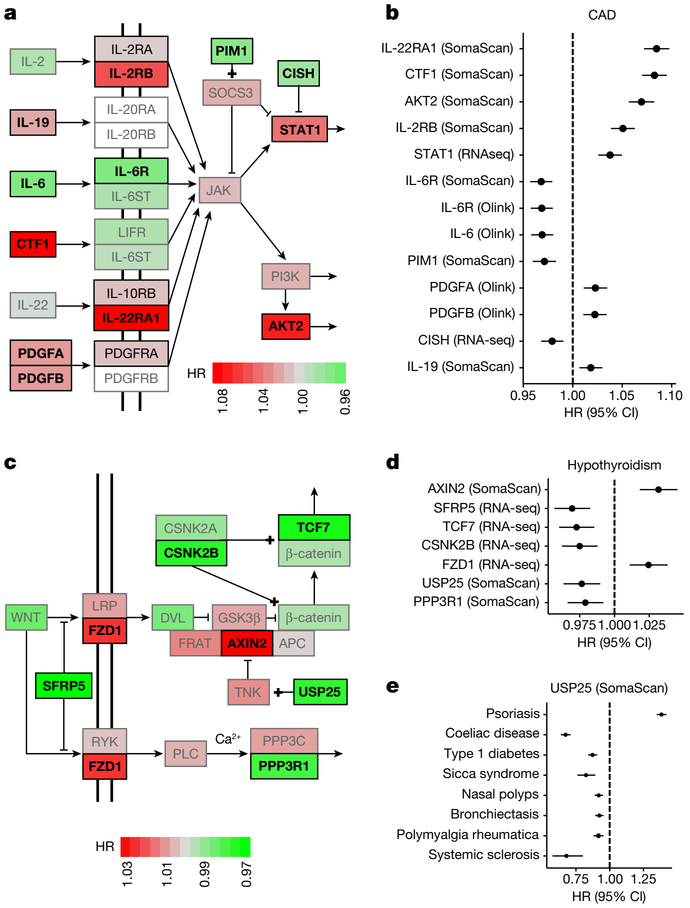


Fig. 5 | JAK–STAT and WNT signalling pathways. a,**c**, Pathway diagrams for JAK–STAT (**a**) and WNT (**c**) signalling. Nodes are coloured on the basis of the HR of the genetic score for CAD (**a**) and hypothyroidism (**c**). Nodes are white if there is not a corresponding genetic score. The most significant HR across omic platforms is used at each node. Nodes are bold if the genetic score had FDR-adjusted $P < 0.05$ (two-sided Wald test and correcting for 11,576 tested traits). **b**,**d**, Forest plots of FDR-significant HRs and 95% confidence intervals (CI) for CAD (**b**; $n = 28,854$ cases and 390,159 controls) and hypothyroidism (**d**; $n = 21,871$ cases and 404,440 controls) for genetic scores in JAK–STAT (**b**) or WNT (**d**) signalling. **e**, Forest plot of HRs and 95% CI for the genetic score of USP25 (SomaScan) across several diseases.

activation of WNT has been shown to impair thyroid development in zebrafish[44], and a risk allele for congenital hypothyroidism has been identified within enhancer regions of two WNT pathway genes[45]. We also found that the genetic score for USP25 (SomaScan) was associated with a decreased risk of hypothyroidism. USP25 is a deubiquitinating enzyme that can activate WNT by stabilizing TNKS1 (ref. 46). USP25 also modulates inflammatory responses[47], contributes to metabolic adaptation to hypoxia[48] and inhibits the degradation of abnormal proteins[49]. Notably, we found that USP25 was also associated with a wide range of diseases, including psoriasis, type 1 diabetes, sicca syndrome, bronchiectasis, polymyalgia rheumatica, nasal polyps and systemic sclerosis, making USP25 a potentially useful biomarker and therapeutic target.

The OmicsPred portal

We developed an online portal (https://www.omicspred.org/) to facilitate open dissemination of the genetic scores, detailed validation results and visualizations. OmicsPred also serves as an online updatable resource, which will allow future expansion of the omics platforms, multi-ancestry transferability and the development of more powerful genetic scores, as well as results from its applications (Extended Data Fig. 9).

The portal presents genetic scores of multi-omic traits by platform, in which users can access summary statistics of the training and validation cohorts, and download the corresponding model files for genetic scores (that is, variants and weights). Users can visualize validation results by selected performance metrics (for example, R^2 or Spearman's rho) and cohort(s), together with detailed trait and validation information, and they can easily search the portal to find multi-omic traits of interest, either by name or through related descriptions. OmicsPred also hosts descriptions and summary results from applications of the genetic scores (for example, the PheWAS above). Moreover, OmicsPred serves as a central resource to which users can submit their multi-omic genetic scores so that they can be openly distributed to the community.

Discussion

We have developed genetic scores for more than 17,000 multi-omic traits across 5 platforms covering proteomics, metabolomics and transcriptomics. The relative predictive values and robustness of the genetic scores were assessed in external validations using cohorts of European, Asian and African American ancestries; longitudinal stabilities of the genetic score performances were established across ancestries; and the utility of the multi-omic genetic scores was demonstrated by elucidating the relative genetic control of biological pathways and by identifying disease associations from a phenome-wide scan of predicted multi-omic data in UKB. Finally, we developed an open resource OmicsPred (https://www.omicspred.org/) to publicly disseminate and continuously enhance the value of multi-omic genetic scores.

Although the utility of predicted transcriptomic data for cohorts with genome-wide genotype data has been shown[1], our work substantially extends these foundations using a large multi-omic cohort, quantifying both the intra- and the inter-ancestry reliability of proteomic and metabolomic genetic scores across multiple platforms. We generated a predicted multi-omic dataset for UKB and showed that PheWASs can uncover many known and novel omic associations with disease. In turn, this raises the question of what is a meaningful predictive value for a genetic score—to which, given each user's own particular application, there is no simple answer. Given that the increase in sample size required to detect an association for a noisy explanatory variable can be estimated by n/R (where n is the sample size required if no measurement error exists and R is the reliability coefficient)[11], even genetic scores of apparently low predictive value may be powerful enough to detect true associations at the sample sizes of current and forthcoming biobanks. This suggests that large biobanks could reliably and efficiently test trait–disease associations using genetically predicted multi-omic data, before committing to (frequently expensive) data generation.

Our study has limitations. Although blood is a key tissue of broad utility, it is likely a correlate and not the main site of causal biomolecular functions. Genetic-score validity was generally consistent across cohorts; however, performance was affected by technical factors (for example, serum versus plasma, batch variations, fasting versus non-fasting samples and genetic variant missingness), participant demographics, genetic factors (for example, allele frequency and LD differences) and environmental factors (for example, dietary differences). Genetic scores might also pick up differences in molecular traits shared by multiple platforms (for example, Olink and SomaScan). Despite genetic scores for most shared proteins being consistently predictive across platforms, large differences can be due to technical factors such as binding affinity (Methods) as assessed in a previous study[15]. The attenuated performance of polygenic scores across ancestries is well-known[50] and our analysis also found this in multi-omics data. Multi-omics for populations of non-European ancestry will become more common, and we see a key role for OmicsPred in facilitating robust genetic scores that enable multi-omic prediction in diverse populations. Given that genetic prediction and its methodology is a rapidly evolving field, we further acknowledge that there are many highly sophisticated machine learning approaches that may improve the performance and/or transferability of genetic scores. We selected Bayesian ridge because it has been shown to perform well relative to other genetic-score approaches in both a previous study[7] and a benchmark performed here. In addition, Bayesian ridge has been shown to scale well to large numbers of traits, thus improving computational efficiency and consistency with green computing[7,13]. Optimal variant-selection thresholds may also vary across traits. Finally, although OmicsPred provides a key first step towards a better understanding of the distributions of clinically or therapeutically important biomarkers under high genetic control, more research is needed to understand to what extent genetic scores for multi-omic traits may one day be of clinical use.

Future avenues for research include the expansion of OmicsPred to additional platforms and/or cohorts, multi-ancestry training for improved prediction and causal inference. In summary, we have developed, validated and applied multi-omic genetic scores for more than 17,000 traits and made them publicly accessible through our OmicsPred resource (https://www.omicspred.org/), facilitating the generation and application of multi-omics data at scale for the wider community.

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/s41586-023-05844-9.

1. 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).
2. Moore, C. et al. The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials* **15**, 363 (2014).
3. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
4. Ritchie, S. C. et al. Integrative analysis of the plasma proteome and polygenic risk of cardiometabolic diseases. *Nat. Metab.* **3**, 1476–1483 (2021).
5. Lambert, S. A. et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* **53**, 420–425 (2021).
6. Adeyemo, A. et al. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat. Med.* **27**, 1876–1884 (2021).
7. Xu, Y. et al. Machine learning optimized polygenic scores for blood cell traits identify sex-specific trajectories and genetic correlations with disease. *Cell Genomics* **2**, 100086 (2022).
8. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
9. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
10. Mosley, J. D. et al. Probing the virtual proteome to identify novel disease biomarkers. *Circulation* **138**, 2469–2481 (2018).
11. Hutcheon, J. A., Chiolero, A. & Hanley, J. A. Random measurement error and regression dilution bias. *Br. Med. J.* **340**, 1402–1406 (2010).
12. Pividori, M., Schoettler, N., Nicolae, D. L., Ober, C. & Im, H. K. Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-wide studies. *Lancet Respir. Med.* **7**, 509–522 (2019).
13. Lannelongue, L., Grealey, J., Bateman, A. & Inouye, M. Ten simple rules to make your computing more environmentally sustainable. *PLoS Comput. Biol.* **17**, e1009324 (2021).
14. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).
15. Pietzner, M. et al. Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).
16. Igl, W., Johansson, A. & Gyllensten, U. The Northern Swedish Population Health Study (NSPHS)—a paradigmatic study in a rural population combining community health and basic research. *Rural Remote Health* **10**, 1363 (2010).
17. McQuillan, R. et al. Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83**, 359 (2008).
18. Kerr, S. M. et al. An actionable *KCNH2* Long QT Syndrome variant detected by sequence and haplotype analysis in a population research cohort. *Sci. Rep.* **9**, 10964 (2019).
19. Tan, K. H. X. et al. Cohort profile: the Singapore Multi-Ethnic Cohort (MEC) study. *Int. J. Epidemiol.* **47**, 699–699j (2018).
20. Katz, D. H. et al. Whole genome sequence analysis of the plasma proteome in black adults provides novel insights into cardiovascular disease. *Circulation* **145**, 357–370 (2021).
21. Fabregat, A. et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).
22. Patrick, et al. Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. *JMIR Med. Inf.* **7**, e14325 (2019).
23. Sarwar, N. et al. Interleukin-6 receptor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies. *Lancet* **379**, 1205–1213 (2012).
24. Haiman, C. A. et al. Levels of β-microseminoprotein in blood and risk of prostate cancer in multiple populations. *J. Natl Cancer Inst.* **105**, 237–243 (2013).
25. Ding, E. L. et al. Sex hormone-binding globulin and risk of type 2 diabetes in women and men. *N. Engl. J. Med.* **361**, 1152–1163 (2009).
26. Saini, V. Molecular mechanisms of insulin resistance in type 2 diabetes mellitus. *World J. Diabetes* **1**, 68 (2010).
27. Qi, L. et al. Genetic variants in ABO blood group region, plasma soluble E-selectin levels and risk of type 2 diabetes. *Hum. Mol. Genet.* **19**, 1856–1862 (2010).
28. Peters, M. C. et al. Plasma interleukin-6 concentrations, metabolic dysfunction, and asthma severity: a cross-sectional analysis of two cohorts. *Lancet Respir. Med.* **4**, 574–584 (2016).
29. Banaganapalli, B. et al. Exploring celiac disease candidate pathways by global gene expression profiling and gene network cluster analysis. *Sci. Rep.* **10**, 16290 (2020).
30. Gagliano Taliun, S. A. et al. Exploring and visualizing large-scale genetic associations by using PheWeb. *Nat. Genet.* **52**, 550–552 (2020).
31. Kim, H. I. et al. Fine mapping and functional analysis reveal a role of SLC22A1 in acylcarnitine transport. *Am. J. Hum. Genet.* **101**, 489 (2017).
32. Tamai, I. Pharmacological and pathophysiological roles of carnitine/organic cation transporters (OCTNs: SLC22A4, SLC22A5 and Slc22a21). *Biopharm. Drug Dispos.* **34**, 29–44 (2013).
33. Chang, H. B., Gao, X., Nepomuceno, R., Hu, S. & Sun, D. Na+/H+ exchanger in the regulation of platelet activation and paradoxical effects of cariporide. *Exp. Neurol.* **272**, 11–16 (2015).
34. de Vries, P. S. et al. Whole-genome sequencing study of serum peptide levels: the Atherosclerosis Risk in Communities study. *Hum. Mol. Genet.* **26**, 3442–3450 (2017).
35. Babaev, V. R. et al. Loss of 2 Akt (protein kinase B) isoforms in hematopoietic cells diminished monocyte and macrophage survival and reduces atherosclerosis in *Ldl* receptor-null mice. *Arterioscler. Thromb. Vasc. Biol.* **39**, 156–169 (2019).
36. Miteva, K. et al. Cardiotrophin-1 deficiency abrogates atherosclerosis progression. *Sci. Rep.* **10**, 5791 (2020).
37. Agrawal, S. et al. Signal transducer and activator of transcription 1 is required for optimal foam cell formation and atherosclerotic lesion development. *Circulation* **115**, 2939–2947 (2007).
38. Peltola, K. J. et al. Pim-1 kinase inhibits STAT5-dependent transcription via its interactions with SOCS1 and SOCS3. *Blood* **103**, 3744–3750 (2004).
39. Khor, C. C. et al. CISH and susceptibility to infectious diseases. *N. Engl. J. Med.* **362**, 2092–2101 (2010).
40. Baldini, C., Moriconi, F. R., Galimberti, S., Libby, P. & De Caterina, R. The JAK–STAT pathway: an emerging target for cardiovascular disease in rheumatoid arthritis and myeloproliferative neoplasms. *Eur. Heart J.* **42**, 4389–4400 (2021).
41. Skah, S., Uchuya-Castillo, J., Sirakov, M. & Plateroti, M. The thyroid hormone nuclear receptors and the Wnt/β-catenin pathway: an intriguing liaison. *Dev. Biol.* **422**, 71–82 (2017).
42. Chen, G. et al. Regulation of GSK-3β in the proliferation and apoptosis of human thyrocytes investigated using a GSK-3β-targeting RNAi adenovirus expression vector: involvement the Wnt/β-catenin pathway. *Mol. Biol. Rep.* **37**, 2773–2779 (2009).
43. Ely, K. A., Bischoff, L. A. & Weiss, V. L. Wnt signaling in thyroid homeostasis and carcinogenesis. *Genes* **9**, 204 (2018).
44. Haerlingen, B. et al. Small-molecule screening in zebrafish embryos identifies signaling pathways regulating early thyroid development. *Thyroid* **29**, 1683–1703 (2019).
45. Narumi, S. et al. GWAS of thyroid dysgenesis identifies a risk locus at 2q33.3 linked to regulation of Wnt signaling. *Hum. Mol. Genet.* **31**, 3967–3974 (2022).
46. Xu, D. et al. USP25 regulates Wnt signaling by controlling the stability of tankyrases. *Genes Dev.* **31**, 1024–1035 (2017).
47. Lin, D. et al. Induction of USP25 by viral infection promotes innate antiviral responses by mediating the stabilization of TRAF3 and TRAF6. *Proc. Natl Acad. Sci. USA* **112**, 11324–11329 (2015).
48. Nelson, J. K. et al. USP25 promotes pathological HIF-1-driven metabolic reprogramming and is a potential therapeutic target in pancreatic cancer. *Nat. Commun.* **13**, 2070 (2022).
49. Blount, J. R., Burr, A. A., Denuc, A., Marfany, G. & Todi, S. V. Ubiquitin-specific protease 25 functions in endoplasmic reticulum-associated degradation. *PLoS One* **7**, e36542 (2012).
50. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [2]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [3]Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge, UK. [4]British Heart Foundation Centre of Research Excellence, School of Clinical Medicine, University of Cambridge, Cambridge, UK. [5]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore, Singapore. [6]Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, UK. [7]MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge School of Clinical Medicine, Cambridge, UK. [8]Computational Medicine, Berlin Institute of Health (BIH) at Charité – Universitätsmedizin Berlin, Berlin, Germany. [9]Precision Healthcare University Research Institute, Queen Mary University of London, London, UK. [10]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK. [11]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK. [12]Division of Cardiovascular Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. [13]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [14]Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia. [15]Department of Immunology and Inflammation, Faculty of Medicine, Imperial College London, London, UK. [16]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [17]BioMarin Pharmaceutical, Novato, CA, USA. [18]Department of Haematology, University of Cambridge, Cambridge, UK. [19]NIHR Blood and Transplant Research Unit in Donor Health and Behaviour, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [20]Genomics Research Centre, Human Technopole, Milan, Italy. [21]Health Data Science Research Centre, Human Technopole, Milan, Italy. [22]Department of Medicine, National University of Singapore and National University Health System, Singapore, Singapore. [23]Departments of Exercise and Nutrition Sciences and Epidemiology, Milken Institute School of Public Health, The George Washington University, Washington, DC, USA. [24]Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK. [25]Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK. [26]Health Data Research UK, London, UK. [27]Broad Institute of Harvard University and Massachusetts Institute of Technology, Cambridge, MA, USA. [28]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [29]Pfizer Worldwide Research, Development and Medical, Stockholm, Sweden. [30]MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK. [31]The Alan Turing Institute, London, UK. ✉e-mail: yx322@medschl.cam.ac.uk; mi336@medschl.cam.ac.uk

