scientific reports

OPEN

Polygenic prediction of human longevity on the supposition of pervasive pleiotropy

M. Reza Jabalameli[1], Jhih-Rong Lin[1], Quanwei Zhang[1], Zhen Wang[1], Joydeep Mitra[1], Nha Nguyen[1], Tina Gao[2], Mark Khusidman[1], Sanish Sathyan[3], Gil Atzmon[1,2,4], Sofiya Milman[1,2], Jan Vijg[1], Nir Barzilai[1,2,5] & Zhengdong D. Zhang[1✉]

The highly polygenic nature of human longevity renders pleiotropy an indispensable feature of its genetic architecture. Leveraging the genetic correlation between aging-related traits (ARTs), we aimed to model the additive variance in lifespan as a function of the cumulative liability from pleiotropic segregating variants. We tracked allele frequency changes as a function of viability across different age bins and prioritized 34 variants with an immediate implication on lipid metabolism, body mass index (BMI), and cognitive performance, among other traits, revealed by PheWAS analysis in the UK Biobank. Given the highly complex and non-linear interactions between the genetic determinants of longevity, we reasoned that a composite polygenic score would approximate a substantial portion of the variance in lifespan and developed the integrated longevity genetic scores (iLGSs) for distinguishing exceptional survival. We showed that coefficients derived from our ensemble model could potentially reveal an interesting pattern of genomic pleiotropy specific to lifespan. We assessed the predictive performance of our model for distinguishing the enrichment of exceptional longevity among long-lived individuals in two replication cohorts (the Scripps Wellderly cohort and the Medical Genome Reference Bank (MRGB)) and showed that the median lifespan in the highest decile of our composite prognostic index is up to 4.8 years longer. Finally, using the proteomic correlates of *iLGS*, we identified protein markers associated with exceptional longevity irrespective of chronological age and prioritized drugs with repurposing potentials for gerotherapeutics. Together, our approach demonstrates a promising framework for polygenic modeling of additive liability conferred by ARTs in defining exceptional longevity and assisting the identification of individuals at a higher risk of mortality for targeted lifestyle modifications earlier in life. Furthermore, the proteomic signature associated with *iLGS* highlights the functional pathway upstream of the PI3K-Akt that can be effectively targeted to slow down aging and extend lifespan.

The human lifespan is a complex trait that reflects the interplay of numerous socioeconomic factors and genetic predispositions. Its narrow-sense heritability (h^2) has been estimated in the range of 15 to 33%[1,2]. Recently, a lower estimate (< 10%) has been reported after correcting for assortative mating[3]. The magnitude of missing heritability suggests that a large portion of heritable variation in human survival may come from the infinitesimal effects of numerous loci spread widely across the entire genome[4]. In fact, findings from moderately powered genome-wide association analysis (GWASs)[5–7] are consistent with the perceived polygenic architecture of human lifespan, healthspan, and longevity. This notion complements other potential explanations like environmental influences and epistatic interactions.

The extensive polygenicity underlying human lifespan has some immediate implications. With a finite number of genes and a theoretically infinite number of age-related traits and endophenotypes, extensive pleiotropy would

[1]Department of Genetics, Albert Einstein College of Medicine, New York, NY, USA. [2]Department of Medicine, Albert Einstein College of Medicine, New York, NY, USA. [3]Department of Neurology, Albert Einstein College of Medicine, New York, NY, USA. [4]Faculty of Natural Sciences, University of Haifa, Haifa, Israel. [5]Institute for Aging Research, Albert Einstein College of Medicine, New York, NY, USA. ✉email: zhengdong.zhang@einsteinmed.edu

be inevitable[8,9]. The shared genetic component across correlated traits tags the common mechanisms amenable for drug development to slow the aging process and increase the healthspan[10,11]. If only a handful of rare alleles would underlie exceptional longevity, then the mutational target would be extremely narrow and, therefore, highly sensitive to population dynamics. In populations with a small effective size (N_e) or appreciable founder effect, the strong impact of genetic drift would have supplanted other evolutionary forces and ultimately either drive the longevity-increasing rare variants to fixation or complete removal from the gene pool[12]. In the absence of earlier observations among bottlenecked and isolated populations that acquired such extraordinary longevity in an evolutionarily short period (in terms of the number of generations), the extent of missing heritability attributable to rare variants remains to be determined.

We posited an infinitesimal model for longevity in which variation in human lifespan is primarily determined by the additive effect of many segregating common variants (MAF > 1%), each with a small effect size. Given the additive model, we reasoned that the polygenic score would approximate a substantial proportion variance in lifespan among long-lived individuals. Leveraging the pervasive pleiotropy, we modeled highly complex and non-linear interactions among segregating variants to construct the integrated longevity genetic scores (iLGSs) for distinguishing differential survival. We studied a cohort comprising 515 Ashkenazi Jewish centenarians and 442 ethnically matched controls with known age at death (Supplementary Fig. S1). We validated the predictive capacity of iLGS for distinguishing enrichment of exceptional longevity among the Wellderly cohort (n = 510) and Medical Genome Reference Bank (MRGB) cohort (n = 2,570) and subsequently converted the model to 3.8 million single nucleotide polymorphism (SNP) weights using the European descent portion of the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort (n = 62,268). Finally, we investigated proteomic correlates of exceptional longevity using iLGS and identified drugs with repurposing potentials for ameliorating the aging process. The schematic overview of the study is provided in Fig. 1.

Results

Tracking the allele frequency changes as a function of viability

Selective pressure early in life favors maximization of the fetal viability and fecundity during adulthood, although later in life, it attenuates maximizing the homeostasis maintenance[13–16]. The adaptive response to changing
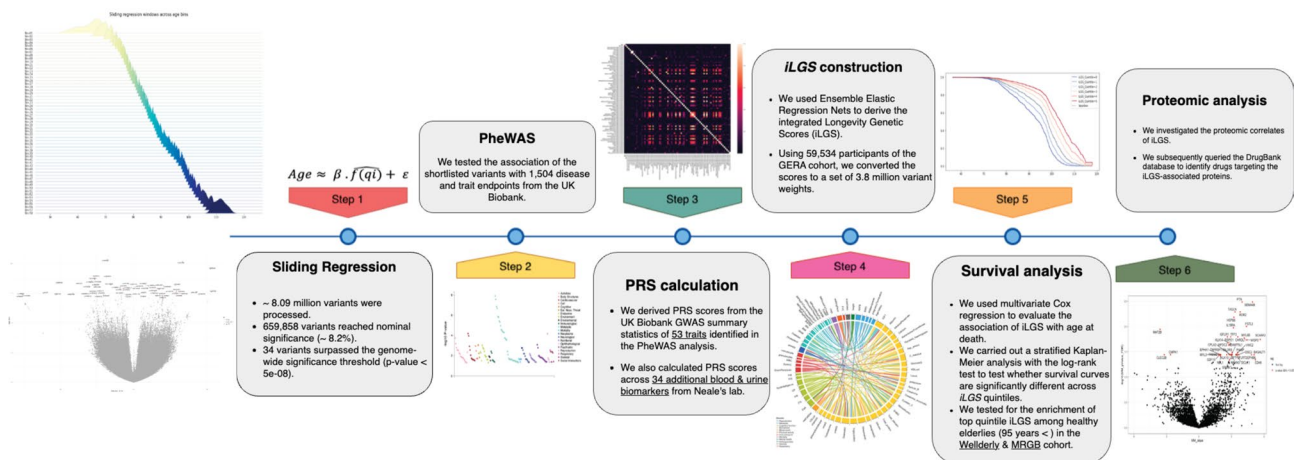

Fig. 1. The schematic overview of the analysis. Our analysis constitutes six major steps; in **Step 01**, we used the sliding regression framework to track allele frequency changes across 48 age bins in the Einstein LonGenity cohort. We fitted a robust linear regression with MM-estimator for each variant to derive the slope and p-value of association with chronological age. Variants surpassing the genome-wide significance threshold (p < 5e − 8) were followed up in the UK Biobank; in **Step 02**, the association of 34 candidate variants with 614 disease endpoints and traits were investigated in a PheWAS analysis. For each variant, we used the Bonferroni-corrected association p-values for shortlisting associated traits; in **Step 03**, we used GWAS summary statistics across the 87 traits (53 associated traits from PheWAS analysis + 34 additional blood and urine biomarkers) to calculate the polygenic scores (PGS) for unrelated participants in the Einstein LonGenity cohort (n = 957); in **Step 04**, we split our cohort into the derivation set (consisting of training and test sets, n = 715) and validation set (n = 242). Using the derivation portion of our cohort, we applied a stacked Elastic-net regression framework to effectively combine PGS across the 87 traits and construct a composite prognostic score to distinguish survival chances. We trained the ensemble model on 65% of the derivation set and tested on the remaining 35%. Our model achieved an AUC of 0.87 in the validation set. Using coefficients derived from our stacked model, we computed the integrated Longevity Genetic Scores (iLGS). Subsequently, using an external cohort (GERA cohort), we converted the scores to a set of 3.8 million variant weights. In **Step 05**, we carried out survival analysis to test the performance of iLGS in predicting survival in the validation set as well as two external cohorts including the Wellderly and MRGB cohorts; in **Step 06**, we carried out a sex-stratified association analysis to identify proteomic correlates of iLGS. Proteins significantly associated with iLGS were subsequently queried in the DrugBank database to identify druggable targets. Drugs targeting or interacting with iLGS associated proteins were investigated for potential repurposing as senolytic.

selective pressure often entails allele frequency changes at many segregating loci influencing the trait[9]. Determinants of the evolutionary trade-off between viability and fecundity characterize the polygenic response that results in the fluctuation of allele frequencies across different age groups[17]. To identify variants underlying exceptional longevity, we adopted a regression framework similar to that of Bergman et al.[18], which tracks allele frequency changes across different age bins. Variants with a negative impact on longevity are expected to be naturally purged from the centenarian gene pool, while those with a positive contribution to healthy aging are expected to be enriched in the gene pools toward the extremities of the human lifespan.

In constructing our sliding regression framework (see "Methods"), we assumed that due to the specific age structure of our discovery cohort (between the ages of 56 and 111), the fitness cost of pro-aging alleles monotonously decreases their presence in the gene pool, while pro-longevity alleles will monotonously increase in frequency. Neutral alleles with no effect on longevity will appear with constant frequency during aging. Hence, the regression slope would reflect the overall direction of the effect conferred at each variant by the minor allele. We tested 8,090,427 variants in total and identified 34 independent lead SNPs at the genome-wide significance threshold ($P < 5E - 8$) (Supplementary Fig. S2 and Supplementary Table S1). Allele frequency of these 34 lead SNPs in centenarians and controls are also displayed in the Supplementary Table S1. These variants are sex independent as allelic frequencies in each age bin were corrected for the sex effect (see Methods). Among the 34 variants, two are coding, and the rest are non-coding (mapping to intronic and intergenic regions), and 20 variants are tagging expression Quantitative Trait Loci (eQTL) (Supplementary Table S2). As proof of principle, we also tracked the frequency of each of the three *APOE* haplotypes (ε2, ε3, and ε4) tagged by rs429358 and rs7412 across directly genotyped and phased samples. As expected, we identified a significant negative effect ($\beta = -21.84$, $P = 4.25E - 7$) for *APOE*-ε4 haplotype (Supplementary Fig. S3).

Several genomic loci included multiple prioritized genes, including two of the previously reported contributors of aging, such as *LPA* and *LDLR*[6]. The strongest association signal for longevity (i.e., rs76430661) was on 5q35.3 where six genes were prioritized (*COL23A1*, *HNRNPAB*, *N4BP3*, *ZNF454*, *ZNF879*, and *PRELID1*. (Supplementary Fig. S4). The variant physically maps to the second intron of *COL23A1* and is an eQTL for all the prioritized genes apart from *ZNF454* which is prioritized based on chromatin interactions. While none of the prioritized genes in the locus has been previously described in the context of aging, upregulation of a closely related family member of *HNRNPAB* is suggested to prevent age-dependent cognitive decline in Alzheimer's disease (AD) mouse models[19].