Methods

INTERVAL cohort and data quality control

The INTERVAL study[2] is a randomized trial of around 50,000 healthy blood donors, who were recruited at 25 centres of England's National Health Service Blood and Transplant (NHSBT) and aged 18 years or older at recruitment. This trial aimed to study the safety of a varying frequency of blood donation, and all of the participants completed an online questionnaire when joining the study about their demographics and lifestyles, including information such as age, sex, weight, height, alcohol intake, smoking habits and diet. This trial is registered with the ISRCTN (ISRCTN24760606). All participants have given informed consent and this study was approved by the National Research Ethics Service (11/EE/0538).

In total, 48,813 INTERVAL samples were genotyped using the Affymetrix UK Biobank Axiom array in 10 batches, which assays approximately 830,000 variants. The variants were phased using SHAPEIT3 and imputed on a combined 1000 Genomes Phase 3-UK10K reference panel. Affymetrix implemented standard quality control (QC) procedures during the genotype calling pipeline, excluding samples with a poor signal intensity (dish QC < 0.82) and samples with a low call rate (<97%) on the basis of around 20,000 high-quality probe sets. Variants were excluded if they had a low call rate (<95%), had more than three clusters (indicative of off-target measurement), had cluster statistics (Fisher's linear discriminant, heterozygous cluster strength or homozygote ratio offset) indicative of poor-quality genotyping or were complicated multi-allelic variants that couldn't easily be called. Then, within-batch sample and variant QC was performed, in which non-best probe sets were excluded to leave a single probe set per variant. As visual inspection of cluster plots had identified that some variants, particularly rare variants, had minor-allele homozygotes incorrectly called owing to the presence of an extreme intensity outlier, we failed variants from a batch if: (1) the variant had fewer than ten called minor-allele homozygotes; (2) the cluster plot contained at least one sample with an intensity at least twice as far from the origin as the next most extreme sample; or (3) the outlying sample (s) had an extreme polar angle (<15° or >75°) in the direction of the minor allele. We excluded duplicate samples and samples that were clearly not of European ancestry using a set of high-quality autosomal variants, defined as those with a minor-allele frequency (MAF) > 0.05, Hardy–Weinberg equilibrium (HWE) $P > 1 \times 10^6$ and $r^2 \leq 0.2$ between pairs of variants. Duplicate samples were defined as those with $\widetilde{\pi} \geq 0.9$ using the PLINK v.1.9 method-of-moment identity-by-descent approach[51] and non-European samples were defined as those with scores < 0 on genetic PC1 or genetic PC2 after a principal component analysis (PCA) including INTERVAL samples with 1000 Genomes major ancestry populations[52]. More details on the genotyping and sample QC for INTERVAL data can be found in the previous study[53]. After QC steps, it finally resulted in 10,572,788 variants for 43,059 samples. The number of valid samples in each platform for genetic-score construction (Extended Data Table 1) excluded samples that did not pass the QC.

Using the aptamer-based SomaScan assay (v.3), this study profiled plasma proteins of 3,562 participants in two batches (n = 2,731 and n = 831), of which 3,175 samples remained for analysis after QC. The detailed steps for measurement and QC of the protein levels using the SomaScan array in INTERVAL have been previously described[4,54]. In summary, the relative concentration of 3,622 proteins (or protein complexes) targeted by 4,034 modified aptamers (SOMAmer reagents, referred to as SOMAmers) on the array were measured from 150-µl aliquots of plasma at SomaLogic (Boulder, CO, USA). QC was performed at the sample and SOMAmer levels by Somalogic, which uses the control aptamers and calibrator samples to correct for systematic variability in hybridization, within-run and between-run technical variability. For this study, we did not exclude protein aptamers with greater than 20% coefficient of variation in either batch, but excluded those aptamers targeting non-human proteins. We also excluded aptamers that, since the original quantification in INTERVAL, had been (1) deprecated by SomaLogic; (2) found to be measuring the fusion construct rather than the target protein; or (3) measuring a common contaminant[4], which finally filtered the data to 3,793 high-quality aptamers targeting 3,442 proteins. Within each batch, the relative protein abundances were natural log-transformed, and then adjusted for age, sex, the first three genetic principal components and duration between blood draw and sample processing (binary, 1 day versus more than 1 day). The protein residuals from this linear regression were finally inverse-rank normalized and used as phenotype values for their genome-wide association study (GWAS), which has been previously reported in detail[54]. These normalized phenotype values were further adjusted for batch effect and the first ten genetic principal components, which were used as the phenotype values for the genetic-score model training and internal validation (Supplementary Table 8).

Using Olink proximity extension assays[55], the INTERVAL study measured the plasma protein abundance of around 5,000 samples on 4 Olink panels: Inflammation-1 (INF-1), Cardiovascular II (CVD-2), Cardiovascular III (CVD-3) and Neurology (NEUR), each of which includes 92 proteins. For the INF-1, CVD-2 and CVD-3 panels, samples were assayed in two equal batches and their protein levels were pre-processed and quality controlled by Olink using NPX Manager software. Protein levels were then regressed on age, sex, sample-measurement plate, time from blood draw to sample processing (number of days) and season (categorical: spring, summer, autumn and winter), and inverse-normal rank-transformed. Details on QC and GWAS for proteins on these three panels have been described previously[56]. Owing to differences in timing and funding, the NEUR panel was treated separately from other three panels for QC purposes. In detail, samples were assayed in one large batch, and trait levels were also processed by the NPX software and final measurements were presented as NPX values on a $\log_2$ scale (that is, a one-unit increase represents a doubling of the protein level). We removed 187 measurements flagged by Olink as potentially having technical issues and 147 samples of potentially non-European origin as determined by PCA, which left 4,811 measurements proceeding to standard QC assessments. We also checked for missing measurements and measurements below the limit of detection. No missing measurements were found. Eight out of 92 proteins had values below the limit of detection (LOD), of which 4 (HAGH, BDNF, GDNF and CSF3) had more than 5% of measurements below the LOD so were not taken forward for further analyses. No participant had more than 4% of protein measurements below the LOD, and we did not observe an overrepresentation of particular proteins below LOD for specific participants. Protein measurements were then adjusted for age, sex, season when the blood sample was drawn (spring, summer, autumn and winter) and the first ten genetic PCs, residuals of which were further inverse-normal rank-transformed for their association analyses. We ran association tests using SNPTEST (v.2.5.2), with method 'expected', filtering out variants with a minor-allele count (MAC) lower than 10 for analyses. It was noted that there are a small number of shared proteins across the four Olink panels (detailed numbers of proteins and participants per panel after QC are given in Supplementary Table 9). To avoid duplication in genetic-score construction, these shared proteins were merged by averaging their protein levels on each sample across panels, and taken as a unique protein. All of the genetic variants identified in GWASs for the same protein across multiple panels were combined (if different) for the development of its genetic score. The normalized protein levels of 308 unique proteins were adjusted for the first ten genetic principal components (if not adjusted previously), which were used as phenotype values for genetic-score model construction and testing in INTERVAL.

The Metabolon HD4 Discovery platform (Metabolon, Durham, NC, USA) was used to measure the plasma metabolites of INTERVAL participants. Four subcohorts of 4,316, 4,637, 3,333 and 4,802 participants were created through random sampling from the INTERVAL study and metabolites were measured within the 4 subcohorts (or batches) separately at two time phases of the study (two batches at each phase). Samples of the first two batches were used as training data for GWAS and the genetic-score development of metabolite traits in the platform, and samples of the other two batches were held out for external validation purposes. The two subsets of INTERVAL data were put through the same QC process as described below before performing training or validation. No significant technical variability was found between batches and hence batches within a subset (phase 1 or 2) were merged before the QC and genetic analysis including batch as a covariate to adjust for any residual batch effects. In the first step, samples with missing values for each of the ion counts for a specific metabolite fragment ('OrigScale') were identified. These sample-specific metabolite values were set to missing within the scaled and imputed data ('ScaledImpData'), which contains for each metabolite the values within the OrigScale median normalized for run day (median set to 1 for run-day batch). Metabolites were then excluded if measured in only one batch or in fewer than 100 samples. Metabolite values were then winsorized to five standard deviations from the mean in cases in which the values exceeded mean ± 5 × s.d. of the metabolite. Each metabolite was then log (natural) transformed before calculating the residuals adjusted for age, sex, Metabolon batch, INTERVAL recruitment centre, plate number, appointment month, the lag time between the blood donation appointment and sample processing, and the first five genetic principal components. Before the genetic analysis, these residuals were standardized to a mean of 0 and standard deviation of 1. GWASs were then performed for each trait using the standardized trait values on samples of the first two batches, the details of which have been described previously[57]. Finally, the standardized metabolite levels of the two INTERVAL subsets (batches 1 + 2 and batches 3 + 4) were further adjusted for the first ten genetic principal components, and then used for genetic-score training and external validation, respectively.

The Nightingale Health NMR platform (Nightingale Health, Helsinki, Finland) was used to assay baseline serum samples of 45,928 INTERVAL participants and quantified 230 analytes in total, which are largely lipoprotein subfractions and ratios, lipids and low-molecular-weight metabolites. This study only focused on the 141 directly measured analytes and excluded those derived from other analytes. Apart from the missing values for low-abundance analytes, the dataset also included zero values for some analytes, which were recoded as missing in our analysis. In addition, those analyte values of participants with particularly high or low values of more than 10 s.d. from the analyte mean across all participants were set as missing. We further excluded participants with more than 30% analyte missingness and duplicate samples. Participants who failed genetic QC (see above) or did not have relevant phenotype data available were also removed, which resulted in 37,359 participants remaining in the analysis. Values of each analyte were log (natural) transformed and adjusted for age, sex, recruitment centre, processing duration, month of donation, appointment time, missing appointment time (Yes or No) and the first ten genetic principal components. The residuals were then inverse-normal rank-transformed, which were finally used to perform GWAS of these traits and their genetic-score development. Details of QC and GWAS for these traits have been described previously[58].

RNA-seq was performed on the NovaSeq 6000 system (S4 flow cell, Xp workflow; Illumina) with 75-bp paired-end sequencing reads (reverse-stranded) in INTERVAL, which were aligned to the GRCh38 human reference genome (Ensembl GTF annotation v.99) using STAR (v.2.7.3.a)[59]. The gene count matrix was obtained using featureCounts (v.2.0.0)[60]. This in total resulted in raw gene-level count data of 60,676 genes (ENSEMBL gene IDs) across 4,778 individuals with 2.03 million–95.55 million uniquely mapped reads (median: around 24 million). Poor-quality samples with an RNA integrity number (RIN) of less than 4 or a read depth of fewer than 10 million uniquely mapped reads were removed. Sample swaps and cross-contamination were assessed using the match BAM to VCF (MBV) method from QTLtools (v.1.3.1)[61], which identified and corrected ten pairs of mislabelled samples; samples with no clear indication of their matching genotype data were also removed. Genes were retained on the basis of their meeting an expression threshold of more than 0.5 counts per million (CPM) in at least 1% of the samples. The filtered count values were converted to trimmed mean of M-values (TMM)-normalized transcript per million mapped reads (FPKM) values[62]. Next, the normalized $\log_2$-FPKM values for each gene were rank-based inverse-normal transformed across samples. We further excluded globin genes, rRNA genes and pseudogenes. After filtering, a total of 4,732 samples and 19,835 genes were retained for further eQTL analysis. Before eQTL mapping, the probabilistic estimation of expression residuals (PEER) method[63] was used to find and correct for latent batch effects and other unknown confounders in the gene expression data. To estimate PEER factors independent of the effects of known variables, a set of 22 covariates of interest was included in the analysis. These were age, sex, body mass index (BMI) and blood cell traits ($n = 19$), including: (1) basophil percentage (of white blood cell count); (2) eosinophil percentage (of white blood cell count); (3) lymphocyte percentage (of white blood cell count); (4) monocyte percentage (of white blood cell count); (5) neutrophil percentage (of white blood cell count); (6) white blood cell (leukocyte) count (reported); (7) immature reticulocyte fraction; (8) haematocrit (volume percentage of blood occupied by red cells); (9) reticulocyte percentage (of red blood cell and reticulocyte count); (10) haemoglobin concentration; (11) mean corpuscular haemoglobin; (12) mean corpuscular haemoglobin concentration; (13) mean corpuscular (red blood cell) volume; (14) red blood cell (erythrocyte) count (reported); (15) red blood cell distribution width; (16) mean platelet volume; (17) plateletcrit; (18) platelet distribution width; (19) platelet count. The eQTL mapping was performed on genome-wide variants using TensorQTL (v.1.0.6)[64], adjusting for age, sex, BMI, the above-mentioned blood cell traits ($N = 19$), the top ten genetic principal components, RIN, sequencing batch, RNA concentration, season (based on month of blood draw) and PEER factors ($n = 20$). The normalized gene-level values were also adjusted for the same set of covariates used in the eQTL mapping for their genetic-score training and validation. Note that we held out the last two batches of samples for external validation purposes and the first four were used for eQTL mapping, genetic-score training and internal validation.