The largest two risk loci on 6q26 and 19q13 (tagged by rs41272112 and rs147053538), each with 16 and 50 genes, respectively, represented the highest density of prioritized coding genes physically located in the locus (Supplementary Table S3). The lead SNP on 6q26 (rs41272112) maps to the exon 26 of *LPA* and the risk locus is an eQTL for 8 genes including *LPA* itself and *PLG*. The locus shows chromatin interactions with *IGF2R*, *AGPAT4*, and *FNDC1* (Supplementary Fig. 5). Exposure to high Lp(a) (lipoprotein A) levels is associated with coronary heart disease[20] and was recently shown to be causally related to shortened parental lifespan in the UK Biobank[21]; nevertheless, the relationship between Lp(a) concentrations and all-cause mortality among patients with established cardiovascular diagnosis is debated[22]. Given that the prioritized missense variant in the locus is not damaging (CADD: 0.06, PrimateAI: 0.30), it is likely that its negative association with longevity ($\beta = -0.21$, $P = 4.2E - 08$) is mediated through a more stable Lp(a) resulting from the c.4262G > A mutation. It is worth noting that the risk locus also shows chromatin interactions with the insulin-like growth factor 2 receptor (*IGF2R*), which has been identified to be associated with parental longevity[7]. Furthermore, coding variant rs3798220, which is in complete LD with the prioritized tag SNP, has been identified to show the most substantial individual-level effect in "lost healthy life years" across the participants of the UK Biobank and FinnGen cohort[23].

The lead SNP rs147053538 on 19q13 maps to the intergenic region between *PSG11* and *PSG7*, and the risk locus is an eQTL for six genes including *ARHGEF1* and shows chromatin interactions with 13 additional genes including *CEACAM1* (Supplementary Fig. S6). Rho Guanine Nucleotide Exchange Factor 1 (*ARHGEF1*) was identified to be crucial in angiotensin II-induced hypertension, and its inactivation has been suggested as an amenable target for treating high blood pressure[24]. Given the proposed association of hypertension with late-stage dementia and Alzheimer's disease[25,26], the net effect of the risk loci in modulating the expression of *ARHGEF1* merits further investigation. Remarkably, age-dependent upregulation of carcinoembryonic antigen-related cell adhesion molecule 1 (*CEACAM1*) is identified to result in endothelial impairments and promotion of atherosclerotic plaque formation during aging[27]. Whether the favorable impact of the locus on longevity is mediated through the downregulation of *ARHGEF1* expression or *CEACAM1* (or both) is not clear, but we speculate that the locus implicates an important cardiovascular component of longevity that can be therapeutically targeted for ameliorating aging.

The most prolific regulatory activity was observed by rs11556579, where 43 genes were prioritized by both eQTL mapping and chromatin interactions, including *MAN1B1* on 9q34 (Supplementary Fig. S7). ER mannosidase I (Man1b1) is involved in the intracellular clearance of misfolded alpha1-antitrypsin[28]. Accumulation of misfolded α1-antitrypsin plaques in the lesions of AD is previously reported[29], and it has been shown that properly folded protein protects against amyloid-β-induced toxicity in microglial cells[30]. Moreover, treatment of type-2 diabetes mouse models with α1-antitrypsin rescues glucose intolerance and normalizes blood glucose levels[31]. In light of recent findings on the significance of heme homeostasis in longevity and the proposed mechanism of action for metformin[32] that mimics α1-antitrypsin protection against heme oxidation, it is highly likely that the favorable impact of the locus is mediated through the regulation of *MAN1B1* expression that ensures conformational acuity of α1-antitrypsin. The expression of the gene is downregulated in AD[33], and we speculate that the risk locus implicates the link between T2D and AD.

Additionally, the lead SNPs on 2q24 and 2q32 (i.e., rs6757605 and rs59642822, respectively) tagged eQTL loci that regulate the expression of *GPD2* and *GLS1* respectively (Supplementary Fig. S8). Suppression of *gpd-2* in C. elegans is shown to further extend the lifespan of *daf-2* mutants[34] and inhibition of *GLS1* ameliorated age-related

3

pathologies by eliminating senescent cells in aged mice[35]. Given that downregulation of both genes results in extended lifespan in animal models, it is likely that the favorable effect of these two loci on human longevity is mediated by the attenuation of gene expression by eQTLs in LD with the lead SNPs.

Overall, positional, eQTL, and chromatin interaction mapping prioritized 332 genes of which 191 genes are exclusively the target of expression alteration by the 20 eQTLs. Gene-set analysis suggests a shared biological function that is enriched in the CD40 pathway, thyroid hormone-mediated pathway, osteoblast proliferation, and apolipoprotein binding; however, significance levels do not withstand the multiple-test correction (Supplementary Table S4).

iLGS model construction

Multimorbidity is a common phenomenon during aging[36]. The shared genetic component among age-related traits (ARTs) and age-related diseases (ARDs) results in genetic correlation and underlies certain multimorbidities[37]. Given the limited size of our discovery cohort, any polygenic derivation of additivity of effect solely based on the few associated signals (from our sliding regression framework) is massively underpowered and not capable of capturing the full spectrum of additive variance underlying longevity. However, given the shared pathways among ARTs and ARDs[38], we reasoned that combining the polygenic effect among age-related pleiotropic traits enables the robust approximation of the shared genetic effect in the common pathways related to human lifespan.

To explore the pleiotropic landscape of the 34 variants identified in the sliding-regression framework, we carried out a phenome-wide association study (PheWAS) using GWAS summary statistics primarily from the UK Biobank (see "Methods"). Overall, we tested the association of 1,504 traits across 28 domains, of which 223 unique traits and diseases across 17 domains were identified to be significantly associated with the candidate variants (Supplementary Table S5). The association signals were significantly enriched across three domains including Skeletal ($P = 1.0E\text{-}05$, OR = 4.39), Metabolic ($P = 6.8E\text{-}05$, OR = 1.90), and Cognitive domains ($P = 6.3E\text{-}05$, OR = 3.20) (Supplementary Table S6). The most significant associations across these three domains included traits such as heel bone mineral density (skeletal domain), impedance measures of body fat percentages (metabolic), and overall cognitive performance.

Given the ubiquitous pleiotropy revealed in the PheWAS analysis, we constructed a model that integrates the complex interplay of pleiotropic ARTs in shaping longevity. Here, we consider genes as the units determining longevity and presume that total contribution to longevity is better approximated by the additive effect of their constituent functional variants on pleiotropic traits determining the evolutionary trade-offs throughout the life history of the species[39]. Traits identified by PheWAS analysis either causally influence longevity or simply arise from the spurious pleiotropy due to linkage disequilibrium (LD)[40]. Regardless of the true nature of the genetic correlation between these traits and longevity, we reasoned that a model summarizing the total additive genetic variance conferred by genetically correlated traits while accounting for multicollinearity among them would yield a descriptive statistic that can be used to distinguish survival.

We applied a stacked ensemble method to construct the integrated longevity genetic score (iLGS). We first randomly split the Einstein LonGenity cohort into the derivation ($n = 715$) and validation ($n = 237$) sets. We then used polygenic risk scores of 53 select PheWAS-identified traits and 34 additional UK Biobank blood and urine biomarker traits (See "Methods"). We trained and evaluated the ensemble model using fivefold cross-validation on the derivation dataset, with each fold consisting of 65% training data and 35% testing data. Our model achieved an area under the curve (AUC) of 0.91 in the derivation set and 0.87 in the validation set. The schematic overview of the model construction is provided in Fig. 2.

The stacked Elastic-net regression framework shrank the coefficients of fifteen traits to zero, and the final score was constructed using the coefficients of the remaining 72 traits (out of 87 initial traits) and 297 interaction terms (See "Methods"). The size of adjusted coefficients in the final model is provided in Supplementary Tables S7 and S8. As expected, the pattern of genetic correlation (r_g) among the traits included in our model revealed several distinct clusters (Fig. 3): (1) The biggest cluster includes different body impedance measures that are significantly correlated with different metabolic and cardiovascular traits (including systolic blood pressure, diabetes, birth weight, and the birth weight of first child), different metabolic traits and anxiety; (2) Impedance measures are negatively correlated with numerous reproductive traits, which include age at menarche, age at first sexual intercourse, age at first birth and age at menopause; (3) Fat impedance measures are also negatively correlated with different measures of fluid intelligence and usual walking pace; (4) Among the biomarker traits the greatest number of intra-domain correlations belong to C-reactive protein and triglyceride (each with 28 significant inter-domain genetic correlation), which accounts for most intra-domain correlations.

Across the individual PGS (polygenic score) terms, high-density lipoprotein (HDL) cholesterol, fluid intelligence (chained arithmetic measure), and forced vital capacity are positive top contributors to longevity, while glycated hemoglobin (HbA1c), apolipoprotein B, and body mass index (BMI) are top negative contributors to longevity. Our model also entails the interactions among traits (Extended Fig. 1 and Supplementary Fig. S9). Since the iLGS is linearly correlated with longevity, traits with multiple interacting partners can potentially tag canonical longevity features. Several interesting insights are immediately discernible from the interacting pairs in our model. Across the 12 domains, traits in the "biomarkers" category have the most interacting partners. Out of the 34 blood and urine biomarkers included in the model, 32 interact with at least three other traits. The insulin-like growth factor 1 (IGF1) has the highest number of interactions (17 partners), followed by rheumatoid factor (15 partners), alkaline phosphatase, and estradiol (both with 14 partners). (Extended Fig. 1). The "cognitive function" traits have the second-highest number of interactions with other traits among the remaining domains. Within this domain, educational attainment interacts with 14 other traits, representing the most important non-biomarker traits in describing the model variance (Extended Fig. 1). Traits in the metabolic, immunological, and
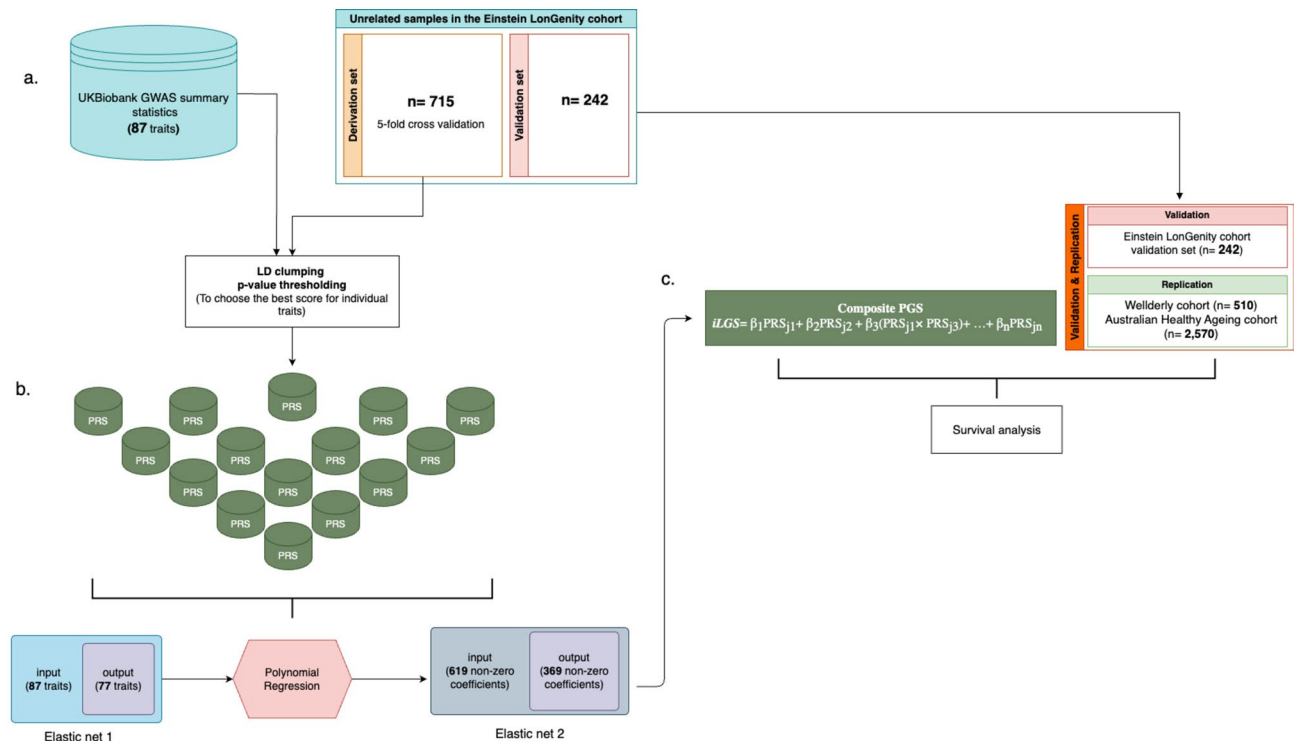
Fig. 2. Overview of the *iLGS* model construction pipeline. We derived integrated Longevity Genetic Scores (*iLGS*) in three steps: (**a**) First, across unrelated individuals in the Einstein LonGenity cohort, we calculated polygenic scores for a total of 87 traits (53 associated traits from PheWAS analysis + 34 additional blood and urine biomarkers). We used LD clumping and p-value thresholding to derive the best score for each trait; (**b**) Next, we applied a stacked Elastic-net regression framework to combine the polygenic scores and derive *iLGS*. This method entails passing the output from one model to the next to increase model accuracy and derive meaningful insight. By sandwiching a polynomial regression between two Elastic-net regressions, we created a search space for the model to select the most informatic PGS and their interactions for predicting longevity. We started with 87 PGS as input and derived 369 PGS and interaction terms with model accuracy of around 91% in the derivation set upon fivefold cross-validation; (**c**) Using the validation portion of the Einstein LonGenity cohort (n = 242) and two additional cohorts, including the Wellderly (n = 510) & Australian Healthy Ageing cohort (MRGB; n = 2,570) we replicated the predictive performance of iLGS for distinguishing differential lifespan. Our model achieved an area under the curve (AUC) of 87% in the validation portion of the Einstein LonGenity cohort.

reproduction domains, each with 45, 34, and 29 total interactions, respectively, account for most of the remaining cross-trait interactions in our model. When traits were considered individually, age at first birth, anxiety, and walking pace has the most interacting partners, following educational attainment (Extended Fig. 1).