Correlation and PCA analysis

This analysis included all of the traits qualified for genetic-score development at each platform and all of the training samples in INTERVAL. The same QC steps and covariate adjustments as were used in genetic-score development were applied before analysis. The adjusted trait levels were used to calculate Pearson's correlation *r* (using scipy v.1.5.4 in Python v.3.6.8) between traits (Supplementary Figs. 14–18) and perform PCA in each platform (Supplementary Figs. 19–23), in which the probabilistic PCA method was used to impute missing trait values and perform the PCA analysis at each platform (using pcaMethods v.1.86.0 in R v.4.1.3)[65]. We then considered traits in each platform as vertices of an undirected graph and vertices were connected via edges if traits were correlated with $r > 0.9$. Thus, subgraphs in this graph were used to identify groups of highly correlated traits in each of the platforms. In total, we identified 2,225, 299, 700, 29, 13,663 (in total 16,916 groups out of 17,227 traits) highly correlated groups of traits in SomaScan, Olink, Metabolon, Nightingale and RNA-seq, respectively (Supplementary Table 10).

External validation cohorts

The FENLAND study profiled the plasma proteins of 12,084 participants using the aptamer-based SomaScan assay (v.4), in which 8,994 participants were genotyped using the same Affymetrix UK Biobank Axiom array as INTERVAL[15]. The later subset of Fenland participants was used for the genetic-score model validation in our study. As FENLAND and INTERVAL applied two different versions of the SomaScan array (versions 3 and 4), we matched aptamers (or SOMAmers) between the two studies by using their unique SomaScan IDs, which resulted in 2,129 matched results. The detailed QC steps for protein measurements, genotype imputation and QC for genotype data in the FENLAND study were described previously[66]. The FENLAND study was approved by the National Health Service (NHS) Health Research Authority Research Ethics Committee (NRES Committee – East of England Cambridge Central, ref. 04/Q0108/19), and all participants provided written informed consent. Both the Orkney Complex Disease Study (ORCADES)[17] and the Northern Sweden Population Health Study (NSPHS)[16] have measured the plasma protein levels of their participants on the four Olink panels that were used in INTERVAL, and whole-genome-sequenced or genotyped participants (Supplementary Table 11). Thus, participants in the two studies were used to validate genetic-score models of Olink proteins considered in our study, in which gene names of proteins were used to match proteins between studies. For those proteins that appeared in two or more Olink panels, their validation measurements were averaged across panels for the protein. Detailed imputation and QC steps for protein abundance measurements and genetic data in the two studies were described in the previous studies[67,68]. Protein levels in ORCADES were adjusted for age, sex, plate, plate row, plate column, sampling year and season, top ten genetic principal components and kinship using a linear additive model. Similarly, protein levels in NSPHS were adjusted for age, age^2, sex, plate number, plate row, plate column and the first ten genetic principal components. The model residuals after adjustment in both cohorts were inverse-rank normalized before being used for validation analyses. The ORCADES study was approved by the South East Scotland Research Ethics Committee, NHS Lothian (reference: 12/SS/0151) and the NSPHS study was approved by the local ethics committee at the University of Uppsala (Regionala Etikprövningsnämnden, Uppsala, Dnr. 2005:325 with approval of extended project period on 2016-03-09). All participants gave their written informed consent in both studies.

In ORCADES, the same platform (Metabolon HD4) as INTERVAL was used to measure 1,143 blood metabolites of 1,046 participants in June 2018. Metabolite measurements were normalized by Metabolon in terms of raw area counts and rescaled to set the median equal to 1. There were 221,102 metabolite values below the limit of detection (18.5%), which were set to zero after the following QC steps. In the QC, we first removed 521 metabolite values that exceeded 10 standard deviations from their respective means (0.04%). At most, a single participant carried no more than 30 such outliers (2.6% of all metabolites), and all individuals were therefore included in the analysis. Next, we identified 94 metabolites of which fewer than 100 participants exceeded the limit of detection (8.2%). These poorly measured metabolites were excluded, leaving 1,049 metabolites measured in 1,046 individuals for analysis. Metabolite levels were adjusted for age, sex, BMI, genotyping array, season of venepuncture, year of venepuncture, sample volume available, sample volume extracted, plate, row, column and top 20 genetic principal components, where genotyping array indicates whether the individual was genotyped using the Illumina Human Hap 300v2, Illumina Omni Express or Illumina Omni 1 arrays; sample volume available is the volume of the blood sample delivered to Metabolon; sample volume extracted is the volume of the blood sample used to measure the metabolite abundance; and plate, column and row refer to the plate box number and sample well position (row and column). Model residuals were then inverse-rank normalized before being used for the validation analysis. A total of 1,007 participants had complete covariates. We used the COMP identifier in the platform to match metabolites between INTERVAL and ORCADES, which resulted in 455 overlapped metabolites.

The UKB, ORCADES and VIKING studies[18] were used as external cohorts to validate genetic scores of Nightingale traits, and trait identifiers provided in the platform were used to successfully match all 141 traits between these studies and INTERVAL. QC for these traits in UKB has been described previously in detail[69], and the levels of these traits were adjusted for sex, age, BMI, use of lipid-lowering medication, top ten genetic principal components and technical variance following the protocol of the previous study[69], and only genetically defined European participants[3] were included in the validation analyses.

In ORCADES, 2,055 participants had 249 blood metabolites measured in December 2020 using the same Nightingale NMR platform as INTERVAL. In total, 2070 samples were measured, with 15 participants having multiple measurements; for these participants, the mean value was used. We removed 22 participants who did not have any valid metabolite measurements. For the remaining 2,033 participants, the vast majority had zero missing metabolite values (1,938; 95%), and a small subset had up to 4% missing metabolite values (95; 5%). Conversely, the highest sample-missing rate per metabolite was 87 participants (4%). Each metabolite was adjusted by the following covariates in a linear model: age, sex, BMI, season of venepuncture, year of venepuncture, genotyping array and top 20 genetic principal components, where genotyping array indicates whether the individual was genotyped using Illumina Human Hap 300v2, Illumina Omni Express, or Illumina Omni 1 arrays. Model residuals were then inverse-rank normalized and used for the validation analysis. A total of 1,884 individuals had complete covariates.

In the VIKING study, 2,104 participants (no duplicates) had 249 blood metabolites measured in December 2020 using the Nightingale NMR platform. We removed 37 participants who did not have any valid metabolite measurements. For the remaining 2,067 participants, the vast majority had zero missing metabolite values (1,911; 92%), and a small subset had up to 4% missing metabolite values (156; 8%). Conversely, the highest sample-missing rate per metabolite was 150 participants (7%). Each metabolite was adjusted by the following covariates in a linear model: age, sex, BMI, season of venepuncture, year of venepuncture and the top 20 genetic principal components. Model residuals were then inverse-rank normalized and used for the validation analysis. A total of 2,046 individuals had complete covariates. Detailed descriptions on the genetic data and its QC in VIKING were provided in the previous study[18]. The study was approved by the Research Ethics Committees in Orkney, Aberdeen (North of Scotland REC), and South East Scotland REC, NHS Lothian (reference: 12/SS/0151). All participants gave written informed consent.

The Multi-Ethnic Cohort (MEC) recruited three major Asian ethnic groups represented in Singapore—Chinese, Malay and Indian individuals— between 2004 and 2010 to better understand how genes and lifestyle influence health and diseases differently in people of different ethnicities[19]. Between 2011 and 2016, the participants were invited for a follow-up. Analyses on the MEC study were approved by the National University Institutional Review Board (NUS-IRB: LN-18-059 and NUS-IRB-2021-812) and Singapore Population Health Studies Scientific Committee. Whole-genome sequencing was performed on 2,902 MEC participants as Phase I of the Singapore National Precision Medicine Programme (https://npm.a-star.edu.sg/)[70]. Samples were whole-genome-sequenced to an average of 15× coverage. Read alignment was performed with BWA-MEM v.0.7.17 and variant discovery and genotyping were performed with GATK v.4.0.6.0. Site-level filtering includes only retaining VQSR-PASS and non-STAR allele variants. At the sample level, samples with a call rate < 95%, BAM cross-contamination rate > 2% or BAM error rate > 1.5%, and at the genotype call level, genotypes with DP (the filtered depth at the sample level) < 5 or genotype quality (GQ) < 20 or allele balance (AB) > 0.8 (heterozygote calls), were set to NULL. Finally, samples with abnormal ploidy were excluded. To determine the genetic ancestry of samples, we first performed the PCA on the variant panel of verifyBamID2 (ref. 71) (1000G, phase 3), and the obtained top 15 genetic principal components and their associated explained variance were used to perform *k*-means clustering ($k = 3$). An ancestry label (Chinese, Malay or Indian) was then assigned to each sample on the basis of the major self-reported ethnicity of each cluster. Both the SomaScan (v.4) and the Nightingale NMR platform were used to assay the baseline and revisit blood samples of participants in MEC. For QC of Nightingale data, participants with more than 10% missing metabolic biomarker values were excluded from subsequent analyses. For participants with biomarker values lower than the detection level, we replaced values of 0 with a value equivalent to 0.9 multiplied by the non-zero minimum value of that measurement. For QC of SomaScan data, protein levels were first normalized to remove hybridization variation within a run. This was followed by median normalization across calibrator control samples to remove other assay biases within the run. Overall scaling and calibration were then performed on a per-plate basis to remove overall intensity differences between runs with calibrator controls. Finally, median normalization to a reference was performed on the individual samples with QC controls. During these standardization steps, multiple scaling factors were generated for each sample or aptamer at each step. The final numbers of samples in each ethnic group used in our validation are shown in Extended Data Table 1. For both SomaScan and Nightingale traits, natural log-transformation was applied before adjusting for age, sex, type 2 diabetes status and BMI (Nightingale traits only) and the first ten genetic principal components. Residuals from the regression were inverse-normalized for correlation analyses with genetic scores trained in INTERVAL.

The Jackson Heart Study (JHS) is a community-based longitudinal cohort study of 5,306 self-identified Black individuals from the Jackson, Mississippi metropolitan statistical area that began in 2000 (refs. 20,72). The participants included in our validation of genetic scores for SomaScan proteins are samples collected at Visit 1 between 2000 and 2004 from 1,852 individuals with whole-genome sequencing and proteomic profiling (SomaScan) performed, QCs of which were detailed in the previous studies[20,73,74]. SomaScan IDs were used to match shared proteins between JHS and INTERVAL, which identified 820 proteins in total. Protein levels were adjusted for age, sex and the first ten principal components of genetic ancestry in JHS, before they were used for evaluating the performance of genetic scores. This study was approved by the JHS Publications and Presentations Subcommittee and the TOPMed Multi-Omics Working group.

In summary, we performed QC in each external cohort to ensure the quality of the omic data used for validation, and adjusted trait levels for covariates to minimize potential validation bias across cohorts, which include age, sex, genetic princiipal components and other cohort- or platform-specific environmental and technical factors (Supplementary Table 11). Note that using Nightingale traits in ORCADES as examples, we found that controlling for family structure (adjustment for kinship) had a very minor effect on the validation results (Supplementary Fig. 24); thus, we did not consider control for this factor essential in the external validation.

Polygenic scoring method

A polygenic or genetic score is most commonly constructed as a weighted sum of genetic variants carried by an individual, in which the genetic variants are selected and their weights quantified by univariate analysis in a corresponding GWAS[75,76]:

$$\hat{\mathrm{PGS}}_{\mathrm{i}} = \sum_{j\in S} \beta_j \times x_{ij}, \tag{1}$$

where PGS is the polygenic score; S is the set of variants, referring to single-nucleotide polymorphisms (SNPs) in this study, that are identified in the variant-selection step described below; β_j *is* the effect size of the SNP *j* that is obtained through the univariate statistical association tests in the GWAS; and x_{ij} is the genotype dosage of SNP *j* of the individual *i*. As the variant set *S* is derived through a LD thinning and *P*-value thresholding process, this method is often named P+T. However, P+T relies on hard cut-off thresholds to remove LD correlations among variants and select associated variants. It is often challenging to balance between keeping predictive variants and removing redundant and uninformative variants that can limit the prediction precision. Also, owing to the inherent linear assumption of the univariate analysis in P+T, this method leaves no modelling considerations for joint effects between variants. To alleviate these limitations, various machine learning-based methods, such as Bayesian ridge (BR), elastic net (EN)[77] and LDpred[78], have been used to construct genetic scores for a wide range of traits and diseases[7]. In particular, BR and EN have been shown to outperform other methods when developing scores for predicting biomolecular traits, such as blood cell traits and gene expression[7,9], which are similar to the types of traits considered in this study. We adopted the BR method for the genetic-score construction of all the biomolecular traits, as BR is more efficient to run in practice (see details below).

BR is a multivariate linear model that assumes that the genetic variants have linear additive effects on the genetic score of the trait[7,79]. In addition, BR also assumes that the genetic score of a trait follows a Gaussian distribution, and the prior for effect sizes of variants is also given by a spherical Gaussian:

$$p(\mathrm{PGS}|\boldsymbol{x}, \boldsymbol{\beta}, \alpha) \approx N(\mathrm{PGS}| \sum_{j\in S} x_j\beta_j, \alpha^{-1}) \tag{2}$$

$$p(\boldsymbol{\beta}|\lambda) \approx N(\boldsymbol{\beta}|0, \lambda^{-1}) \tag{3}$$

where *S* is the set of input variants and α and λ are coefficients of the model and subject to two gamma distributions: Gamma(α_1, α_2) and Gamma(λ_1, λ_2). These two prior gamma distributions can be set via a validation step.

Genetic-score training and evaluation

The explained variance (R^2) and Spearman's rank correlation coefficient (Rho) were used to measure the performance of constructed genetic scores in the INTERVAL training samples and external cohorts (or INTERVAL withheld subset), in which R^2 scores were calculated using the squared Pearson correlation coefficient (*r*). The Python (v.3.6.8) package scipy v.1.5.4 was used to derive Rho and *r* scores, and statistical significance was calculated using a two-sided t-test for *r* and a two-sided Mann-Whitney U test for Rho. We adopted a similar strategy for sample partition when training and evaluating genetic scores within the training samples to that in previous studies[7,9] that used learning-based methods to construct genetic scores for molecular traits. The training samples of a trait were randomly and equally partitioned to five subsets, from which any four subsets were used as true-training data to learn a genetic-score model of the trait, and tested the model's performance on the remaining 20% of samples (Extended Data Fig. 1). Given a genetic-scoring method and a trait, we obtained five different genetic-score models of the trait, and the mean of their performance measurements in the corresponding testing samples in INTERVAL was reported (internal validation). Note that, owing to the high similarities between the five genetic-score models trained for most traits, only one model was randomly selected from the five and evaluated in the external cohorts (or INTERVAL withheld set for Metabolon and RNA-seq).

When training genetic-score models using the BR method, we need to select two appropriate prior gamma distributions; that is, α_1, α_2, λ_1 and λ_2. To do so, a grid search across a set of optional hyperparameters is often performed; however, this searching process is resource- and time-intensive, which makes it challenging to run for tens of thousands of multi-omic traits. To address this problem, we randomly selected subsets of SomaScan, Olink and Metabolon traits (20 each), on which we trained and internally validated genetic scores on any α_1, α_2, λ_1 and λ_2 taken from $\{0, 10^{-10}, 10^{-5}, 10^{-3}, 10^{-1}, 10, 10^3 10^5, 10^{10}\}$. The results suggested that using non-informative priors[80] (α_1, α_2, λ_1 and $\lambda_2 \in \{0, 10^{-10}, 10^{-5}, 10^{-3}\}$)

performed as well as that using the best-performing hyperparameters selected through extensive searching (Supplementary Figs. 25–28). We further externally validated the performance of genetic scores developed using non-informative priors (α_1, α_2, λ_1 and $\lambda_2 = 10^{-5}$) and the best-performing priors selected in internal validation for each of the 20 Metabolon traits on the INTERVAL withheld set, which showed that they had a nearly identical R^2 performance (Supplementary Fig. 28b). Therefore, we adopted the non-informative priors (α_1, α_2, λ_1 and $\lambda_2 = 10^{-5}$) in BR for genetic-score development of all other traits. We further note that our approach minimizes the risk of collider bias; for example, by using BR to re-estimate the weights for all genetic variants that pass univariate genome-wide significance, and then performing external validation using only minimal covariates (sex, age and genetic principal components).

Variant selection and comparison of methods

Selecting a proper set of variants and feeding into a polygenic scoring method are a key step for effective genetic-score construction. To do so and further confirm the superiority of the BR method, we looked at the performance of BR and P+T on a variety of variant-selection schemes for the traits in three platforms (SomaScan, Olink and Metabolon). The Python (v.3.6.8) packages scikit-learn v.0.21.2, pandas v.1.1.5 and numpy v.1.19.5 were used to implement BR for genetic-score training.

To ensure the generalizability of genetic-score models when applied to other cohorts, a variant-filtering step was first performed for all the traits considered, which applied an MAF threshold of 0.5% and excluded all multi-allelic variants as well as ambiguous variants (that is, A/T or G/C) in INTERVAL. To remove LD dependencies among variants, a follow-up LD thinning step was carried out at an r^2 threshold of 0.8 on all the variants for both BR and P+T methods using indep-pairwise in Plink v.2.0 (ref. 51). The remaining variants were then filtered at given *P*-value thresholds (from their GWAS summary statistics conducted on the INTERVAL training data) for a trait in different platforms as inputs of BR and P+T. To identify an appropriate variant-selection scheme for the use of all the biomolecular traits, we attempted the following four *P*-value thresholding schemes for protein traits in the Olink and SomaScan platforms: (1) $P < 5 \times 10^{-8}$ on all the variants; (2) $P < 5 \times 10^{-8}$ on variants in the *cis* region only (within 1 Mb of the corresponding gene's transcription start site); (3) all the *cis* variants only; (4) all the *cis* variants and $P < 1 \times 10^{-3}$ on the *trans* variants; and the two different *P*-value thresholds on the genome-wide variants for metabolite traits in the Metabolon platform (as they do not distinguish *cis* and *trans* regions): (1) $P < 5 \times 10^{-8}$; (2) $P < 1 \times 10^{-3}$.

Then, we compared the performance of BR and P+T on these variant sets in the internal validation (Supplementary Figs. 1–3). With regard to the proteomic traits (SomaScan and Olink), the two variant-selection schemes: (1) $P < 5 \times 10^{-8}$ on genome-wide variants and (2) all the *cis* variants and $P < 1 \times 10^{-3}$ on the *trans* variants, were shown to be the best-performing schemes with either of the methods; the BR method largely outperformed P+T across the two variant-selection schemes. Meanwhile, it was noted that the two selection schemes led to greatly different performances, with the latter scheme achieving an unrealistic mean R^2 of around 0.74 across all the proteins (only around 0.09 for the former scheme). Similarly, for the metabolomic traits (Metabolon), the applied two variant-selection schemes significantly differed in their performance in internal validation, and BR was also shown to be a better-performing method.

To further identify the optimal variant-selection scheme for BR, we also looked at the performance of genetic-score models trained with the two identified (for proteins) or all the two applied (for metabolites) schemes using the BR method for Olink traits and Metabolon traits (Fig. 2 and Supplementary Fig. 4) in external cohorts (NSPHS and ORCADES) or withheld INTERVAL data. Despite the second scheme (all the *cis* variants and $P < 1 \times 10^{-3}$ on the *trans* variants for proteins, or $P < 1 \times 10^{-3}$ on genome-wide variants for metabolites) showing an outstanding performance in internal validation, its performance saw a marked decline in external validation for almost every trait validated (Supplementary Fig. 4). This indicates that this variant-selection scheme caused an overfitting problem in genetic-score training, which is consistent with previous findings when using overly lenient *P*-value thresholds for variant selection[7].

The performance of BR (variant set with a *P*-value threshold of 5×10^{-8}) was further benchmarked alongside P+T (*P*-value thresholds of 5×10^{-8} and 1×10^{-3}) and LDpred2 (ref. 81) for a random subset of 20 Metabolon traits in the INTERVAL withheld set. We used the LDpred2-auto model to train genetic scores, where the R (v.3.6.1) package bigsnpr v.1.10.8 was used to implement LDpred2-auto, and summary statistics from GWAS in the training samples and the recommended Hapmap3 variant set were used as model inputs. All of the INTERVAL samples, excluding those withheld for independent validation, were used to obtain the variant–variant correlation matrix for LDpred2. Our results showed that BR outperformed P+T. Although LDpred2 showed a similar R^2 to BR for most traits, some were substantially attenuated in the withheld set (Extended Data Fig. 2). In addition, our benchmark results showed that BR, P+T and LDpred2-auto had an average running time of 3.1 seconds (2 CPU cores), 2.9 s (2 CPU cores) and 51 min (20 CPU cores) per trait respectively on the Cambridge Service for Data-Driven Discovery platform (https://www.hpc.cam.ac.uk/), showing that BR performed well in both performance and scalability.

These results suggested that the BR method with the variant-selection scheme of $P < 5 \times 10^{-8}$ on genome-wide variants was the optional method (of those tested) for genetic-score development of these biomolecular traits; thus, we applied this approach to all other traits for their genetic-score development in this study. We noted that the optimal variant set had been selected using a much larger *P*-value threshold in the previous study[7], which could be due to there being an order of magnitude difference in training sample size and greater polygenicity of the traits as compared to the current study.

Longitudinal stability of genetic scores

Within the MEC, 1,739 individuals were measured at both baseline and revisit with a mean length of follow-up of 6.31 years (s.d. 1.45 years). This allowed the longitudinal assessment of the stability of genetic scores for SomaScan ($n = 403$ Chinese, 356 Indian and 353 Malay) and Nightingale ($n = 721$ Chinese, 376 Indian and 363 Malay) platforms. For SomaScan traits, we found a strong consistency between the predictive capacity of genetic scores between baseline and revisit samples (Pearson's $r = 0.99$ for Chinese, 0.98 for Indian and 0.98 for Malay populations), and little difference in longitudinal stability between ancestries (Extended Data Fig. 7d–f). For Nightingale traits, despite variation in the predictive capacity of genetic scores between baseline and revisit samples, the longitudinal stability was still largely consistent between Indian and Malay ancestries (Pearson's $r = 0.60$ for Chinese, 0.84 for Indian and 0.85 for Malay populations; Extended Data Fig. 7a–c).

Cross-platform performance of genetic scores

SomaScan and Olink used two different technologies for measuring protein levels. The two platforms measured many proteins in common, among which there are 169 unique proteins whose genetic scores we have validated. To check the effect of technologies on genetic prediction, we looked at how the genetic scores trained on one platform can predict protein levels from the other platform on the INTERVAL training samples (Supplementary Fig. 29). We confirmed that the performance of these overlapped genetic scores trained on the other platform was generally consistent with that of the scores trained on their original platform. However, we observed, in some cases, that the genetic scores trained on the two platforms could lead to very different predictions, which we found to be due mainly to the differences in what the two platforms were actually quantifying. For example, among the 169 proteins, there were 11 proteins in SomaScan that had an $R^2 > 0.3$ in internal validation, in which 10 proteins also achieved an $R^2 > 0.3$ but the remaining protein (CHI3L1) received a poor $R^2 < 0.1$ when predicting with Olink genetic scores. We found that the remaining protein received the lowest Pearson's *r* score among the 11 proteins between their actual protein levels measured in the two platforms. In INTERVAL, there were around 700 participants (depending on the protein) who were assayed by both SomaScan and Olink, which allowed us to calculate the correlations between the actual protein levels measured by the two platforms for the same protein. These results suggested—despite great consistency—that genetic scores of the same protein trained on two platforms can represent distinct aspects of protein biology, and that integrating diverse proteomic techniques might enable better genetic scores to be developed for these proteins[82]. Similarly, we also investigated the predictive performance of our Nightingale genetic scores on the biochemistry assay data in UKB for overlapping biomarkers. We found that the performances of these INTERVAL-trained genetic scores were largely robust with respect to measurement technology (Supplementary Fig. 30).

Pathway coverage analysis of proteins

In this analysis, SomaScan and Olink proteins were combined on the basis of their Uniprot ID, and duplicate proteins were removed if identified. We only kept proteins with $R^2 > 0.01$ in internal validation, resulting in a total of 2,205 unique proteins for the analysis. We used pathway data for *Homo sapiens* curated at Reactome[21] and conducted analyses to uncover the coverage of these proteins in the pathways. In detail, this analysis looked at the percentages of these proteins in annotated physical entities of each super-pathway, and the percentages of the lowest-level pathways these proteins covered among all the lowest-level pathways of each super-pathway. In cases in which at least one protein in this study was included in the entities of a lowest-level pathway, we considered this pathway to be covered by proteins of this study.

Phenome-wide association analysis in UKB

We included biomolecular traits with $R^2 > 0.01$ in internal validation in this analysis (11,576 traits in total) and considered only participants of European ancestry in UKB (the white British subset). We used version 3 of the imputed and quality controlled genotype data for UKB, which were detailed previously[3]. Using version 1.2 of the PheWAS Catalog[22], we extracted the curated phenotype definitions of all phecodes. Each phecode is provided as a set of WHO International Classification of Diseases (ICD) diagnosis codes in versions 9 (ICD-9) and 10 (ICD-10) of the ontology to define individuals with the phenotype of interest, and a set of related phecodes that should be excluded from the control cohort of unaffected individuals. To define cases for each phecode, we searched for the presence of any of the constituent ICD-9 or ICD-10 codes in linked health records (including in-patient Hospital Episode Statistics data, cases of invasive cancer defined in the cancer registry and primary and secondary cause of death information from the death registry), and converted the earliest coded date to the age of phenotype onset. Individuals without any codes for the phenotype of interest were recorded as controls, and censored according to the maximum follow-up of the health linkage data (31 January 2020) or the date of death, whichever came first. To define the cohort for testing molecular genetic-score associations with the age of onset of each phenotype, we used the set of events and censored individuals described above and removed any individuals with related phenotypes (according to definitions from the PheWAS Catalog), restricting analyses to be sex-specific (for example, ovarian and prostate cancer) where required. To ensure a well-powered study, we restricted the PheWAS analysis to phenotypes with at least 200 cases in the 409,703 individuals of European ancestry whose reported sex matched the genetically inferred sex from the UKB quality controlled genotype data[3], resulting in a set of 1,123 phecodes included in the final analysis. The association of the genetic score for biomolecular traits with the onset of each phenotype was assessed by using a Cox proportional hazards model with age as timescale, stratified by sex and adjusted for genotyping array and ten principal components of genetic ancestry. The association between genetic scores and each phecode is reported in terms of its effect size (Hazard ratio) and corresponding significance (*P* value), and significant results were defined as Benjamini–Hochberg FDR-corrected $P < 0.05$ for all the tested traits (two-sided Wald test). Statistical analyses were performed in Python (v.3.6.8) and the Cox model was implemented using the lifelines v.0.26.0 package[83].

Carbon impact and offsetting

We used GreenAlgorithms v.1.0 (ref. 84) to estimate that the main computational work in this study had a carbon impact of at least 1,004 kg of CO_2 emissions (CO_2e), corresponding to 94 tree-years. As a commitment to the reduction of carbon emissions associated with computation in research, we consequently funded the planting of 45 trees through a local Australian charity, which across their lifetime will sequester a combined estimated 12,000 kg of CO_2e, or 12 times the amount of CO_2e generated by this study.

Reporting summary

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

Data availability

All of the genetic-score models trained in this study and GWAS summary statistics used to develop genetic scores are publicly accessible through the OmicsPred portal (https://www.omicspred.org/) under accession codes OPGS000001–OPGS017227. INTERVAL study data from this paper are available to bona fide researchers from helpdesk@intervalstudy.org.uk and information, including the data access policy, is available at http://www.donorhealth-btru.nihr.ac.uk/project/bioresource.

Code availability

The original codes used to train the genetic scores with INTERVAL data, internally validate these scores and benchmark the performance of different genetic-score construction methods are available at https://github.com/xuyu-cam/atlas_genetic_scores_omic_traits.

51. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
52. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
53. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 (2016).
54. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
55. Lundberg, M., Eriksson, A., Tran, B., Assarsson, E. & Fredriksson, S. Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood. *Nucleic Acids Res.* **39**, e102 (2011).
56. Folkersen, L. et al. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.* **2**, 1135–1148 (2020).
57. Surendran, P. et al. Rare and common genetic determinants of metabolic individuality and their effects on human health. *Nat. Med.* **28**, 2321–2332 (2022).
58. Karjalainen, M. K. et al. Genome-wide characterization of circulating metabolic biomarkers reveals substantial pleiotropy and novel disease pathways. Preprint at *medRxiv* https://doi.org/10.1101/2022.10.20.22281089 (2022).
59. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
60. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
61. Fort, A. et al. MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets. *Bioinformatics* **33**, 1895–1897 (2017).
62. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
63. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
64. Taylor-Weiner, A. et al. Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).