Based on our model, the interaction between fluid intelligence and serum total bilirubin has the most favorable contribution to longevity, and the interaction between IGF1 and direct bilirubin has the most negative effect on longevity. Across the individual PGS terms, high-density lipoprotein (HDL) cholesterol, fluid intelligence (chained arithmetic measure), and forced vital capacity were revealed as the top contributors to increased longevity, and Glycated hemoglobin (HbA1c), Apolipoprotein B, and body mass index (BMI) were identified as the top negative contributors to longevity.

iLGS distinguishes differential survival

Using the coefficients derived from our stacked model, we computed the *iLGS* across 59,534 individuals with European ancestry in the GERA cohort and converted the scores to a set of 3.8 million variant weights (See "Methods" and Supplementary Fig. S10). We subsequently assessed the risk prediction performance of *iLGS* for distinguishing differential longevity in three independent cohorts.

Using the validation portion of our in-house cohort (n = 242), we applied a multivariate Cox proportional hazard model to evaluate the association of *iLGS* with age at death while controlling for sex, *APOE4* status, and their interaction with *iLGS* (See "Methods"). *iLGS* was associated with age at death with a hazard ratio (HR) of 0.09 (95% CI = [0.05, 0.18], $P = 6.9E-13$) per standard deviation of *iLGS* (Fig. 4a). This is equivalent to a 3% decrease in the baseline mortality hazard per one unit increase in the *iLGS*. The role of sex has been widely recognized as a factor that contributes to differences in exceptional longevity[41], and, expectedly, we identified a significantly increased hazard rate for mortality among males against the baseline hazard (Fig. 4a; HR = 1.84, 95% CI [1.25, 2.73], likelihood ratio test in Cox proportional hazard model, $P = 2.10E-3$). However, no statistical interaction
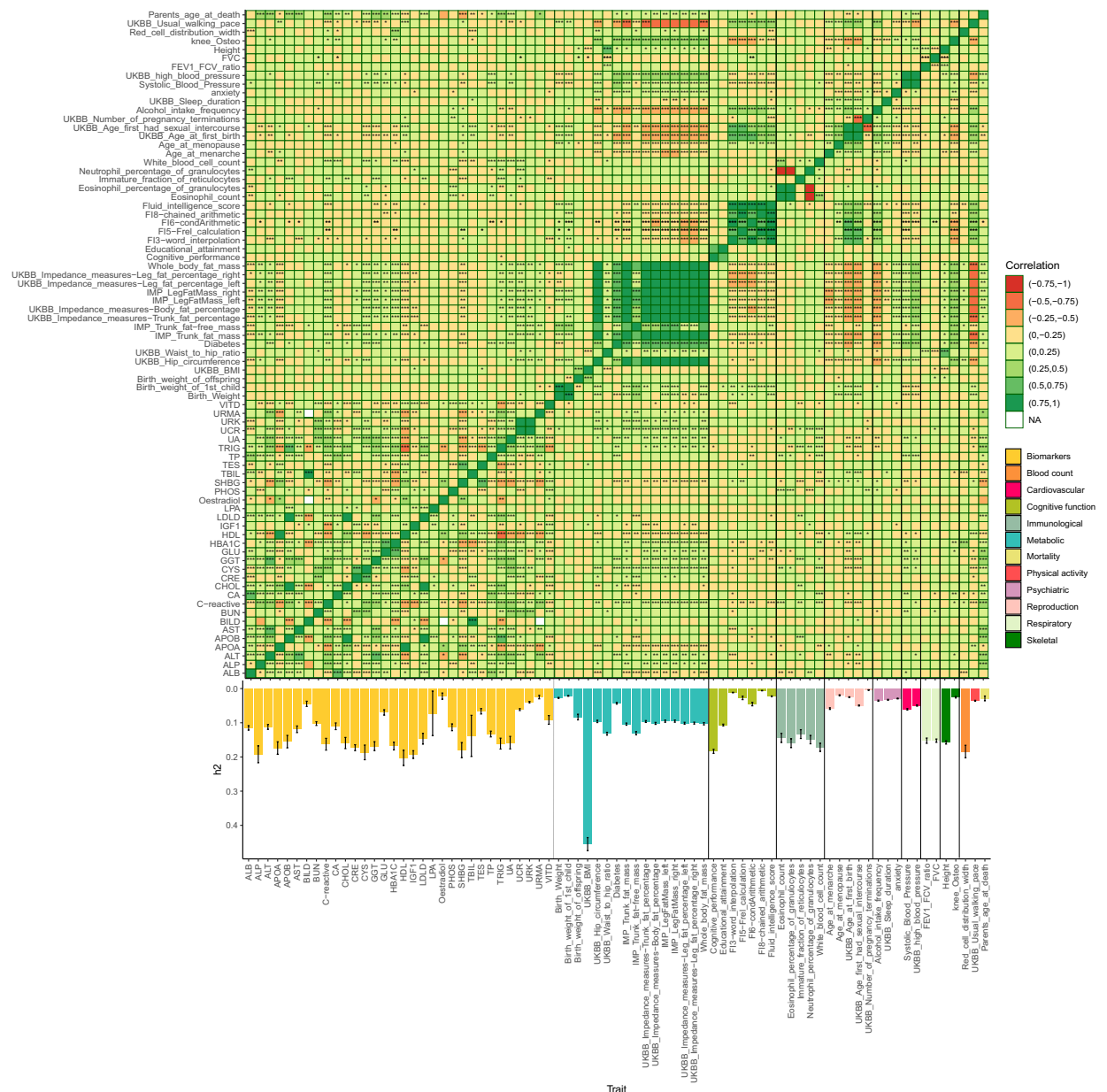
Fig. 3. Heatmap plot showing the pattern of genetic correlation and the magnitude of narrow-sense heritability across the traits used for constructing the *iLGS*. Pairwise genetic correlation (r_g, n = 2,556) across the 72 traits (out of 87 initial traits) used in constructing *iLGS* was calculated using bivariate LD score regression. Positive and negative genetic correlations are shown in green and red respectively. P-values for the significance of genetic correlations are FDR corrected using the Benjamini–Hochberg method. The magnitude of the r_g significance is indicated by the asterisks. The bars at the bottom of the plot show the magnitude of narrow-sense heritability across the 72 traits. The colour of each bar corresponds to the respective domain of the trait; (ALB: Albumin, ALP: Alkaline phosphatase, ALT: Alanine aminotransferase, APOA: Apolipoprotein A, APOB: Apolipoprotein B, AST: Aspartate aminotransferase, BILD: Direct bilirubin, BUN: Urea, CA: Calcium, CHOL: Cholesterol, CRE: Creatinine (in serum), CYS: Cystatin C, GGT: Gamma glutamyltransferase, GLU: Glucose, HBA1C: Glycated haemoglobin, HDL: HDL cholesterol, IGF1: Insulin growth factor-1, LDLD: LDL cholesterol, LPA: Lipoprotein A, PHOS: Phosphate, SHBG: Sex hormone binding globulin, TBIL: Total bilirubin, TES: Testosterone, TP: Total protein, TRIG: Triglycerides, UA: Urate, UCR: Creatinine in urine (enzymatic), URK: Potassium in urine, URMA: Microalbumin in urine, VITD: Vitamin D, FVC: Forced vital capacity, FEV1: Volume that has been exhaled at the end of the first second of forced expiration).

between *iLGS* and sex was identified for mortality hazard (Fig. 4a; HR = 0.43, 95% CI [0.17, 1.10], *P* = 7.76E − 2), which indicates that the association of *iLGS* with survival is largely sex-independent and cross-gender difference in overall survival is not influenced by any sex-specific effects of *iLGS*. Furthermore, given the well-known
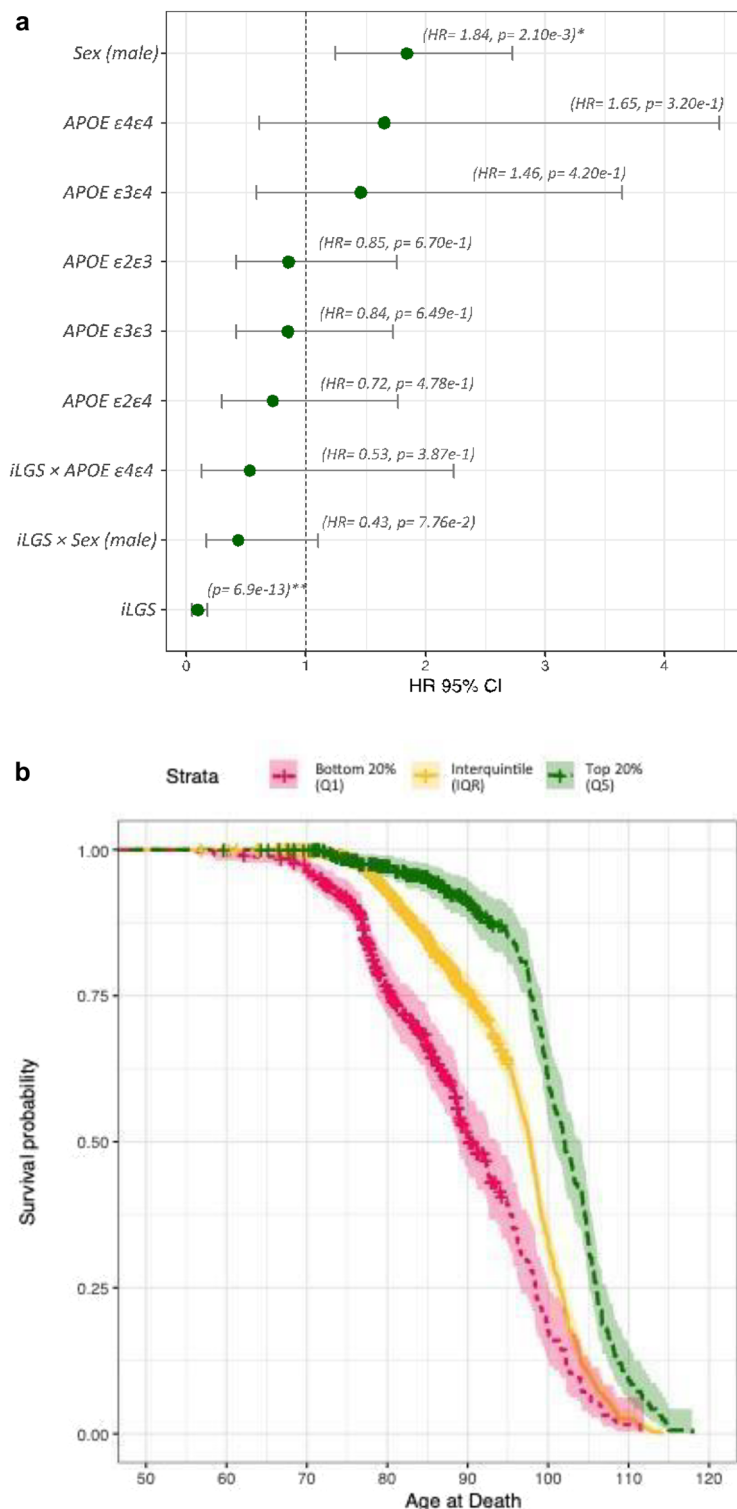
Fig. 4. The association of *iLGS* as compared with other risk factors with longevity and differential survival across the quintiles of *iLGS*. (**a**) Cox regression hazard ratios (HR) for the association of sex, *APOE* haplotypes and *iLGS* with age at death. An HR greater than 1 suggests an increased risk of death, and an HR below 1 indicates a smaller risk. Expectedly, the risk of earlier death is increased among males. Neither of the *APOE* haplotypes was revealed to be significantly associated with age at death in the study population. However, the direction of HRs was consistent with the unfavourable effect of *APOE* $\varepsilon 4$ haplotypes on longevity. We also modelled the interaction of iLGS with *APOE* $\varepsilon 4$ haplotype and sex to confirm that the association is independent of the sex effect of *APOE* haplotype status. (**b**) Stratified Kaplan–Meier curves showing differential survival across the top (Q_{iLGS}^{5}) and bottom quintile of iLGS (Q_{iLGS}^{1}) in comparison to the survival of individuals with *iLGS* in the interquintile range (log-rank test between the top quintile (Q_{iLGS}^{5}) and IQR: p = 2.47e − 5, log-rank test between the bottom quintile (Q_{iLGS}^{1}) and IQR: p = 2.75e − 27). (**c**) Plot depicting the Cox regression HRs for the association of *iLGS* quintile with age at death. Given that our inhouse cohort is primarily ascertained to include centenarians, we adopted the Q_{iLGS}^{1} as the baseline hazard and calculated the HRs for the remaining quintile against the hazards of Q_{iLGS}^{1}. We also included sex as a covariate to adjust for the sex effect.
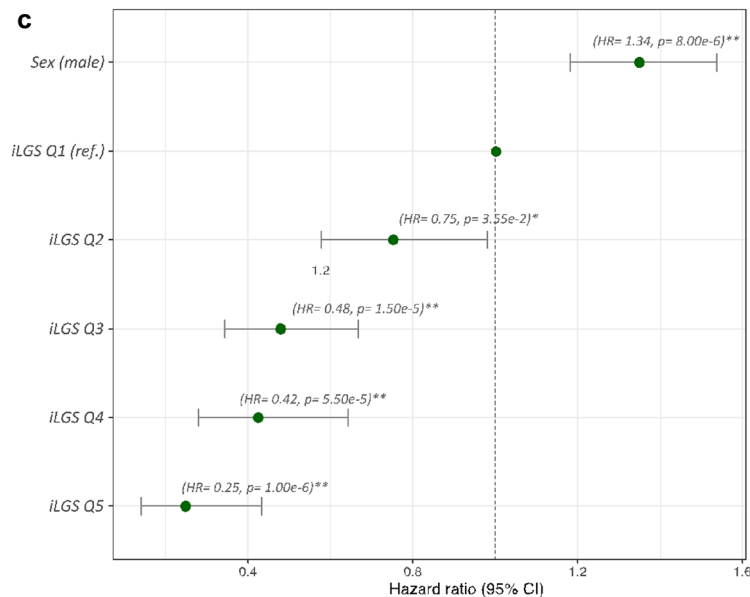
Fig. 4. (continued)

effect of *APOE* haplotypes on human longevity[42,43], we also investigated the association of all these haplotypes with age at death and modeled the interaction of *APOE*-ε4ε4 genotype with the *iLGS* in our multivariate Cox regression model (See "Methods"). Neither of the *APOE* haplotypes was significantly associated with mortality in our validation set (Fig. 4a). Similarly, the effect of *iLGS* on survival appears to be independent of *APOE*-ε4 status as we did not identify any evidence of statistical interaction between *iLGS* and the *APOE*-ε4ε4 haplotype (HR = 0.53, 95% CI = [0.12, 2.24], P = 3.87E-1) (Fig. 4a). We carried out a stratified Kaplan–Meier analysis across the quintiles of *iLGS* to investigate if survival curves are significantly different across the top and bottom quintiles compared to the interquintile range (IQR) (Fig. 4b and Extended Fig. 2). The cumulative incidence of death was significantly different between the top quintile (Q_{iLGS}^5) and IQR (log-rank test, P = 247E − 5) and between the bottom quintile (Q_{iLGS}^1) and IQR (P = 2.75E − 27). A comparison of Kaplan–Meier functions across deciles revealed that, on average, individuals with *iLGS* in the top decile live up to 4.8 years longer. Furthermore, the risk of death for individuals in the bottom two quintiles (i.e., bottom 40%) is maximum before the age of 85, while for individuals in the top quintiles, the risk of death does not peak until over the age of 95 years (Extended Fig. 2). The better survival outcome across the top quintile of *iLGS* is consistent with the coefficient estimates from the Cox proportional hazards model assessing HRs across the quintiles of *iLGS* (Fig. 4c).

We replicated the association of the top and bottom *iLGS* quintiles with three age categories (age at last contact: ≥ 95, [90, 94], and < 90) in two independent cohorts, which include the Scripps Wellderly cohort[44] and the Medical Genome Reference Bank (MRGB) cohorts, two independent cohorts specifically ascertained to comprise a healthy aging population. However, since the absolute "age at death" for participants of these two cohorts are not available, we formulated the null hypothesis (H_0) as no association between the top (Q_{iLGS}^5) and bottom (Q_{iLGS}^1) quintiles with age category at the last point of contact, and tested for evidence against H_0 using a Fisher's Exact test. In the Wellderly cohort, the top quintile (Q_{iLGS}^5) was significantly associated with both the ≥ 95 and [90, 94] age categories (Fig. 5a). Individuals in Q_{iLGS}^5 are 3.67 times more likely to surpass 95 years of age (95% CI [1.68, 7.95], P = 6.68E-4) and 2.49 times more likely to fall into the [90, 94] age bin (95% CI [1.70, 3.73], P = 6.89E-6) than the rest of the *iLGS* strata. Here, the definitive lifespan is not known, and "age at last contact" is used as an approximate proxy for lifespan. As such, age categories further away from the two oldest age bins (i.e., ≥ 95 and [90, 94]) are necessarily comprised of individuals who have already achieved their maximum lifespan and individuals who will live longer and proceed to higher age bins. Therefore, the odds of individuals with favorable polygenic background (i.e., Q_{iLGS}^5) is expected to be attenuated in the lower age bins and continue to increase as we move to upper age-bins. This is consistent with the observed 1.47-fold increase in the odds of association between the Q_{iLGS}^5 and the ≥ 95 age category (compared to the [90, 94] age bin). Since individuals of the Wellderly cohort are essentially devoid of age-related conditions, the lack of association between the bottom quintile *iLGS* (Q_{iLGS}^1) and age categories is not surprising (Fig. 5a). In the MRGB cohort, Q_{iLGS}^5 is significantly associated with both the ≥ 95 and [90, 94] age bins (Fig. 5b). Here, individuals in Q_{iLGS}^5 are 4.27 times more likely to surpass 95 years of age (95% CI [3.43, 5.33], P = 7.88E − 47) and 3.66 times more likely (95% CI [2.96, 4.53], P = 2.37E − 30) to reach the [90, 94] age category than the rest of the population. Individuals with *iLGS* in the bottom quintile (Q_{iLGS}^1) were significantly depleted from the [90, 94] age bin (OR: 0.50, 95% CI [0.25, 0.90], P = 3.02E − 2). Conversely, Q_{iLGS}^1 was significantly enriched among individuals in the < 90 age bin (OR: 1.38, 95% CI [1.08, 1.76], P = 9.79E − 3).
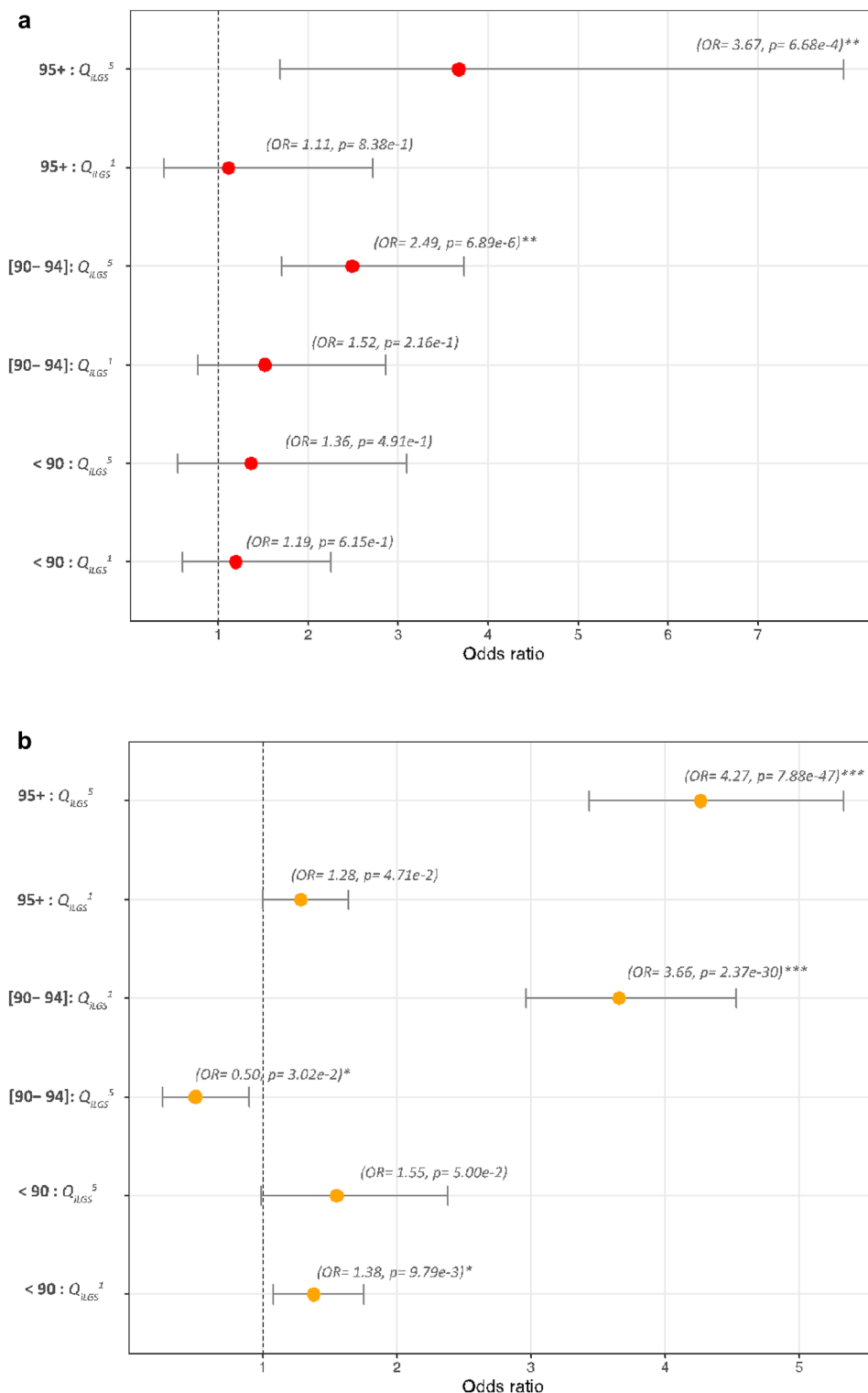
Fig. 5. Replication of *iLGS* association with age at last contact in the Wellderly and MRGB cohort. (**a**) Plot illustrating enrichment of top and bottom quintile iLGS scores (Q^5_{iLGS} & Q^1_{iLGS}) across three age bins (age at last contact: ≥ 95, [90–94] and < 90) of the Wellderly cohort. (asterisks identify significant associations). (**b**) Enrichment of top and bottom quintile iLGS scores (Q^5_{iLGS} & Q^1_{iLGS}) across three age bins (age at last contact: ≥ 95, [90–94] and < 90)) of the MRGB cohort. (asterisks identify significant associations).

In both replication cohorts, the strength of association between Q_{iLGS}^5 and the older age categories is significantly stronger. Given that age at death in these cohorts is undetermined, we speculate the stronger signal toward exceptional longevity arises from the more homogenous cohort construct among individuals achieving the upper limits of lifespan. Using the available clinical data in the MRGB cohort, we further investigated the association between Q_{iLGS}^5 and three age-related traits: extreme obesity (BMI > 40 kg/m^2), treatment history for high blood pressure and treatment history for high cholesterol. Despite the expected direction of effect across the three traits (Supplementary Fig. S11), the protective effect of Q_{iLGS}^5 is only statistically significant against treatment for high blood pressure (OR: 0.64, 05% CI [0.98, 0.42], $P = 3.82E - 2$).