65. Stacklies, W., Redestig, H., Scholz, M., Walther, D. & Selbig, J. pcaMethods—a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* **23**, 1164–1167 (2007).
66. Pietzner, M. et al. Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nat. Commun.* **11**, 6397 (2020).
67. Bretherick, A. D. et al. Linking protein to phenotype with Mendelian randomization detects 38 proteins with causal roles in human diseases and traits. *PLoS Genet.* **16**, e1008785 (2020).
68. Kierczak, M. et al. Contribution of rare whole-genome sequencing variants to plasma protein levels and the missing heritability. *Nat. Commun.* **13**, 2532 (2022).
69. Ritchie, S. C. et al. Quality control and removal of technical variation of NMR metabolic biomarker data in ~120,000 UK Biobank participants. *Sci. Data* **10**, 64 (2023).
70. Wong, E. et al. The Singapore National Precision Medicine strategy. *Nat. Genet.* **55**, 178–186 (2023).
71. Zhang, F. et al. Ancestry-agnostic estimation of DNA sample contamination from sequence reads. *Genome Res.* **30**, 185–194 (2020).
72. Taylor, H. A. J. et al. Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. *Ethn. Dis.* **15**, S6-4-17 (2005).
73. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
74. Ngo, D. et al. Aptamer-based proteomic profiling reveals novel candidate biomarkers and pathways in cardiovascular disease. *Circulation* **134**, 270–285 (2016).
75. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
76. Chatterjee, N., Shi, J. & Garcia-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).
77. Okser, S. et al. Regularized machine learning in the genetic prediction of complex traits. *PLoS Genet.* **10**, e1004754 (2014).
78. Vilhjálmsson, B. J. et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
79. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
80. Tipping, M. E. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **1**, 211–244 (2001).
81. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2021).
82. Pietzner, M. et al. Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nat. Commun.* **12**, 6822 (2021).
83. Davidson-Pilon, C. lifelines: survival analysis in Python. *J. Open Source Softw.* **4**, 1317 (2019).
84. Lannelongue, L., Grealey, J. & Inouye, M. Green algorithms: quantifying the carbon footprint of computation. *Adv. Sci.* **8**, 2100707 (2021).
85. Di Angelantonio, E. et al. Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. *Lancet* **390**, 2360–2371 (2017).

Acknowledgements Participants in the INTERVAL randomized controlled trial were recruited with the active collaboration of NHS Blood and Transplant England (https://www.nhsbt.nhs.uk/), which has supported field work and other elements of the trial. DNA extraction and genotyping were co-funded by the National Institute for Health and Care Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk) and the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). The academic coordinating centre for INTERVAL was supported by core funding from the: NIHR Blood and Transplant Research Unit (BTRU) in Donor Health and Genomics (NIHR BTRU-2014-10024), NIHR BTRU in Donor Health and Behaviour (NIHR203337), UK Medical Research Council (MR/L003120/1), British Heart Foundation (SP/09/002; RG/13/13/30194; RG/18/13/33946) and NIHR Cambridge BRC (BRC-1215-20014; NIHR203312). A complete list of the investigators and contributors to the INTERVAL trial is provided in a previous study[85]. The academic coordinating centre would like to thank blood donor centre staff and blood donors for participating in the INTERVAL trial. RNA-seq was funded as part of an alliance between the University of Cambridge and the AstraZeneca Centre for Genomics Research (AZ ref: 10033507) and by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). INTERVAL SomaLogic assays were funded by Merck and the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). INTERVAL Olink Proteomics assays (Neurology panel) were funded by Biogen. INTERVAL Metabolon assays were funded by the NIHR BioResource, Wellcome Trust grant number 206194, BioMarin Pharmaceutical and the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). INTERVAL Nightingale Health NMR assays were funded by the European Commission Framework Programme 7 (HEALTH-F2-2012-279233). UKB data access was approved under projects 7439, 11193 and 19655, and all the participants gave their informed consent for health research. The MEC is funded by individual research and clinical scientist award schemes from the Singapore National Medical Research Council (NMRC, including MOH-000271-00) and the Singapore Biomedical Research Council (BMRC), the Singapore Ministry of Health (MOH), the National University of Singapore (NUS) and the Singapore National University Health System (NUHS). This work on omics polygenic score transferability is supported by the NUS–Cambridge Seed Grant July 20201 (NUSMEDIR/Cambridge/2021-07/001). The metabolite biomarkers data were generated in collaboration with Nightingale Health. The protein biomarker data were generated in collaboration with Somalogic. The MEC whole-genome sequencing data made use of data generated as part of the Singapore National Precision Medicine (NPM) program funded by the Industry Alignment Fund (Pre-Positioning) (IAF-PP: H17/01/a0/007). NPM made use of data and samples collected in the following cohorts in Singapore: (1) the Health for Life in Singapore (HELIOS) study at the Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore (supported by grants from a Strategic Initiative at Lee Kong Chian School of Medicine, the Singapore Ministry of Health (MOH) under its Singapore Translational Research Investigator Award (NMRC/STaR/0028/2017) and the IAF-PP: H18/01/a0/016); (2) the Growing up in Singapore Towards Healthy Outcomes (GUSTO) study, which is jointly hosted by the National University Hospital (NUH), KK Women's and Children's Hospital (KKH), the National University of Singapore (NUS) and the Singapore Institute for Clinical Sciences (SICS), Agency for Science Technology and Research (A*STAR) (supported by the Singapore National Research Foundation under its Translational and Clinical Research (TCR) Flagship Programme and administered by the Singapore Ministry of Health's National Medical Research Council (NMRC), Singapore-NMRC/TCR/004-NUS/2008; NMRC/TCR/012-NUHS/2014. Additional funding is provided by SICS and IAF-PP H17/01/a0/005); (3) the Singapore Epidemiology of Eye Diseases (SEED) cohort at Singapore Eye Research Institute (SERI) (supported by NMRC/CIRG/1417/2015; NMRC/CIRG/1488/2018 and NMRC/OFLCG/004/2018); (4) the MEC cohort (supported by NMRC grant 0838/2004, BMRC grants 03/1/27/18/216, 05/1/21/19/425 and 11/1/21/19/678, Ministry of Health, Singapore, National University of Singapore and National University Health System, Singapore); (5) the SingHealth Duke–NUS Institute of Precision Medicine (PRISM) cohort (supported by NMRC/CG/M006/2017_NHCS, NMRC/StaR/0011/2012, NMRC/StaR/ 0026/2015, Lee Foundation and Tanoto Foundation); (6) the TTSH Personalised Medicine Normal Controls (TTSH) cohort (supported by NMRC/CG12AUG17 and CGAug16M012). The views expressed are those of the author(s) and not necessarily those of the National Precision Medicine investigators, or institutional partners. We are grateful to all Fenland volunteers and to the general practitioners and practice staff for assistance with recruitment. We thank the Fenland Study Investigators, Fenland Study Co-ordination team and the Epidemiology Field, Data and Laboratory teams. Proteomic measurements were supported and governed by a collaboration agreement between the University of Cambridge and SomaLogic. The Fenland Study (10.22025/2017.10.101.00001) is funded by the Medical Research Council (MC_UU_12015/1). We further acknowledge support for genomics from the Medical Research Council (MC_PC_13046). ORCADES was supported by the Chief Scientist Office of the Scottish Government (CZB/4/276 and CZB/4/710), a Royal Society URF to J.F.W., the MRC Human Genetics Unit quinquennial programme 'QTL in Health and Disease', Arthritis Research UK and the European Union framework program 6 EUROSPAN project (contract no. LSHG-CT-2006-018947). DNA extractions were performed at the Edinburgh Clinical Research Facility, University of Edinburgh. We would like to acknowledge the contributions of the research nurses in Orkney, the administrative team in Edinburgh and the people of Orkney. The Viking Health Study Shetland (VIKING) was supported by the MRC Human Genetics Unit quinquennial programme grant 'QTL in Health and Disease'. DNA extractions and genotyping were performed at the Edinburgh Clinical Research Facility, University of Edinburgh. We would like to acknowledge the contributions of the research nurses in Shetland, the administrative team in Edinburgh and the people of Shetland. We acknowledge support from the MRC Human Genetics Unit programme grant, 'Quantitative traits in health and disease' (U. MC_UU_00007/10). Whole-genome sequencing for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Whole-genome sequencing for 'NHLBI TOPMed: Jackson Heart Study' (phs000964) was performed at the Northwest Genomics Center (HHSN268201100037C). 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 were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We acknowledge the studies and participants who provided biological samples and data for TOPMed. 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). We also thank the staff and participants of the JHS. JHS disclaimer: 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 US Department of Health and Human Services. This work was also funded by the Swedish Research Council (2019-01497) and the Swedish Heart-Lung foundation (20200687). Y.X. and M.I. were supported by the UK Economic and Social Research Council (ES/T013192/1). S.C.R., L.L., C.F. and E.P. are funded by a BHF Programme Grant (RG/18/13/33946). C.L., M.P. and J.L. were funded by the Medical Research Council (MC_UU_00006/1 – Aetiology and Mechanisms). S.A.L. was supported by a Canadian Institutes of Health Research postdoctoral fellowship (MFE-171279). U.A.T. is supported by a US National Institutes of Health Mentored Clinical Scientist Development Award program (1K08HL161445-01A1). C.F. was supported by the Health Data Research UK. E.P. was funded by the EU/EFPIA Innovative Medicines Initiative Joint Undertaking BigData@Heart grant 116074, NIHR BTRU in Donor Health and Genomics (NIHR BTRU-2014-10024) and NIHR BTRU in Donor Health and Behaviour (NIHR203337). J.E.P. was supported by a Medical Research Foundation grant (MRF-042-0001-RG-PETE-C0839). E.E.D. is supported by a Wellcome Trust grant (206194, 220540/Z/20/A). R.E.G. is supported by a US National Institutes of Health grant for proteomics in the Jackson Heart Study (R01 HL133870) and an NIH contract to perform proteomics and metabolomics in multiple cohorts (HHSN268201600034I). J.D. holds a British Heart Foundation Professorship and a NIHR Senior Investigator Award. M.I. is supported by the Munz Chair of Cardiovascular Prediction and Prevention and the NIHR Cambridge Biomedical Research Centre (NIHR203312). This study was supported by the Victorian Government's Operational Infrastructure Support (OIS) program. This work was supported by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. This research was supported by an HDRUK Director's Innovation Award (HDRUK2022.0130). We acknowledge B. Sun and T. Jiang for previous analyses of INTERVAL SomaScan and genotype QC, respectively. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any author accepted manuscript version arising from this submission. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, NHSBT or the Department of Health and Social Care.

Author contributions Y.X. and M.I. conceived and designed the study. Y.X. and S.C.R. performed the genetic-score training and internal validation analyses. Y.X., Y.L., P.R.H.J.T., M.P., U.A.T., S.M.-W., Å.J., P.S. and S.D. performed the external validation analyses. S.C.R., A.P.N., E.P., J.E.P., C.O.-W. and B.P. performed the data QC and GWAS in INTERVAL. Y.X. and L.L. performed the methods benchmarking analyses. Y.X. and S.A.L. performed the PheWAS. Y.X. performed the pathway coverage, correlation and PCA analyses. Y.X. and S.C.R. performed the cross-platform validation analyses. S.C.R. performed the genetic-score polygenicity analysis. C.F. and M.I. interpreted the biological insights. Y.X. developed and maintained the online portal. M.I., A.S.B., J.F.W., C.L., X.S., J.D., R.E.G., D.S.P., E.E.D., R.M.v.D., E.S.T., E.D.A., N.S., L.B. and J.L. acquired the resources and datasets; M.I., A.S.B., J.F.W., C.L., X.S., J.D., A.M., R.E.G., C.Y., D.S.P., E.E.D., H.P. and N.P. supervised the work. Y.X. and M.I. wrote the original manuscript. All authors reviewed and approved the final paper.

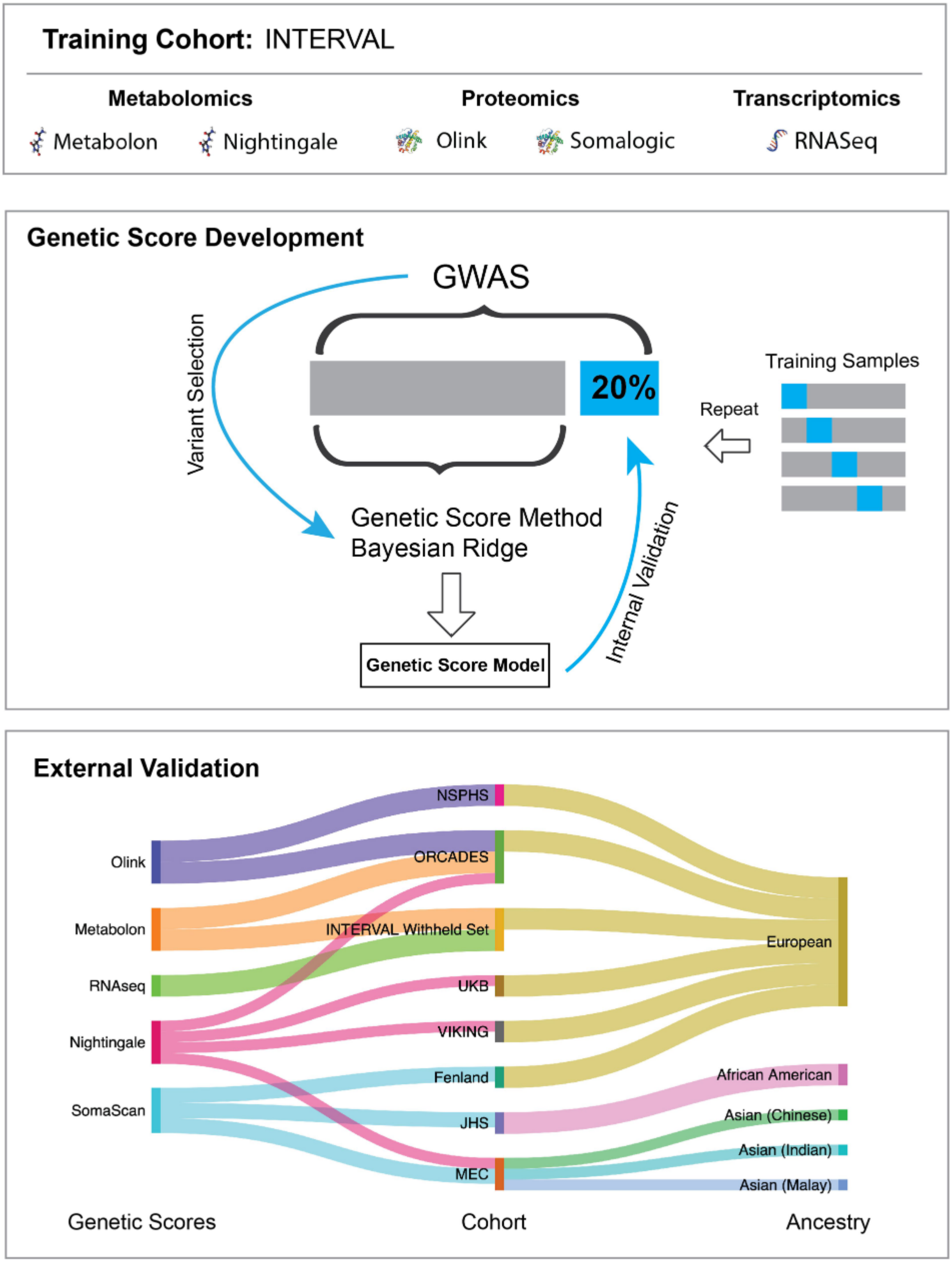
Competing interests During the drafting of the manuscript, P.R.H.J.T. became a part-time employee of BioAge Labs, P.S. became a full-time employee of GSK and D.S.P. became a full-time employee of AstraZeneca. L.B. is an employee of BioMarin. J.D. serves on scientific advisory boards for AstraZeneca, Novartis and UK Biobank, and has received multiple grants from academic, charitable and industry sources outside of the submitted work. A.M. is an employee of Pfizer. A.S.B. reports institutional grants from AstraZeneca, Bayer, Biogen, BioMarin, Bioverativ, Novartis, Regeneron and Sanofi.