Burden of rare variants across Q_{iLGS}^1 and Q_{iLGS}^5

Given the association of Q_{iLGS}^5 with an extended lifespan, we compared the burden of rare pathogenic variants (PrimateAI score ≥ 0.9 and alternative allele frequency < 1%) among the top versus bottom quintile *iLGS* carriers across centenarians and non-centenarians in our longevity cohort. A significant difference in the number of rare pathogenic variants between the Q_{iLGS}^5 and Q_{iLGS}^1 centenarians was observed (p < 0.03, Supplementary Fig. S12). It appears centenarians in the top quintile *iLGS* collectively carry a higher burden of rare pathogenic variants compared to centenarian carriers of Q_{iLGS}^1. This perhaps reflects the buffering effect of the favorable polygenic background that offsets the higher burden of rare pathogenic variants among Q_{iLGS}^5 carriers. This observation is consistent with a liability threshold model, where the accumulation of liabilities from pathogenic variants is favorably offset by the higher threshold rendered by the polygenic background[45]. To assess whether the observed significant difference (i.e., the p-value) in rare-variant burden between the two groups of centenarians was attributed to *iLGS* scores, we conducted a permutation test with 1000 iterations of randomization, in each of which a pseudo p-value was calculated to construct the null distribution. These pseudo p-values were calculated based on the difference in rare-variant burden between the top and bottom-quintile centenarians using shuffled *iLGS* scores. The permutation p-value (p = 0.03) suggested that the significant difference was indeed attributed to *iLGS* scores.

Proteomics correlates of *iLGS* and anti-aging drug repurposing

The quantitative size of *iLGS* is determined at birth and remains constant over an individual's life course. Given the association of *iLGS* with longevity, we searched for the proteomic correlates of the score among the 238 individuals (123 males and 115 females with available proteomic data) in our validation set. In doing so, we reasoned that proteomic predictors of aging and longevity could be detected via the patterns of association between the proteome and the *iLGS*. Briefly, we tested the association between 4,265 proteomic markers and *iLGS* among males and females separately and identified significant associations with FDR < 0.05 (See "Methods"). Among females, 37 proteins were positively associated with *iLGS*, while three—MAT2B, CLEC2B, and CMPK1—show negative association (Fig. 6a). Among these 40 proteins, 30 were associated with chronological age among female Ashkenazi Jews[46] (Supplementary Table S9), and ten were differentially expressed among centenarians with a consistent direction of effect[47] (Supplementary Table S9). The top proteins most significantly associated positively
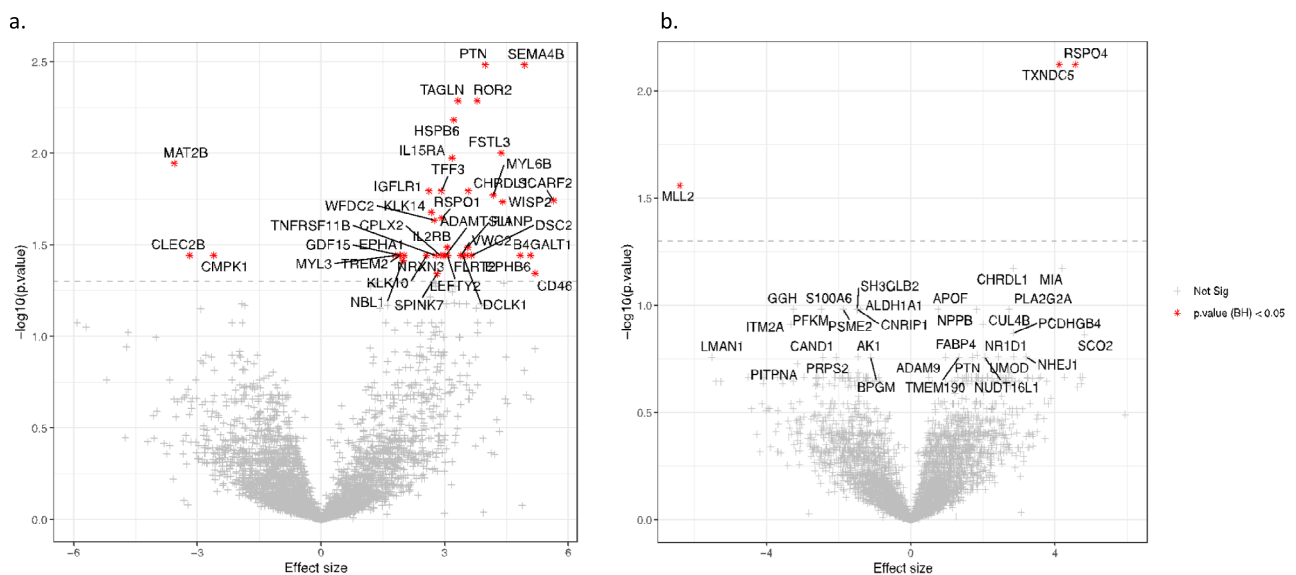

Fig. 6. Proteomic correlates of *iLGS*. Volcano plot showing the association of plasma proteins with *iLGS* among female (**a**) and male (**b**) subsets of the validation set with available proteomic data (n = 115 and 123 respectively). The x-axis denotes the "effect size", and the y-axis shows − log10 Benjamini–Hochberg corrected p-value of association. Proteins clustered to the right are positively correlated with *iLGS* and proteins clustered to the left are negatively associated with *iLGS*. Proteins surpassing the significance threshold at BH < 0.05 are identified by red asterisks.

with *iLGS* were semaphorin 4B (SEMA4B, $\beta = 4.94$, $P = 1.49E-6$) and Pleiotrophin (PTN, $\beta = 3.98$, $P = 1.49E-6$) (Fig. 6a). Previous studies have shown the association of pleiotrophin with both chronological age and extended lifespan[47–49] and identified the association between the expression of *SEMA4B* and favorable proteomic signature of intermittent fasting[50]. On the other hand, methionine adenosyltransferase 2B (MAT2B) showed the most significant negative association with *iLGS* ($\beta = -3.56$, $P = 2.05E-5$) (Fig. 6a). The connection of MAT2B to aging has not been studied, although its overexpression has been attributed to poor prognosis of triple-negative breast cancer[51] and critical illness in COVID-19[52].

Among male subjects ($n = 123$), only three proteins were significantly associated with *iLGS* (Fig. 6b). R-spondin 4 (RSPO4) showed the most significant positive association ($\beta = 4.82$, $P = 1.85E - 6$), consistent with the increased expression of *RSPO4* among centenarians as previously reported[47]. We detected a positive association for Thioredoxin domain-containing protein 5 (TXNDC5, $\beta = 4.25$, $P = 3.28E-6$). While the association of endoplasmic reticulum (ER) protein TXNDC5 with human longevity has not been previously reported, its transcriptional upregulation has been shown to be associated with aging in long-lived red sea urchins[53]. We also identified a negative association for lysine methyltransferase 2D (MLL2, $\beta = - 6.38$, $P = 4.19E - 5$). Among the 43 proteins significantly associated with *iLGS* in either males or females, the direction of effect was inconsistent between the sexes for only three proteins: MAT2B, CELEC2B, and CMPK1 (although not significant in males) (Supplementary Table S9). This inconsistency might indicate the sex differences in the proteomic signature of aging.

We queried the DrugBank database (version 5.1.8)[54] to identify drugs targeting or interacting with the 43 *iLGS*-associated proteins. We only considered drugs with ongoing clinical trial status (see "Methods"). Our rationale was that the underlying mechanism of action for drugs interacting with *iLGS*-associated proteins might implicate the pathways relevant to longevity and they could be repurposed as anti-aging treatments. We identified 25 unique drugs interacting with 11 proteins (Table 1). One noteworthy drug, fostamatinib, interacts with at least four longevity-promoting proteins. It is a selective inhibitor of spleen tyrosine kinase (Syk)[55] that acts upstream of the PI3K-Akt pathway, which is upregulated during aging and whose downregulation promotes improved cell survival in neurons[56]. The suppression of Syk attenuates the PI3K-Akt signaling, which inhibits

Target protein	Drug ID	Drug Name	Drug effect	Therapeutic indications	Clinical phase
PTN↑	DB08865	Crizotinib	ALK inhibitor	Non-small cell lung cancer (NSCLC)	Approved
	DB09063	Ceritinib	ALK antagonist	Non-small cell lung cancer (NSCLC)	Approved
	DB11363	Alectinib	ALK inhibitor	Non-small cell lung cancer (NSCLC)	Approved
	DB12267	Brigatinib	ALK inhibitor	Non-small cell lung cancer (NSCLC)	Approved
	DB12010	Fostamatinib	ALK inhibitor	Chronic immune thrombocytopenia (ITP)	Approved
	DB12130	Lorlatinib	ALK inhibitor	Non-small cell lung cancer (NSCLC)	Approved
TAGLN↑	DB11638	Artenimol	ligand	Plasmodium falciparum infection	Approved
MAT2B↓	DB00134	Methionine	Enzyme, substrate	Dietry supplementation	Approved
MYL6B↑	DB08378	4-[4-(2,5-DIOXO-PYRROLIDIN-1-YL)-PHENYLAMINO]-4-HYDROXY-BUTYRIC ACID	Unknown	–	Experimental
IL2RB↑	DB00041	Aldesleukin	Agonist/ Modulator	Renal cell carcinoma	Approved
	DB00074	Basiliximab	Chimeric antibody (IL-2 receptor antagonist)	Kidney transplant rejection	Approved
	DB00111	Daclizumab	Chimeric antibody (IL-2 receptor antagonist)	Relapsing forms of multiple sclerosis	Investigational
	DB00004	Denileukin diftitox	agonist	Cutaneous T-cell lymphoma	Approved
EPHB6↑	DB12010	Fostamatinib	ALK inhibitor	Chronic immune thrombocytopenia (ITP)	Approved
DCLK1↑	DB12010	Fostamatinib	ALK inhibitor	Chronic immune thrombocytopenia (ITP)	Approved
CMPK1↓	DB00441	Gemcitabine	Target, inhibitor	Ovarican cancer/NSCLC/Breast & lung cancer	Approved
	DB00709	Lamivudine	Enzyme, substrate	HIV and hepatitis B infections	Approved
	DB08934	Sofosbuvir	Enzyme, substrate	Hepatitis C infection	Approved
	DB01262	Decitabine	Enzyme, substrate	Myelodysplastic syndromes	Approved
EPHA1↑	DB12010	Fostamatinib	ALK inhibitor	Chronic immune thrombocytopenia (ITP)	Approved
B4GALT1↑	DB00141	N-Acetylglucosamine	Target	Osteoarthritis	Approved
	DB02696	6-Aminohexyl-uridine-C1,5'-diphosphate	Target	–	Experimental
	DB03013	N-acetyl-beta-D-glucosaminyl-(1->4)-N-acetyl-beta-D-glucosamine	Target	–	Experimental
	DB03501	Galactose-uridine-5'-diphosphate	Target	–	Experimental
	DB03685	Uridine monophosphate	Target	–	Experimental
	DB03814	2-(N-morpholino)ethanesulfonic acid	Target	–	Experimental
MLL2↓	DB00494	Entacapone	Unknown	Idiopathic Parkinson's Disease	Approved
	DB01752	S-adenosyl-L-homocysteine	Unknown	–	Experimental

Table 1. List of 25 unique drugs interacting with 11 protein correlates of iLGS.

mTOR and activates FOXO, leading to the activation of proapoptotic processes and cell cycle arrest. Although its precise mechanism of action is unclear, the active metabolite in fostamatinib has been recently found as a novel senolytic agent[57]. Given the extent of its interaction with longevity-promoting proteins, this drug merits further investigation as a potential anti-aging treatment.

Discussion

In this study, we developed a predictive genetic score for longevity (*iLGS*) based on GWAS summary statistics for a compendium of aging-related traits. We posited that the highly complex and polygenic architecture of human longevity is well approximated by a composite polygenic score that models the cumulative liabilities from ARTs. We assessed the performance of our model for distinguishing differential survival and showed, at the population level, *iLGS* predicts prolonged lifespan independent of sex or *APOE* haplotype status. We also established that the top quintile of *iLGS* is significantly enriched among individuals surpassing 95 years in the extant populations. Finally, we identified protein correlates of *iLGS* and proposed drugs that could be repurposed for potential anti-aging treatments.

In building our model, a few key assumptions were made. First, we assumed a highly polygenic architecture underlying human longevity. Our assumption of high polygenicity and genome-wide distribution of associated variants is supported by the proportional correlation of lifespan heritability with chromosomal length (Supplementary Fig. S13), which indicated that genomic loci with a non-neutral net effect on longevity are numerous and widely distributed across the genome. This is consistent with the overwhelmingly polygenic architecture inferred from a wide variety of complex traits[58,59]. Second, we posited that the deviation of an individual's lifespan from the population mean is primarily due to the additive liability conferred by common variants and can be well approximated by an additive model. Our assumption of "additivity of effects" aligns with the "omnigenic model" of complex traits[60], which posits that a significant portion of heritability is derived from common variants, each contributing a modest effect size[61]. Obviously, rare variants with sizable effects may still be relevant in the context of longevity, although such variants are primarily implicated for traits under a strong selective pressure[9,62]. Given the attenuated selective pressure during the aging[13,15], variants with a substantially detrimental effect on survival are expected to be under a more lenient purifying selection. Consequently, their frequency is not necessarily bound to lower limits, and therefore, rare variants with large effect sizes on aging and longevity are expected to be scarce. Third, we modeled lifespan as a function of a set of traits that are directly or indirectly under selection. Of note, lifespan is the ultimate trait at the top of the fitness hierarchy and its variance is determined by the interplay of other fitness components, which include fertility, fecundity, and inclusive fitness. We used polygenic scores of nested traits (as a reasonable proxy for unmeasured trait values) to model the perceived interplay and derive a single composite genetic score for longevity. Unlike other biological age predictors (such as epigenetic clocks, telomere length, transcriptomic, and metabolic predictors) that are dependent on the actual chronological age and may vary over time, *iLGS* is stable throughout life and can be assessed from birth. Thus, the score can be incorporated with other quantitative predictors of age later in life when they become available to give a more accurate prediction of biological age.

Our results showed that *iLGS* generalizes well in predicting lifespan in the extant populations when the age at death is not yet available. Furthermore, we showed that in the Wellderly and MRGB cohorts of healthy aging, the odds of individuals with top quintile *iLGS* (Q^5_{iLGS}) increased sharply in the ≥ 95 age bin. Although we identified a protective association between Q^5_{iLGS} and the treatment history for high blood pressure in the MRGB cohort, we did not observe a significant association between the *iLGS* and the age at onset (or post-intervention survival) for cardiovascular complications or three cancer types (breast, prostate, and skin) in a small portion (n = 390) of our in-house cohort with available data. A number of possibilities may explain this lack of association: (1) The age distribution in our test set is relatively narrow (95 ~ 111 years) (Supplementary Fig. S1), and therefore, our ability to discriminate individuals based on *iLGS* strata is reduced accordingly. In particular, these complications occur in such a narrow range after age 95 that the age at onset and the survival time are indistinguishable across *iLGS* strata; (2) Heterogeneity of the investigated conditions may play an important role that is underappreciated by our model. Specifically, the underlying cause of differences in the age at onset or survival time may not be completely polygenic in nature (hence not distinguishable by our model) and simply arise due to the acquired somatic mutations in the case of cancers and lifestyle differences in the case of cardiovascular complications. This is quite an interesting possibility since studies into the congestive heart failure diagnosis have also questioned the validity of prognostic markers among centenarians[63]; and (3) Given the limited sample size, our study was underpowered.

The specific pattern of interaction among individual traits in our model provides new insights into the approximate relationship between different traits and longevity. We observed that a great majority of these interactions are mediated through the insulin-like growth factor 1 (IGF1). This recapitulates the well-established significance of this hormone in human aging and underscores the relevance of interventions targeting IGF1-extended signaling network for longevity[64]. Consistent with this, we highlight the repurposing potential of fostamatinib as a senolytic drug candidate. It interacts with four *iLGS*-associated proteins (PTN, EPHA1, EPHB6, and DCLK1) and attenuates the PI3K-Akt signaling, which leads to a range of favorable outcomes that collectively promote longevity[64]. Further, we highlight the significance of our finding for proteins associated with *iLGS*. Unlike earlier studies into the proteomic correlates of longevity, where it is not clear whether differential protein levels are truly causal or simply fluctuate due to aging, our results are not confounded by the chronological age.

Nevertheless, our results should be considered in light of a few limitations. First, in constructing our model, we used GWAS summary statistics from individuals of primarily European ancestry. Polygenic scores are sensitive to allele frequency differences and patterns of LD that vary with ancestry[65,66]. Thus, our model will likely underperform in populations of non-European descent. Second, it remains to be determined whether the association

of *iLGS* with lifespan is independent of lifestyle factors. Although we showed that the association of *iLGS* with longevity is independent of gender and *APOE* status, using the available data, we could not rule out the confounding effect of lifestyle choices. If the effect of *iLGS* on longevity is mediated through a confounder such as dietary pattern or medication intake, we expect that protein correlates of *iLGS* to be also confounded. However, we could not test for this confounding in a replication cohort with available lifestyle data like the UK Biobank dataset since summary statistics for constructing *iLGS* were primarily obtained from UK Biobank GWAS studies.

Taken together, we showed how pervasive pleiotropy could be leveraged in constructing a composite polygenic score for longevity. The application of *iLGS* to the extant populations offers an opportunity for better risk management among individuals at the lower end of the score distribution. These individuals are more likely to develop age-related pathologies earlier in life and, therefore, benefit from lifestyle modifications. Moving forward, in expectation of more powerful GWAS studies in diverse populations, we believe our method offers a promising framework for stratifying life expectancy in the extant population.

Methods

Ethics statement

In this study, we used datasets from the UK Biobank (application number 58069), as approved by the UK Biobank board, the Wellderly cohort (the Scripps Institute, La Jolla, USA) obtained from the European Genome-Phenome Archive (EGAS00001002306), the Medical Genome Reference Bank cohort obtained from the Garvan Institute of Medical Research, Australia, and the Genetic Epidemiology Research on Adult and Aging (GERA) obtained through dbGaP (request number 82588-1). The study and data access were approved by Albert Einstein Institutional Review Board, protocols 2019–9922 (reference number: 05294). All methods were carried out in accordance with the relevant guidelines and regulations, and informed consent was obtained from the research participants.

Data sets

Einstein LonGenity cohort
We performed our analysis on a sub-cohort of 1,740 (722 males and 1018 females) offspring of Ashkenazi Jewish centenarians and controls recruited longitudinally in the LonGenity project at the Albert Einstein College of Medicine[67]. These individuals were genotyped on a custom array at 635,623 SNPs and subsequently underwent whole-exome sequencing (WES). Details pertaining to DNA sample preparation and sequencing are explained elsewhere[68].

Wellderly cohort
For replication, we used the whole-genome sequence data from 510 healthy elderly individuals (aged between 80 and 105 years) without any chronic conditions or medication history recruited through the Scripps Institute Wellderly study. Details pertaining to the cohort demography and the genome sequencing procedure are explained in an earlier study[44].

MRGB cohort
For replication, we also used the whole-genome sequence data from 2,570 individuals (1,251 males & 1,319 females) in the Medical Genome Reference Bank (MRGB) cohort. This cohort comprises Australians who lived at least 70 years of age without any history of cancer, dementia, and cardiovascular disorders at baseline entry or study follow-ups. Gender-stratified age distributions across participants are provided in Supplementary Fig. S14. The cohort demography and genome sequencing details are explained elsewhere[69].

GERA cohort
We used a GERA (Genetic Epidemiology Research on Aging) sub-cohort of 62,268 individuals with European ancestry to convert *iLGS* to SNP weights. We used pre-computed principal component scores released by the GERA consortium to filter out 16,151 individuals with non-European ancestry.

Quality control (QC)

We applied standard GWAS QC measures[70] implemented in PLINK v1.90b[71]. Briefly, we removed individuals with a mismatch between the reported and inferred gender from the sex chromosomes. We restricted our analysis to unrelated individuals for all cohorts by excluding the younger participants from each pair of close relationships (up to second-degree). Our QC measures at the sample level also included removing individuals with < 95% call rate, individuals with extensive runs of homozygosity, or non-European ancestry as determined by the PCA analysis. The Einstein LonGeity cohort analysis exclusively focused on the Ashkenazi Jewish cluster. At the variant level, our QC measure involved removing SNPs with < 1% minor allele frequency (MAF), SNPs with < 95% call rate among all individuals, or an extreme deviation from the Hardy–Weinberg equilibrium (HWE) at $P \leq$ 1E-06, which most likely represents poor genotyping.

We imputed the QC'ed genotypes across filtered individuals using the Michigan Imputation server[72] and 1000 Genomes phase 3 haplotypes as the reference panel. For the UK Biobank sub-cohort, we directly obtained version 3 of the imputed genomes from the consortium. Post-imputation QC measures across all cohorts included removing SNPs with IMPUTE2 info score < 0.5, MAF < 1%, or an extreme deviation from HWE at $P \leq$ 1E−06. For the Einstein LonGenity cohort, we merged the post-imputation genotypes with the exome data while correcting for the strand orientation. Prior to merging, genotypes from the exome data were restricted to bi-allelic loci only,

and INDELs were removed. Wherever the imputed genotype was different from the exome data, we replaced the inferred alleles with the observed allele from the whole-exome sequencing (WES) data.

Principal component analysis (PCA)

To avoid stratification bias, we restricted our analysis to a single ancestry (ASJ for the Einstein LonGenity cohort and European ancestry for the replication cohorts). We used pre-imputed genotypes to calculate principal components in KING v2.2.5[73]. Before running KING, we applied a more stringent QC measure using PLINK (v.1.90b) to remove SNPs and individuals with < 99% call rate. Samples were pruned to include founders only. Additionally, SNPs with MAF < 5% were removed and pruned to include uncorrelated SNPs at pairwise $r^2 < 0.2$. The top three PCAs were projected to that of 1000 Genomes phase 3 data, and samples with European ancestry were selected for downstream analysis.

Sliding regression framework

Unrelated ASJ samples from the Einstein LonGenity cohort ($n = 957$) were divided into 48 overlapping bins according to their reported age at death. Bins were constructed to contain 30 individuals and an overlapping size of 15 individuals between the neighboring windows. The choice of parameters (window size and overlap size) selected and robustness of model was confirmed by carrying out simulation analysis across twelve different configurations in simulated model (window size (i.e. $N_{w.s} \in \{20, 30, 40, 50\}$) and the overlap size (i.e. Noverlap $\in \{5, 10, 15\}$). The model consists of 100 genomic loci across 1000 individuals with age-at-death distribution resembling our discovery cohort. The allelic frequency across 8,090,428 imputed SNPs was calculated in each bin. Details about the age distribution of participants in each bin are provided in the Supplementary Table S10. To remove the sex effect in the ultimate model and correct for the different male/female ratios across bins, calculated allele frequencies were regressed on to the number of females in each bin and residuals were then used in the final regression model so that:

$$\gamma = \alpha + \beta_j e_j$$

where γ is the age at death, β_j is the slope of regression, e_j is the sex-corrected frequency from the prior regression, and α is the intercept of regression defined by the expected mean age at death in our discovery cohort (~70 years). We tested for an effect of gender-corrected allele frequency (e_j) by a two-tailed t-test where we constructed our test hypothesis as below:

$$\{H_0 : (\beta_j = 0) H_1 : (\beta_j \neq 0)$$

We interpreted positive β_j and negative β_j as pro-longevity and pro-aging, respectively. For each β_j, we calculated the standard error using the formula below:

$$SE_{\beta_j} = \frac{\sqrt{\sum (\gamma - \widehat{\gamma})^2 / n - 2}}{\sqrt{\sum (e_j - \overline{e}_j)^2}}$$

where, γ is the age at death, $\widehat{\gamma}$ is the estimated age at death, n is the total number of bins, e_j is the frequency residual and $\overline{e}_j$ is the mean of frequency residuals. Variants surpassing the genome-wide significant threshold ($P < 5E-8$) were selected for replication analysis in the UK Biobank dataset and downstream analysis.

The number of variants in all age bins was consistent across all autosomes.