Additional information

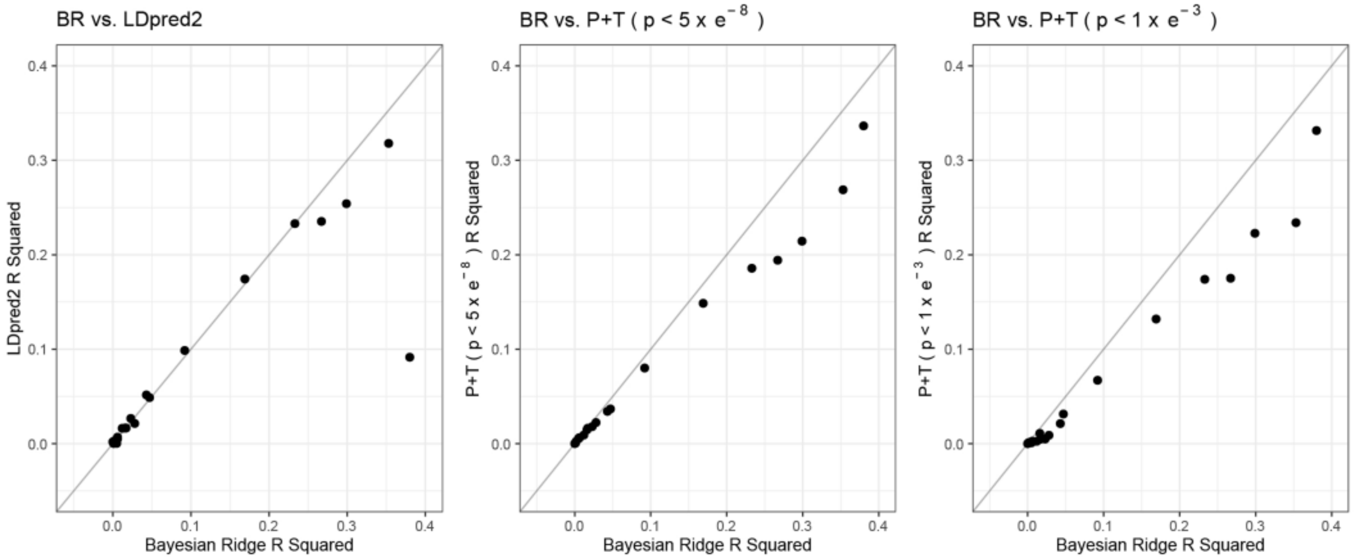
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05844-9.

Correspondence and requests for materials should be addressed to Yu Xu or Michael Inouye.

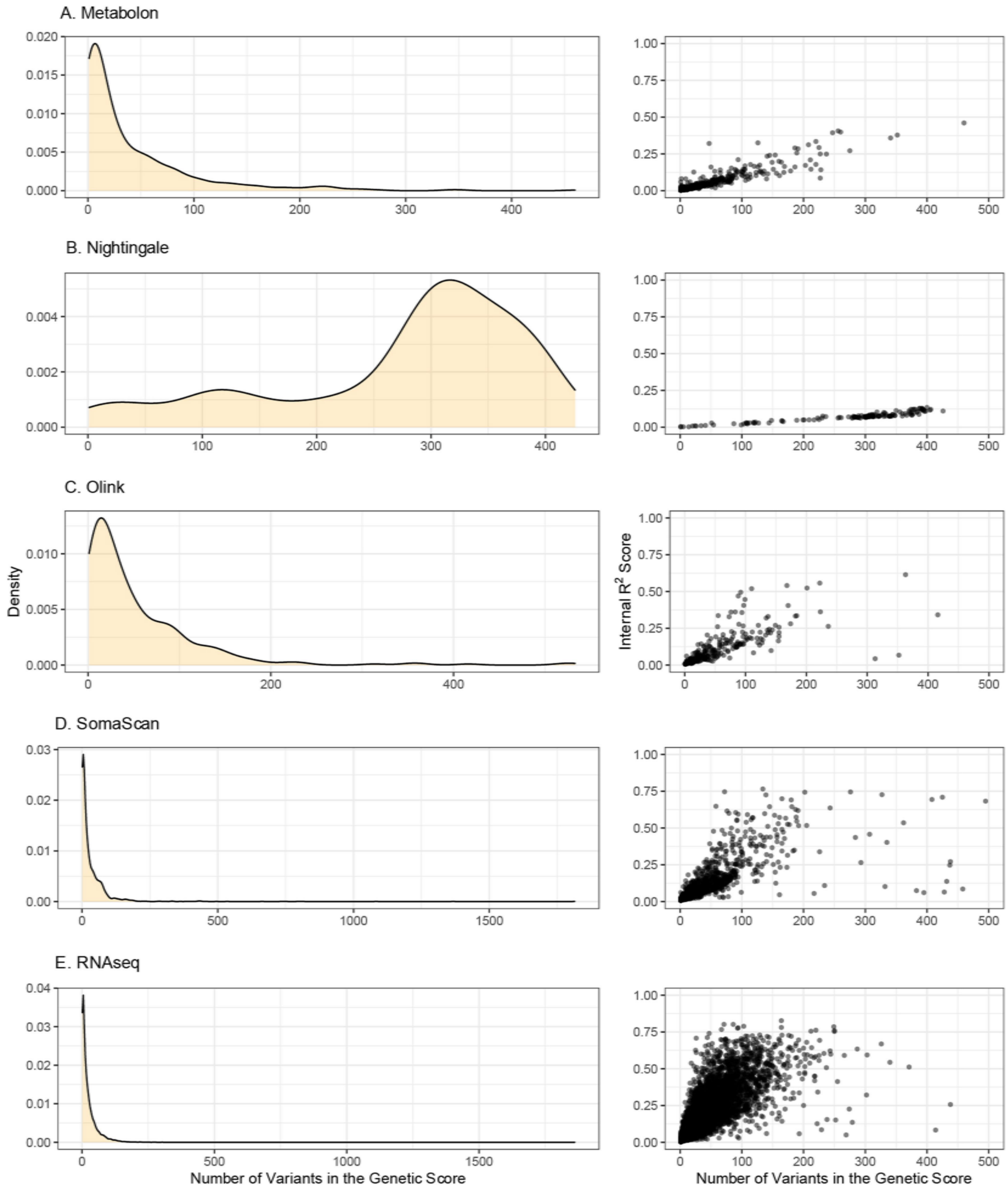
Peer review information *Nature* thanks Heiko Runz, Bjarni Vilhjálmsson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



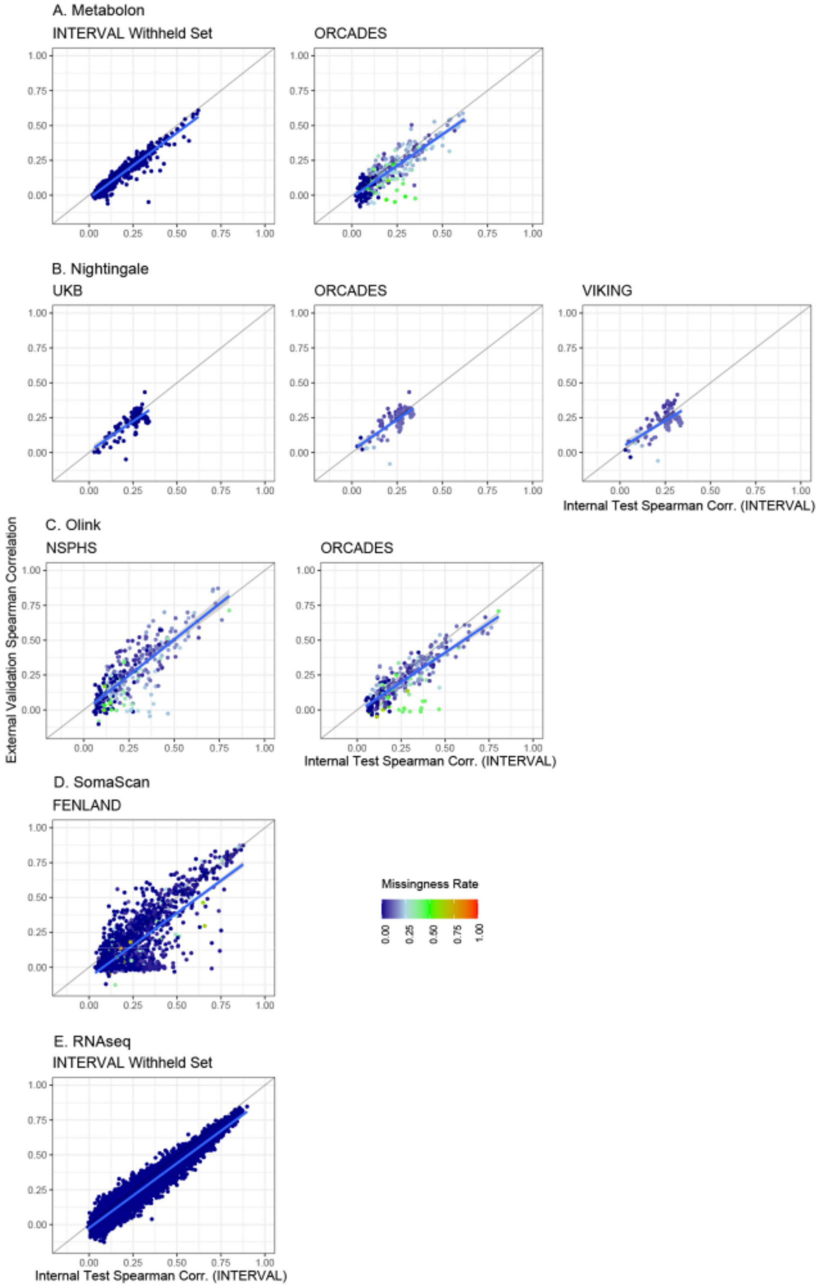
Extended Data Fig. 1 | Schematic framework for the development and validation of multi-omic genetic scores. This figure presents the overall study design for the development of genetic scores for multi-omic traits across five platforms (Nightingale, Metabolon, Olink, SomaScan and RNA-seq) using INTERVAL data as well as their validation in seven external cohorts of multiple ancestries (European, Asian-Chinese, Asian-Malay, Asian-Indian and African American).



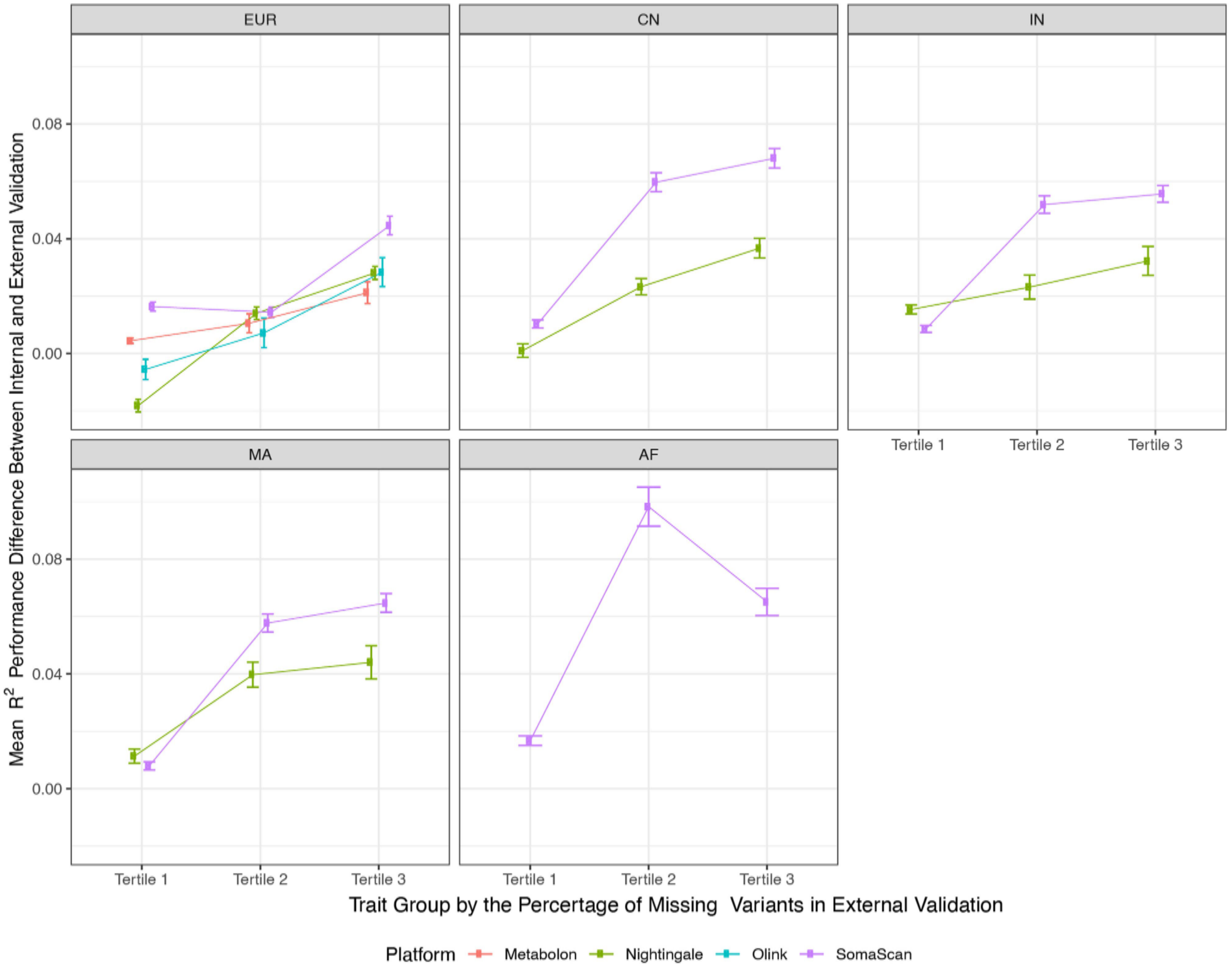
Extended Data Fig. 2 | R^2 performance comparison between Bayesian ridge, LDpred2 and P+T for Metabolon traits in external validation (INTERVAL withheld set). This figure compares the R^2 performance between BR (on the set of genome-wide variants with p-value $< 5 \times 10^{-8}$; x-axis) and LDpred2 (Hapmap3 variant set), and between BR and P+T (variant sets of two p-value thresholds: 5×10^{-8} and 1×10^{-3}) for 20 randomly selected Metabolon traits in external validation (INTERVAL withheld set; **Methods**). P-values in the GWAS for omic traits were derived by t-test in linear regression and all tests were two-sided.



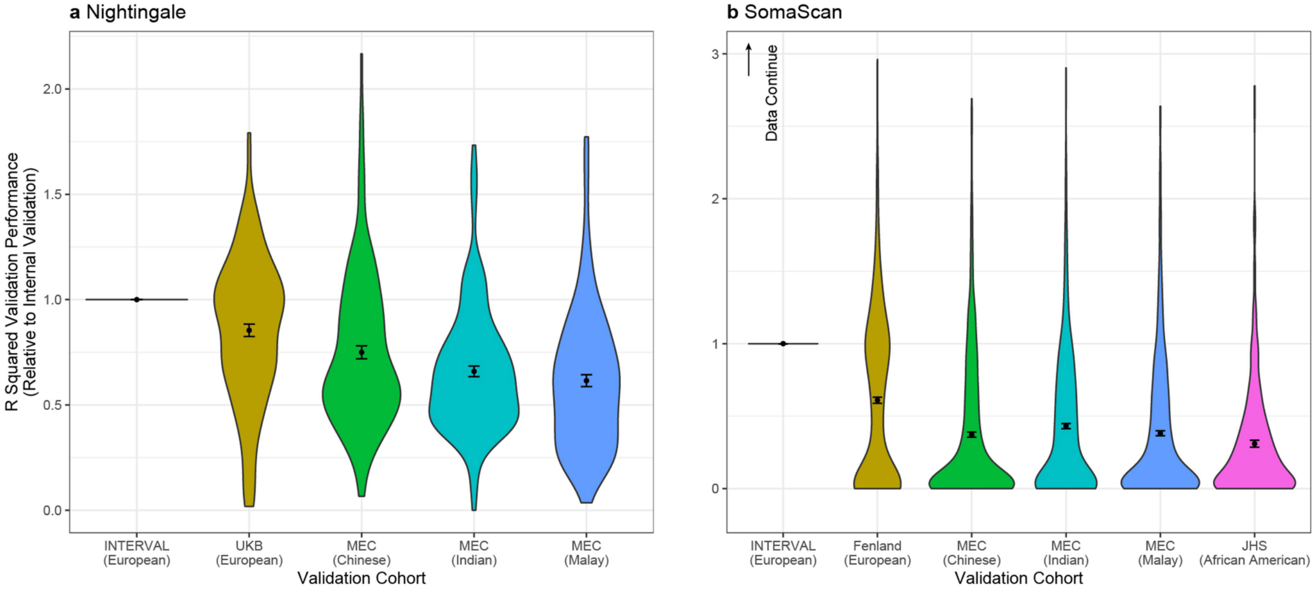
Extended Data Fig. 3 | Distribution of the number of variants in the genetic scores and the correlations between performance (R^2) of genetic scores and the number of variants comprising the score. The density plots show the distribution of the number of variants comprising the genetic scores at each platform. The scatter plots show the change of R^2 score in the internal validation by the number of variants in the genetic-score model.



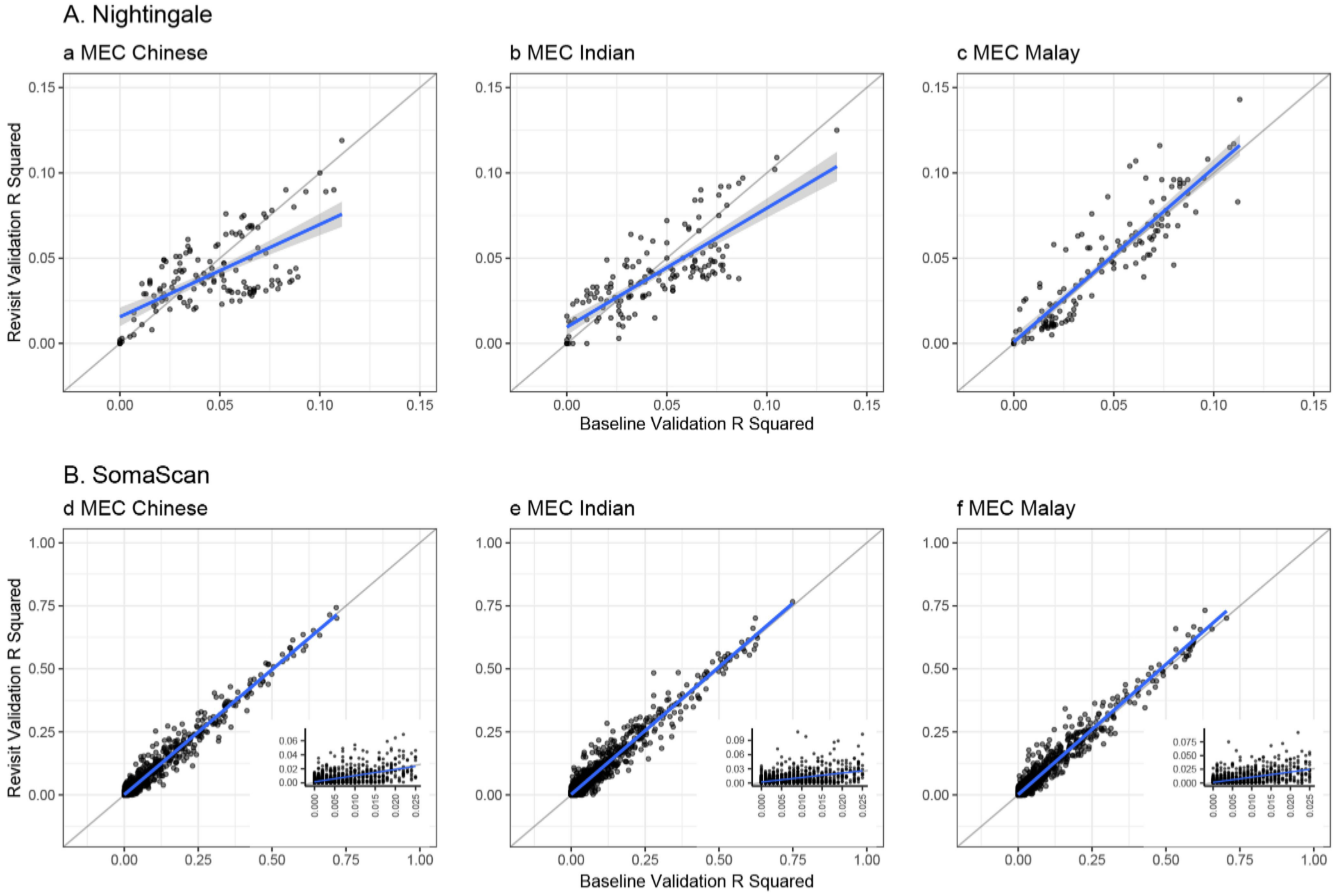
Extended Data Fig. 4 | Validation of genetic scores in external European cohorts. The scatter plots compare the spearman correlation scores between internal validation and external validation with a European cohort on each platform, in which points are coloured by the variant missingness rate in the external cohort and the blue line shows the linear models fitting the data points. This analysis included all the developed genetic scores in this study.



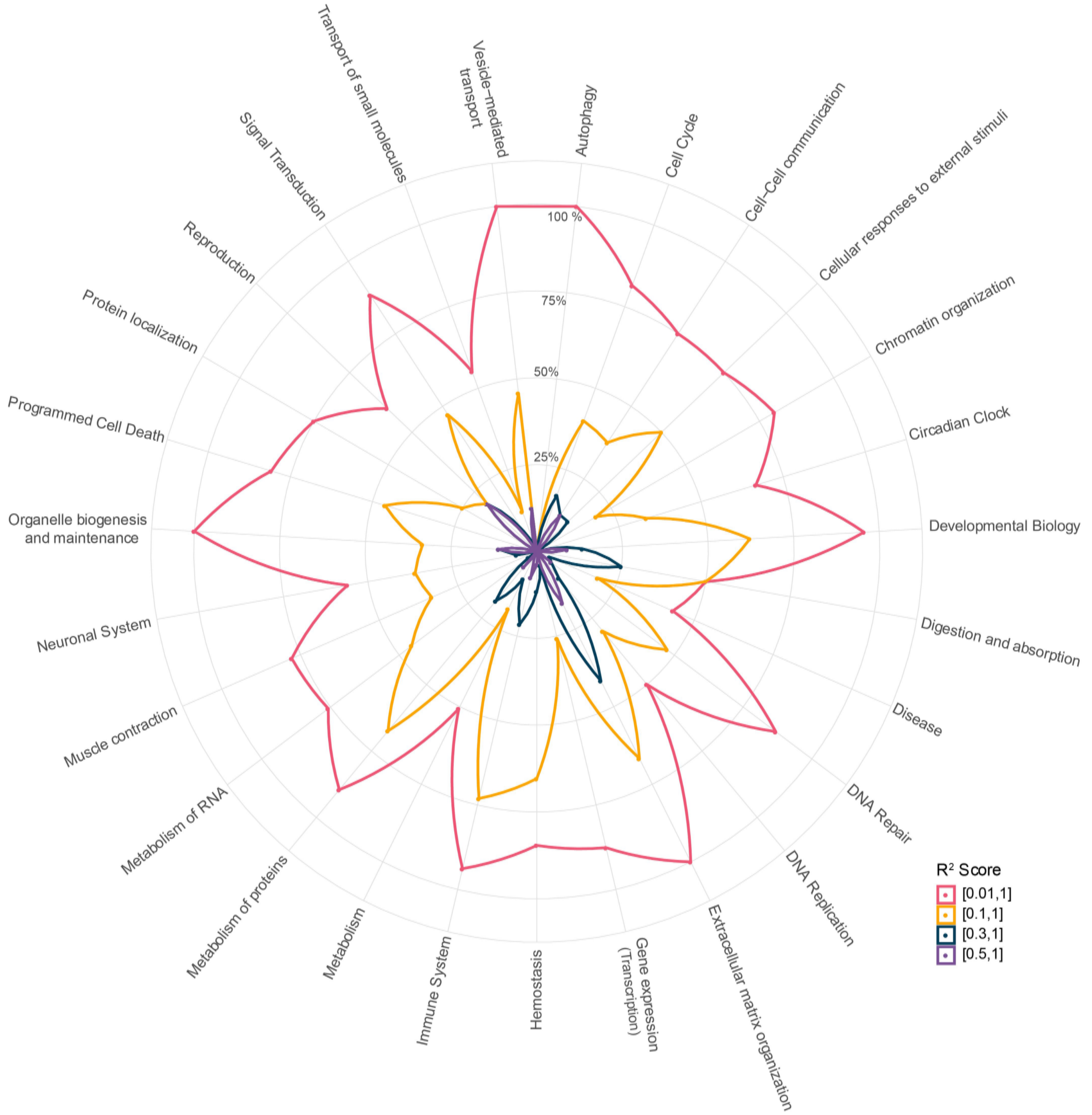
Extended Data Fig. 5 | Validation of the performance change of genetic scores by their variant missing rates in external cohorts of different ancestries. External validation results in European cohorts were merged in each platform to increase the statistical power in this analysis, which include NSPHS and ORCADES validations for Olink, and ORCADES and VIKINGS validations for Nightingale. Note that INTERVAL withheld subset validations and UKB validation for Nightingale traits were excluded in this analysis due to there is no or nearly no variant missingness in the external cohort (or INTERVAL withheld subset). Validation results in each platform were ranked by their variant missing rate of genetic-score models in the external cohort and grouped into tertiles, where variant missing rate is the number of variants missing in the validation cohort / the total number of variants in the genetic score. This figure presents the mean and standard error (SE) of R^2 performance change of genetic scores between internal and external validation across tertiles of validation results. The analysis included validation results of 2,129 SomaScan, 603 Olink, 455 Metabolon and 423 Nightingale traits (traits can be overlapped for the same platform across multiple validation cohorts) for European (EUR); 2,047 SomaScan and 139 Nightingale traits for Chinese (CN), Indian (IN) and Malay (MA); 820 SomaScan traits for African American (AF).



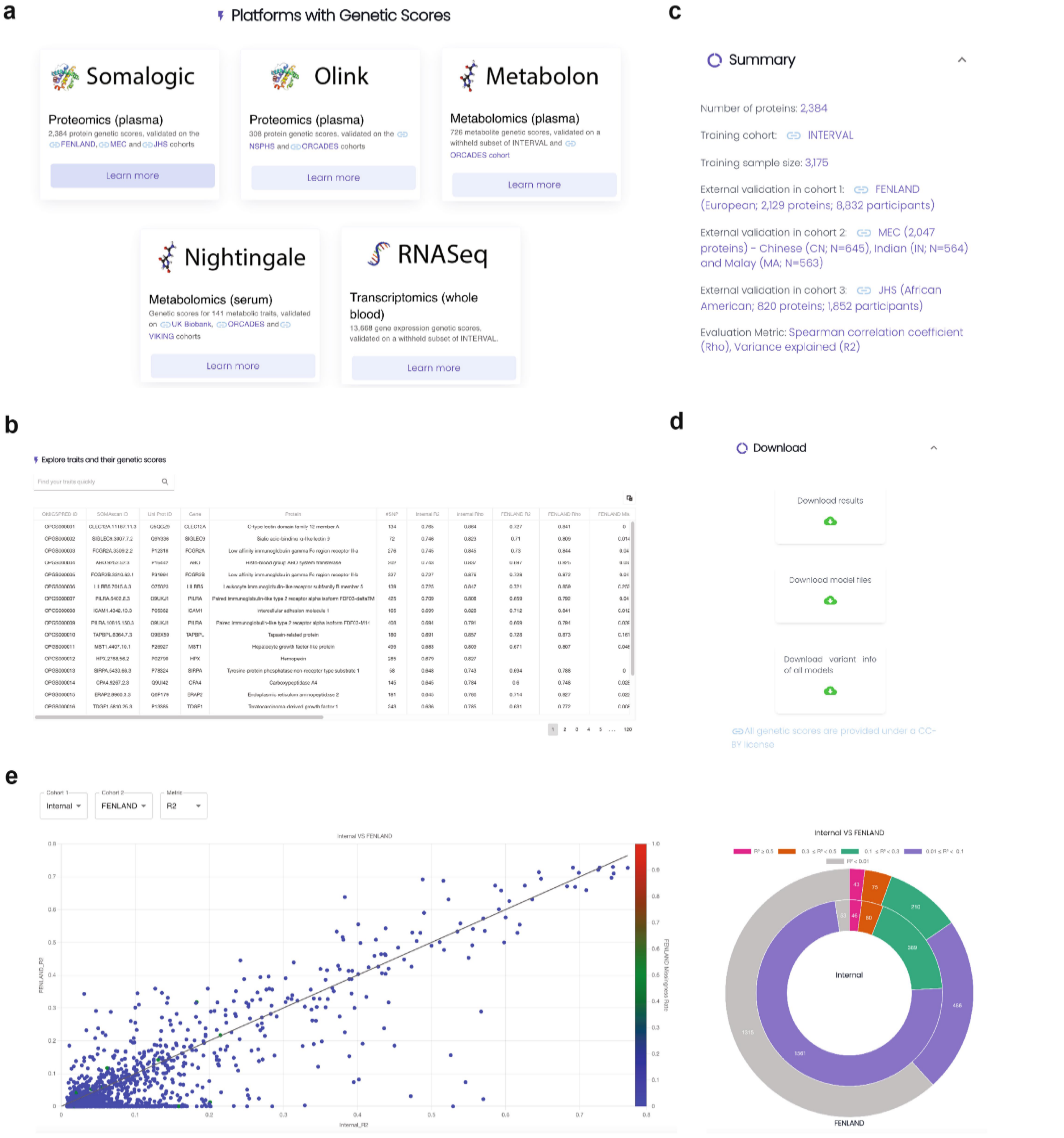
Extended Data Fig. 6 | Performance (R^2) of genetic scores for Nightingale and SomaScan in external cohorts of various ancestries relative to R^2 in internal validation (INTERVAL). a, Nightingale; **b**, SomaScan. Transferability was only tested if the genetic score had a significant (two-sided t-test; Bonferroni corrected p-value < 0.05 for all the 17,227 omic traits tested) association with the directly measured molecular trait in internal validation (n = 1631, 7471, 964, 635 and 827 for Metabolon, Nightingale, Olink, SomaScan and RNA-seq traits, respectively). This resulted in 137, 136 Nightingale metabolic traits for UKB (n = 98,245 participants) and MEC (Chinese, n = 1,067; Indian, n = 654; Malay, n = 634) respectively and 949, 1052, 378 SomaScan proteins for FENLAND (n = 8,832), MEC (Chinese, n = 645; Indian, n = 564; Malay, n = 563) and JHS (n = 1,852). Violin plots show distributions of the ratio of R^2 values. Black points show mean values and error bars are standard errors.