Functional enrichment analysis

To identify functionally relevant SNPs, we used FUMA[74] to analyze the summary statistics from the sliding regression analysis. Variant coordinates in all analyses were defined according to the Hg19/GRCh37 assembly. We identified significant independent SNPs as variants surpassing the genome-wide significant threshold ($P < 5E-08$) and are independent of each other at $r^2 < 0.1$. We defined LD blocks surrounding the lead SNP according to the whole-genome LD maps of the European population[75] and included all known SNPs in the LD block (regardless of being included in the sliding regression) that are in LD with the lead SNP at $r^2 \geq 0.5$ for eQTL and chromatin interaction mapping. The defined risk loci, therefore, may contain SNPs that were not available in the summary statistic input from the sliding regression but are linked to the significant independent SNPs according to the 1000G reference panel. For each risk locus, we only retained SNPs with MAF ≥ 0.1 in the Ashkenazi Jewish population. Additionally, where the distance between two risk loci in neighboring LD blocks was less than 250 kb, the neighboring LD blocks were merged to a single larger risk locus. We mapped SNPs to the nearest gene within a 10 kb window for positional mapping. For eQTL and Chromatin interaction mapping, annotation was carried out over the entire risk region and genes within the risk locus, and those located outside that are linked to the SNPs in the genomic risk locus were prioritized.

Phenome wide association analysis (PheWAS)

We tested the association of the shortlisted variants with 1,504 well-powered traits and disease endpoints ascertained primarily from the UK Biobank[76]. Each SNP was interrogated for association independently, and traits surpassing the Bonferroni corrected threshold for multiple testing were reported as significantly associated. To test whether the extent of observed pleiotropy is higher than expected in specific trait domains, we defined the

null distribution as an equal number of positive signals across all domains (under the null hypothesis of no differences in domain-specific pleiotropy). We used a one-sided Fisher's exact test to compute odd ratios.

Polygenic risk score (PRS) calculation

We calculated the additive effect of common variants on the genetic liability of traits associated with candidate variants. We used the clumping and thresholding strategy implemented in PRSice software. We computed polygenic scores for each trait independently using the formula below:

$$PRS_p = \sum_j^M \beta_j \times \psi; \, \psi \in \{0, 1, 2\},$$

where β_j is the effect size (or log odds ratio (OR) for binary traits) of the allele j, ψ is the allele dosage for SNP j, and M is the total number of SNPs after pruning.

We followed the best practice recommendation for PRS calculation[77]. Briefly, we calculated the chip heritability (h^2_{SNP}) across each independent trait using the LD score regression[78] and confirmed $h^2_{SNP} > 0.05$. Designations of the effect allele across all summary statistics were harmonized to match those of the UK Biobank. We used PRSice-2[79] to calculate PRSs. We implemented the standard "clumping + thresholding" method to remove highly correlated SNPs ($r^2 > 0.1$) and retain only the most significant associations. Since the optimal SNP association P-value for inclusion in the PRS calculation is unknown a priori, we generated PRS across a range of p-value thresholds and selected the threshold providing the highest Nagelkerke R^2.

We calculated the PRS correlation (r_{PRS}) as Pearson's correlation of standardized PRS across the 87 traits. As expected, PRS scores for traits within the same domain and related biomarkers showed a higher correlation.

Derivation of composite polygenic scores

We modeled the age at death using the scaled PRS scores across the 87 traits (53 traits from PheWAS analysis and 34 blood and urine biomarker traits). All PRSs were standardized to the unit scale (i.e., zero mean and unit standard deviation) across the entire dataset. We split our cohort into derivation ($n = 715$) and validation set ($n = 242$). To increase the statistical power of the model for predicting longevity, the derivation set was enriched for samples with age at death over 95 years.

We used a stacked model with five-fold cross-validation to predict age at death based on traits' PRSs. From the 87 distinct traits and endophenotypes, we first fitted an Elastic-net regression to remove non-informative PRSs. Next, the remaining PRSs were used in a polynomial regression to derive the informative interactions between the PRS scores, and finally, all interaction terms and individual PRS scores were plugged into the final elastic-net regression to remove the non-informative interaction terms. In all penalized regression steps, we tested a range of penalties to decide the optimum penalty threshold and selected the best-performing model in terms of cross-validated AUC. We derived the iLGS from the formula below:

$$iLGS = \sum_{i=1}^m \beta_i PRS_i + \sum_{i,j=1}^n \beta_{ij}(PRS_i \times PRS_j),$$

where β_i is the coefficient of PRS for trait i in the ultimate model, and β_{ij} is the non-zero coefficients for the interaction terms between PRSs i and j drawn from $c\,(n, 2)$, in which $n = 77$ is the total number of PRSs with non-zero coefficient from the primary elastic-net regression (Fig. 2).

Derivation of *iLGS* SNP weights

We used a subsample of 62,268 individuals with European ancestry from the GERA cohort to derive the SNP weights. We imputed genotypes across autosomes using the Michigan Imputation Server[72] and 1000 Genomes phase 3 haplotypes as the reference panel. We excluded SNPs with IMPUTE2 info score < 0.5 and carried out standard QC on the remaining SNPs. To enhance the calibration of statistics derived from our generalized linear model, we removed loci with minor allele counts less than 20 in PLINK (version 2.0). We calculated kinship coefficients (φ_k) in KING (version 2.2.5)[73] and removed related individuals up to the second degree, which is defined as $\varphi_k > 0.088$. Imputed and QC'ed genotypes from the remaining 59,534 individuals were used in a generalized linear regression formwork to derive *iLGS* SNP weights according to the formula below:

$$iLGS = G_j\beta_j + (PC_1 + \cdots + PC_3 + sex + B.C + (sex \times B.C))\beta' + \varepsilon$$

where $iLGS$ is the vector of scores, G_j is the dosage matrix for variant j, $(PC_1 + \cdots + PC_3 + sex + B.C + (sex \times B.C))$ is the fixed-covariate matrix correcting for the top 10 principal components, sex, birth cohort (B.C), and the interaction of sex with birth cohort and ε is the residual error subject to least-square minimization.

Survival regression analysis

We evaluated the utility of *iLGS* in distinguishing differential longevity in the validation set (n = 242). We used the Lifelines package[80] (version 0.25.11) to fit a non-parametric Cox's proportional hazard model according to the formula below:

$$h(x) = h_0(t)exp(iLGS.\beta_1 + sex.\beta_2 + (iLGS \times sex).\beta_3$$

where, $h_0(t)$ is the baseline hazard estimated using Breslow's method. To increase the stability of estimates we included a penalizer term of 0.0005 and L_1-ratio of 1.0 to shrink the magnitude of $\widehat{\beta_i}$ according to:

$$0.0005 \times \left(\frac{1 - L_1 ratio}{2} ||\beta||_2^2 + L_1 ratio ||\beta||_1 \right).$$

The significance of individual regression coefficients was investigated using the likelihood ratio test. The absence of deviation from proportional hazard assumptions was validated using the global test for Schoenfeld residuals. We also calculated the cumulative risk of death (as a function of age) across the five quintiles of iLGS based on the equation below:

$$1 - \hat{\eta}(t) = 1 - \exp\left[\hat{h}_0(t) \cdot \exp\left(iLGS^{Q*} \cdot \beta_1 + sex \cdot \beta_2 \right) \right],$$

where $\widehat{\eta(t)}$ is the cumulative survival at the given time (t), $\widehat{h_{0(t)}}$ is the estimated baseline hazard at time t and $iLGS^{Q*}$ is the quintile of the iLGS score.

Evaluation of the *iLGS*
We estimated survival curves across different quintiles of iLGS using the Kaplan–Meier estimator:

$$\widehat{S(t)} = \prod\nolimits_{i:t_i < t} \frac{\delta_i - \theta_i}{\delta_i},$$

where, θ_i is the number of death events at a particular time t_i, and δ_i is the fraction of the population at risk of death prior to time t_i. We investigated whether there is a significant difference in survival probability across quintiles of iLGS using the log-rank test.

We tested the performance of iLGS for distinguishing the "age at diagnosis" of three late-life cardiovascular pathologies (i.e., angina, arrhythmia, and interventional cardiac procedures) and three types of cancers (i.e., breast, prostate, and skin cancer) across 242 Ashkenazi Jewish individuals with available clinical and genotypic data. Kaplan–Meier survival curves across the top and bottom quintile of iLGS were compared using the log-rank test. We calculated the survival time post-diagnosis for each trait as the difference between the "age at death" and the "age at diagnosis". In addition, we investigated the association of dummy *iLGS* quintiles with post-diagnosis medical intervention survival time using an ordinary least square regression while controlling for the gender effect.

Validating iLGS performance in additional cohorts
We investigated the performance of iLGS to distinguish survival in two additional cohorts. We used whole-genome sequences from 2570 participants (male 1251; female 1319) of the Medical Genome Reference Bank (MGRB) cohort[69]. This cohort comprises healthy elderly individuals depleted for aging-related disorders and cancers. We derived the age of participants based on their respective dates of birth. At the time of analysis, the age of participants ranged from 75 to 102 years old (mean = 86; SD = 5.13). We corroborated participants' reported sexes with X-chromosome heterozygosity estimates. Prior to the PRS calculation, we removed variant calls with read depth (DP) $\leq$ 10 or FisherStrand bias (FS > 60). Furthermore, we excluded INDELs and multi-allelic variants and subjected the data to the standard QC procedure as explained above. This resulted in 8,496,911 autosomal SNPs being used to compute iLGS. We also replicated the performance of iLGS across the 510 participants (194 males and 316 females) of the Wellderly cohort[44] (Supplementary Table S10). Since these samples were sequenced using the Complete Genomics platform (Complete Genomics), we initially converted the "transvariants" calls to the conventional VCF format. We subsequently removed indels, multi-allelic SNPs, and variants with low read-depth DP $\leq$ 10 or variants flagged as "VQLOW". Upon standard QC, 3,383,869 autosomal SNPs were used for *iLGS* calculation. Since age at death in these two cohorts was undetermined, we first binned participants of each cohort into three age categories: < 90, [90, 94], and $\geq$ 95. Then, we applied a two-sided Fisher's exact test to investigate the association of top (Q_{iLGS}^5) and bottom quintile *iLGS* (Q_{iLGS}^1) with age categories in each cohort independently.

Proteomic analysis across quintiles
The LonGenity study used the 5 k SomaScan Assay V4, which includes 5,209 SOMAmer reagents targeting human proteins. SomaScan data standardization was carried out by SomaLogic and has been described in detail before[46]. In summary, the study measured the relative concentration of 5,209 human-specific SOMAscan aptamers and 75 non-human and control aptamers in 1,027 participants of the Einstein LonGenity cohort. Normalized aptamer concentrations, measured in relative fluorescent units (RFU), were used to eliminate proteins and individuals with significant variation across array runs, as described by Candia et al.[81]. After conducting quality checks, 960 sequences that failed the checks were removed. After excluding non-human proteins, deprecated markers, and "*noncleavable*", "*nonbiotin*", and spurious markers, 4,265 SOMAmer reagents remained available for proteomic analysis.

We measured the association between the protein levels and the *iLGS* using a robust regression with MM-estimator in the validation set (with available proteomic data) for each gender group separately (123 males and 115 females). To enhance the reliability of the regression in the presence of anomalous protein-levels, we used Cook's distance to remove outliers with a residual value > 4/n. We carried out a gender-stratified comparison between the Q_{iLGS}^5(top quintile *iLGS*) versus Q_{iLGS}^1 (bottom quintile *iLGS*) across individuals in the $\geq$ 95 age bin and applied the empirical Bayes moderated t-test using the Limma package[82] to compare normalized RFU intensities. All *P*-values were corrected for multiple hypothesis testing using the Benjamini–Hochberg method (FDR < 0.05). Significant proteins were utilized to search for potential drug candidates for repurposing in the

DrugBank database (version 5.1.8)[54]. We restricted our analysis to drugs with clinical phase assignment as "experimental", "investigational", or "approved" status, corresponding to clinical trial phase I, II, and III, respectively.