Extended Data Fig. 7 | Performance (R^2) of genetic scores between longitudinal samples and across ancestries in the MEC cohort. Paired samples include a baseline and a revisit sample from each individual run on SomaScan and Nightingale for MEC Chinese (N = 403 and 721 individuals), MEC Indian (N = 356 and 376) and MEC Malay (N = 353 and 363). Blue lines denote linear models fitted to each set of data points and the shaded areas represent 95% confidence intervals where applicable. There is no Nightingale genetic scores with a $R^2 > 0.15$ in both internal and MEC validation, so **a**–**c** only show R^2 in the range of [0, 0.15] for clarity. The sub-box plots at the right bottom of **d**–**f** show the validation results of these traits with baseline validation performance (R^2) between 0 and 0.025 in each ancestry.



Extended Data Fig. 8 | Coverage analysis for blood proteins in the lowest-level pathways. This analysis looked at all the lowest-level pathways of super-pathways curated at Reactome. Where at least one protein genetic score are included in the entities of a lowest-level pathway, we consider this pathway is covered by proteins of this study. This figure shows the percentage of the lowest-level pathways a group of proteins (by R^2 in internal validation) covered among all the lowest-level pathways of each super-pathway.



Platforms with Genetic Scores

Somalogic

Proteomics (plasma)

2,384 protein genetic scores, validated on the FENLAND, MEC and JHS cohorts

Learn more

Olink

Proteomics (plasma)

308 protein genetic scores, validated on the NSPHS and ORCADES cohorts

Learn more

Metabolon

Metabolomics (plasma)

726 metabolite genetic scores, validated on a withheld subset of INTERVAL and ORCADES cohort

Learn more

Nightingale

Metabolomics (serum)

Genetic scores for 141 metabolic traits, validated on UK Biobank, ORCADES and VIKING cohorts

Learn more

RNASeq

Transcriptomics (whole blood)

13,668 gene expression genetic scores, validated on a withheld subset of INTERVAL.

Learn more

Summary

Number of proteins: 2,384

Training cohort: INTERVAL

Training sample size: 3,175

External validation in cohort 1: FENLAND (European; 2,129 proteins; 8,832 participants)

External validation in cohort 2: MEC (2,047 proteins) - Chinese (CN; N=645), Indian (IN; N=564) and Malay (MA; N=563)

External validation in cohort 3: JHS (African American; 820 proteins; 1,852 participants)

Evaluation Metric: Spearman correlation coefficient (Rho), Variance explained (R2)

Explore traits and their genetic scores

Find your traits quickly

OMICSPRED ID	SOMAscan ID	Uni Prot ID	Gene	Protein	#SNP	Internal R2	Internal Rho	FENLAND R2	FENLAND Rho	FENLAND Mis
OPGS000001	CLEC12A.11187.11.3	Q5QGZ9	CLEC12A	C-type lectin domain family 12 member A	134	0.765	0.864	0.727	0.841	0
OPGS000002	SIGLEC9.3007.7.2	Q9Y336	SIGLEC9	Sialic acid-binding Ig-like lectin 9	72	0.746	0.823	0.71	0.809	0.014
OPGS000003	FCGR2A.3309.2.2	P12318	FCGR2A	Low affinity immunoglobulin gamma Fc region receptor II-a	276	0.745	0.845	0.73	0.844	0.04
OPGS000004	ABO.9253.52.3	P16442	ABO	Histo-blood group ABO system transferase	302	0.743	0.837	0.687	0.825	0.03
OPGS000005	FCGR2B.3310.62.1	P31994	FCGR2B	Low affinity immunoglobulin gamma Fc region receptor II-b	327	0.727	0.876	0.728	0.872	0.04
OPGS000006	LILRB5.7015.8.3	O75023	LILRB5	Leukocyte immunoglobulin-like receptor subfamily B member 5	139	0.725	0.847	0.721	0.859	0.252
OPGS000007	PILRA.5402.8.3	Q9UKJ1	PILRA	Paired immunoglobulin-like type 2 receptor alpha isoform FDF03-deltaTM	425	0.709	0.808	0.659	0.792	0.04
OPGS000008	ICAM1.4342.10.3	P05362	ICAM1	Intercellular adhesion molecule 1	165	0.699	0.826	0.712	0.841	0.012
OPGS000009	PILRA.10816.150.3	Q9UKJ1	PILRA	Paired immunoglobulin-like type 2 receptor alpha isoform FDF03-M14	408	0.694	0.791	0.669	0.794	0.036
OPGS000010	TAPBPL.6364.7.3	Q9BX59	TAPBPL	Tapasin-related protein	180	0.691	0.857	0.728	0.873	0.161
OPGS000011	MST1.4407.10.1	P26927	MST1	Hepatocyte growth factor-like protein	495	0.683	0.809	0.671	0.807	0.046
OPGS000012	HPX.2768.56.2	P02790	HPX	Hemopexin	285	0.679	0.827			
OPGS000013	SIRPA.5430.66.3	P78324	SIRPA	Tyrosine-protein phosphatase non-receptor type substrate 1	58	0.648	0.743	0.694	0.788	0
OPGS000014	CPA4.9267.2.3	Q9UI42	CPA4	Carboxypeptidase A4	145	0.645	0.784	0.6	0.748	0.028
OPGS000015	ERAP2.8960.3.3	Q6P179	ERAP2	Endoplasmic reticulum aminopeptidase 2	181	0.645	0.780	0.714	0.827	0.022
OPGS000016	TDGF1.5810.25.3	P13385	TDGF1	Teratocarcinoma-derived growth factor 1	243	0.636	0.785	0.631	0.772	0.006

1 2 3 4 5 … 120

Download

Download results

Download model files

Download variant info of all models

All genetic scores are provided under a CC-BY license

Cohort 1: Internal

Cohort 2: FENLAND

Metric: R2

Extended Data Fig. 9 | Key features of the OmicsPred portal for accessing genetic scores of multi-omic traits. a, Organization of genetic scores on the portal. **b**, Example of how biomolecular traits and their genetic-score-related information can be explored. **c**, Example of how summary statistics of training and validation cohorts are presented. **d**, Example of how validation results and genetic-score models can be downloaded. **e**, Example of how validation results and trait-related information can be visualized.

Extended Data Table 1 | Demographic statistics of training and validation samples for the construction of genetic scores of blood biomolecular traits by platform

Platform	Cohort	Ancestry	#Traits	#Samples	%Men	Age (years)	BMI (kg/m²)
Training and Internal Validation							
Metabolon	INTERVAL	European	726	8,153	51.0%	43.9 ± 14.1	26.4 ± 4.6
Nightingale			141	37,359	51.0%	43.7 ± 14.1	26.4 ± 4.6
Olink			308	4,822	59.3%	59.0 ± 6.7	26.5 ± 4.1
SomaScan			2,384	3,175	50.8%	43.6 ± 14.2	26.3 ± 4.7
Illumina RNAseq			13,668	4,136	56.4%	54.6 ± 11.6	26.6 ± 4.4
External Validation							
Metabolon	INTERVAL withheld subset	European	527	8,114	49.4%	47.9 ± 13.8	26.5 ± 4.6
	ORCADES		455	1,007	43.9%	54.0 ± 15.3	27.7 ± 4.9
Nightingale	UKB	European	141	98,245	45.8%	56.5 ± 8.1	27.4 ± 4.8
	ORCADES		141	1,884	40.0%	53.9 ± 15.0	27.8 ± 5.0
	VIKING		141	2,046	39.9%	49.8 ± 15.2	27.4 ± 4.9
	MEC	Chinese	139	1,067	47.2%	52.1 ± 9.9	23.5 ± 3.8
		Indian	139	654	43.7%	44.5 ± 11.6	26.4 ± 5.1
		Malay	139	634	42.9%	44.9 ± 11.1	26.9 ± 5.1
Olink	NSPHS	European	302	872	47.6%	49.6 ± 20.2	26.7 ± 4.8
	ORCADES		301	1,052	44.1%	53.8 ± 15.7	27.7 ± 4.9
SomaScan	FENLAND	European	2,129	8,832	47.1%	48.8 ± 7.4	26.9 ± 4.8
	MEC	Chinese	2,047	645	46.0%	51.9 ± 10.9	23.5 ± 3.9
		Indian	2,047	564	45.0%	44.0 ± 12.0	26.3 ± 5.3
		Malay	2,047	563	43.9%	44.4 ± 11.3	26.9 ± 5.2
	JHS	African American	820	1,852	39.0%	55.7 ± 12.8	31.6 ± 7.3
Illumina RNAseq	INTERVAL withheld subset	European	12,958	598	49.5%	45.0 ± 13.1	26.8 ± 4.8

The table shows the mean ± s.d. of age and BMI for participants in each cohort or cohort subset.

nature portfolio

Corresponding author(s): Professor Michael Inouye
Dr. Yu Xu

Last updated by author(s): Feb 3, 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

No software was used.

Data analysis

The following softwares and versions were used to perform the analyses:

- SHAPEIT3 (https://jmarchini.org/software/)
- PLINK v1.90b6.11 64-bit (24 Oct 2019) (www.cog-genomics.org/plink/1.9/)
- PLINK v2.00a2.3LM 64-bit Intel (24 Jan 2020)   (www.cog-genomics.org/plink/2.0/)
- Olink NPX Manager software (https://olink.com/products-services/data-analysis-products/npx-manager/)
- SNPTEST v.2.5.2 (https://www.well.ox.ac.uk/~gav/snptest/)
- STAR v2.7.3.a (https://github.com/alexdobin/STAR)
- featureCounts v2.0.0 (http://subread.sourceforge.net/)
- QTLtools v1.3.1 (https://qtltools.github.io/qtltools/)
- TensorQTL v1.0.6 (https://github.com/broadinstitute/tensorqtl)
- BWA-MEM v0.7.17 (https://github.com/lh3/bwa)
- GATK v4.0.6.0 (https://gatk.broadinstitute.org/hc/en-us)
- bigsnpr version 1.10.8 in R version 3.6.1
- pcaMethods version 1.86.0 in R version 4.1.3

- Python version 3.6.8 with the following Python packages:
  - numpy version 1.19.5
  - pandas version 1.1.5
  - scikit-learn version 0.21.2
  - scipy version 1.5.4

- lifelines version 0.26.0

The original codes used to train the genetic scores with INTERVAL data, internally validate these scores, and benchmark the performance of different genetic score construction methods are available at https://github.com/xuyu-cam/atlas_genetic_scores_omic_traits.

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

All the genetic score models trained in this study and GWAS summary statics used to develop genetic scores are publicly accessible through the OmicsPred portal (www.omicspred.org; accession codes OPGS000001-OPGS017227). INTERVAL study data from this paper are available to bona fide researchers from helpdesk@intervalstudy.org.uk and information, including the data access policy, are available at http://www.donorhealth-btru.nihr.ac.uk/project/bioresource.

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	This study used all the samples with the omic data available in each cohort. The sample sizes of each cohort (or cohort subset) used were given in the Extended Data Table 1 of the manuscript. In both internal and external validation, statistical significance was estimated using two-sided t-test and Mann-Whitney U test for all the developed genetic scores (Supplementary Tables 1-5).
Data exclusions	We excluded samples and variants based on the standard quality control for GWAS of the omic traits, which were described in our manuscript.
Replication	The explained variance and Spearman's rank correlation coefficient were used to measure the performance of constructed genetic scores across seven external cohorts (or INTERVAL withheld subset). Validation results were presented in detail in our manuscript.
Randomization	This question does not apply because all samples with omic data/phenotype available were used in the analysis.
Blinding	Not applicable because the analysis was performed using continuous omic data, and no case/control status were used.

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

Human research participants

Policy information about studies involving human research participants

Population characteristics

Population characteristics for each training/validation cohort (or cohort subset) used in this study were given in the Extended Data Table 1 of the manuscript.

Recruitment

Participants in all used cohorts, including INTERVAL Study, Fenland Study, Jackson Heart Study, Singapore Multi-Ethnic Cohort, Northern Swedish Population Health Study, Orkney Complex Disease Study, UK Biobank, VIKING Health Study, were recruited as part of previous studies. Our study has no influence or control on recruitment.

Ethics oversight

INTERVAL study was approved by the National Research Ethics Service (11/EE/0538) and all participants have given informed consent. The Fenland study was approved by the National Health Service (NHS) Health Research Authority Research Ethics Committee (NRES Committee – East of England Cambridge Central, ref. 04/Q0108/19) and all participants provided written informed consent. The ORCADES study was approved by Research Ethics Committees in Orkney, Aberdeen (North of Scotland REC), and South East Scotland REC, NHS Lothian (reference: 12/SS/0151), and all participants gave written informed consent. The VIKING health study was approved by the South East Scotland Research Ethics Committee, NHS Lothian (reference: 12/SS/0151), and all participants gave informed consent. The NSPHS study was approved by the local ethics committee at the University of Uppsala (Regionala Etikprövningsnämnden, Uppsala, Dnr 2005:325) in compliance with the Declaration of Helsinki, and all participants gave their written informed consent to the study. UK Biobank data access was approved under projects 7439, 11193 and 19655, and all the participants gave their informed consent for health research. Ethics approvals for MEC were provided by the SingHealth Centralised Institutional Review Board (IRB) and the National University of Singapore IRB, and all participants gave their written informed consent to the study. This study was approved by the Jackson Heart Study Publications and Presentations Subcommittee and the TOPMed Multi-Omics Working group.

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