Data availability

Summary statistics for the longevity association of analyzed variants in our Ashkenazi Jewish longevity cohort and SNP weights for polygenic scores derived in this study will be available upon request. Due to privacy concerns for our research participants, individual-level genetic data from the Einstein LonGenity study are not publicly available; however, anonymized data will be shared upon request from a qualified academic investigator, providing the data transfer is approved by the Institutional Review Board and regulated by a material transfer agreement. The whole genome data from the Medical Genome Reference Bank and the Wellderly cohort is accessible through the European genome-phenome archive under study IDs EGAS00001003511 and EGAD00001003941, respectively. The genotyping array data for the GERA cohort (Resource for Genetic Epidemiology Research on Aging) is available through dbGaP Study Accession: phs000674.v2.p2. We used genotypes and summary statistics from publicly available databases to estimate linkage disequilibrium correlations and construct polygenic scores. We used the genetic and clinical data from the UK Biobank under research project ID 58,069 (https://www.ukbiobank.ac.uk/enable-your-research/approved-research/evolutionary-analysis-of-ageing-and-age-related-disorders-in-the-extant-human-population). The various summary statistics used to calculate genetic correlations are available from GeneAtlas (http://geneatlas.roslin.ed.ac.uk/), NealeLab (http://www.nealelab.is/uk-biobank) and GWASAtlas (https://atlas.ctglab.nl). We used DrugBank v.5.1.8 to identify drugs with repurposing potential as gerotherapeutics which are available at (https://go.drugbank.com/releases/5-1-8). Source data for figures in this study are available in the supplementary documents and upon request from the corresponding author.

Code availability

The code used for the generation of the Integrated Longevity Genetic Scores (*iLGS*) model and the accompanying sensitivity analysis is available on GitHub at: https://github.com/RezaJF/iLGS.

References

1. Kaplanis, J. *et al.* Quantitative analysis of population-scale family trees with millions of relatives. *Science* **360**, 171–175 (2018).
2. Kerber, R. A., O'Brien, E., Smith, K. R. & Cawthon, R. M. Familial excess longevity in Utah genealogies. *J. Gerontol. Ser. A Biol. Sci. Méd. Sci.* **56**, B130–B139 (2001).
3. Ruby, J. G. *et al.* Estimates of the heritability of human longevity are substantially inflated due to assortative mating. *Genetics* **210**, 1109–1124 (2018).
4. Fisher, R. A. XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1919).
5. Deelen, J. *et al.* A meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nat. Commun.* **10**, 3669 (2019).
6. Timmers, P. R. H. J., Wilson, J. F., Joshi, P. K. & Deelen, J. Multivariate genomic scan implicates novel loci and haem metabolism in human ageing. *Nat. Commun.* **11**, 3570 (2020).
7. Timmers, P. R. *et al.* Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *eLife* **8**, e39856 (2019).
8. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).
9. Sella, G. & Barton, N. H. Thinking about the evolution of complex traits in the era of genome-wide association studies. *Annu. Rev. Genom. Hum. Genet.* **20**, 1–33 (2015).
10. Jabalameli, M. R. & Zhang, Z. D. Unravelling genetic components of longevity. *Nat. Aging* **2**, 5–6 (2022).
11. Zhang, Z. D. *et al.* Genetics of extreme human longevity to guide drug discovery for healthy ageing. *Nat. Metab.* **2**, 663–672 (2020).
12. Pengelly, R. J., Vergara-Lope, A., Alyousfi, D., Jabalameli, M. R. & Collins, A. Understanding the disease genome: Gene essentiality and the interplay of selection, recombination and mutation. *Brief. Bioinform.* **20**, 267–273 (2019).
13. Kirkwood, T. B. L. Evolution of ageing. *Nature* **270**, 301–304 (1977).
14. Kirkwood, T. B. L. & Rose, M. R. Evolution of senescence: Late survival sacrificed for reproduction. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* **332**, 15–24 (1991).
15. Williams, G. C. Pleiotropy, natural selection, and the evolution of senescence. *Evolution* **11**, 398 (1957).
16. Kirkwood, T. B. L. Understanding the odd science of aging. *Cell* **120**, 437–447 (2005).
17. Mostafavi, H. *et al.* Identifying genetic variants that affect viability in large cohorts. *PLoS Biol.* **15**, e2002458 (2017).
18. Bergman, A., Atzmon, G., Ye, K., MacCarthy, T. & Barzilai, N. Buffering mechanisms in aging: A systems approach toward uncovering the genetic component of aging. *PLoS Comput. Biol.* **3**, e170 (2007).
19. Zhang, Q. *et al.* Nuclear speckle specific hnRNP D-like prevents age- and AD-related cognitive decline by modulating RNA splicing. *Mol. Neurodegener.* **16**, 66 (2021).
20. Langsted, A., Kamstrup, P. R. & Nordestgaard, B. G. High lipoprotein(a) and high risk of mortality. *Eur. Hear. J.* **40**, 2760–2770 (2019).
21. Arsenault, B. J. *et al.* Association of long-term exposure to elevated lipoprotein(a) levels with parental life span, chronic disease-free survival, and mortality risk. *JAMA Netw. Open* **3**, e200129 (2020).
22. Zewinger, S. *et al.* Relations between lipoprotein(a) concentrations, LPA genetic variants, and the risk of mortality in patients with established coronary heart disease: A molecular and genetic association study. *Lancet Diabetes Endocrinol.* **5**, 534–543 (2017).
23. Jukarainen, S. *et al.* Genetic risk factors have a substantial impact on healthy life years. *Nat. Med.* **28**, 1893–1901 (2022).
24. Guilluy, C. *et al.* The Rho exchange factor Arhgef1 mediates the effects of angiotensin II on vascular tone and blood pressure. *Nat. Med.* **16**, 183–190 (2010).
25. Gabin, J. M., Tambs, K., Saltvedt, I., Sund, E. & Holmen, J. Association between blood pressure and Alzheimer disease measured up to 27 years prior to diagnosis: The HUNT study. *Alzheimer's Res. Ther.* **9**, 37 (2017).
26. Ding, J. *et al.* Antihypertensive medications and risk for incident dementia and Alzheimer's disease: A meta-analysis of individual participant data from prospective cohort studies. *Lancet Neurol.* **19**, 61–70 (2020).
27. Kleefeldt, F. *et al.* Aging-related carcinoembryonic antigen-related cell adhesion molecule 1 signaling promotes vascular dysfunction. *Aging Cell* **18**, e13025 (2019).

28. Sun, A. H., Collette, J. R. & Sifers, R. N. The cytoplasmic tail of human mannosidase Man1b1 contributes to catalysis-independent quality control of misfolded alpha1-antitrypsin. *Proc. Natl. Acad. Sci.* **117**, 24825–24836 (2020).
29. Eriksson, S., Janciauskiene, S. & Lannfelt, L. Alpha 1-antichymotrypsin regulates Alzheimer beta-amyloid peptide fibril formation. *Proc. Natl. Acad. Sci.* **92**, 2313–2317 (1995).
30. Gold, M. *et al.* α 1-antitrypsin modulates microglial-mediated neuroinflammation and protects microglial cells from amyloid-β-induced toxicity. *J. Neuroinflamm.* **11**, 165 (2014).
31. Rodríguez-Comas, J. *et al.* Alpha1-antitrypsin ameliorates islet amyloid-induced glucose intolerance and β-cell dysfunction. *Mol. Metab.* **37**, 100984 (2020).
32. Li, X., Wang, X. & Snyder, M. P. Metformin affects heme function as a possible mechanism of action. *G3 Genes Genomes Genet.* **9**, 513–522 (2019).
33. Dharshini, S. A. P., Taguchi, Y.-H. & Gromiha, M. M. Investigating the energy crisis in Alzheimer disease using transcriptome study. *Sci. Rep.* **9**, 18509 (2019).
34. Dong, M.-Q. *et al.* Quantitative mass spectrometry identifies insulin signaling targets in *C. elegans*. *Science* **317**, 660–663 (2007).
35. Johmura, Y. *et al.* Senolysis by glutaminolysis inhibition ameliorates various age-associated disorders. *Science* **371**, 265–270 (2021).
36. Marengoni, A. *et al.* Aging with multimorbidity: A systematic review of the literature. *Ageing Res. Rev.* **10**, 430–439 (2011).
37. Amell, A. *et al.* Disease networks identify specific conditions and pleiotropy influencing multimorbidity in the general population. *Sci. Rep.* **8**, 15970 (2018).
38. Johnson, S. C., Dong, X., Vijg, J. & Suh, Y. Genetic evidence for common pathways in human age-related diseases. *Aging Cell* **14**, 809–817 (2015).
39. Keightley, P. D. & Hill, W. G. Variation maintained in quantitative traits with mutation–selection balance: Pleiotropic side-effects on fitness traits. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **242**, 95–100 (1990).
40. van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H. & Wray, N. R. Genetic correlations of polygenic disease traits: From theory to practice. *Nat. Rev. Genet.* **20**, 567–581 (2019).
41. Deluty, J. A., Atzmon, G., Crandall, J., Barzilai, N. & Milman, S. The influence of gender on inheritance of exceptional longevity. *Aging (Albany NY)* **7**, 412–418 (2015).
42. Barzilai, N., Atzmon, G., Derby, C. A., Bauman, J. M. & Lipton, R. B. A genotype of exceptional longevity is associated with preservation of cognitive function. *Neurology* **67**, 2170–2175 (2006).
43. Sebastiani, P. *et al.* APOE alleles and extreme human longevity. *J. Gerontol. Ser. A* **74**, 44–51 (2019).
44. Erikson, G. A. *et al.* Whole-genome sequencing of a healthy aging cohort. *Cell* **165**, 1002–1011 (2016).
45. Fahed, A. C. *et al.* Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat. Commun.* **11**, 3635 (2020).
46. Sathyan, S. *et al.* Plasma proteomic profile of age, health span, and all-cause mortality in older adults. *Aging Cell* **19**, e13250 (2020).
47. Sebastiani, P. *et al.* Protein signatures of centenarians and their offspring suggest centenarians age slower than other humans. *Aging Cell* **20**, e13290 (2021).
48. Menni, C. *et al.* Circulating proteomic signatures of chronological age. *J. Gerontol. Ser. A Biomed. Sci. Méd. Sci.* **70**, 809–816 (2015).
49. Lehallier, B. *et al.* Undulating changes in human plasma proteome profiles across the lifespan. *Nat. Med.* **25**, 1843–1850 (2019).
50. Mindikoglu, A. L. *et al.* Intermittent fasting from dawn to sunset for 30 consecutive days is associated with anticancer proteomic signature and upregulates key regulatory proteins of glucose and lipid metabolism, circadian clock, DNA repair, cytoskeleton remodeling, immune system and cognitive function in healthy subjects. *J. Proteom.* **217**, 103645 (2020).
51. Xu, J., Wu, D., Wang, S. & Wang, Z. MAT2B expression correlates with poor prognosis in triple-negative breast cancer. *Cancer Manag. Res.* **11**, 5501–5511 (2019).
52. Pairo-Castineira, E. *et al.* Genetic mechanisms of critical illness in COVID-19. *Nature* **591**, 92–98 (2021).
53. Polinski, J. M., Kron, N., Smith, D. R. & Bodnar, A. G. Unique age-related transcriptional signature in the nervous system of the long-lived red sea urchin Mesocentrotus franciscanus. *Sci. Rep.* **10**, 9182 (2020).
54. Wishart, D. S. *et al.* DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
55. Braselmann, S. *et al.* R406, an orally available spleen tyrosine kinase inhibitor blocks Fc receptor signaling and reduces immune complex-mediated inflammation. *J. Pharmacol. Exp. Ther.* **319**, 998–1008 (2006).
56. Chen, Y. *et al.* Aging-induced Akt activation involves in aging-related pathologies and Aβ-induced toxicity. *Aging Cell* **18**, e12989 (2019).
57. Cho, H.-J. *et al.* Identification of SYK inhibitor, R406 as a novel senolytic agent. *Aging (Albany NY)* **12**, 8221–8240 (2020).
58. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
59. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the genetic architecture of 30 complex traits from summary association data. *Am. J. Hum. Genet.* **99**, 139–153 (2016).
60. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: From polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
61. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
62. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat. Genet.* **46**, 220–224 (2014).
63. Rasmussen, S. H. *et al.* Diagnosing heart failure in centenarians. *J. Geriatr. Cardiol.* **16**, 1–11 (2019).
64. Barzilai, N., Huffman, D. M., Muzumdar, R. H. & Bartke, A. The Critical Role of Metabolic Pathways in Aging. *Diabetes* **61**, 1315–1322 (2012).
65. Martin, A. R. *et al.* Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **107**, 788–789 (2020).
66. Vega, F. M. D. L. & Bustamante, C. D. Polygenic risk scores: A biased prediction?. *Genome Med.* **10**, 100 (2018).
67. Barzilai, N. *et al.* Unique lipoprotein phenotype and genotype associated with exceptional longevity. *JAMA* **290**, 2030–2040 (2003).
68. Sin-Chan, P. *et al.* Exome-wide association studies in general and long-lived populations identify genetic variants related to human age. *bioRxiv* https://doi.org/10.1101/2020.07.19.188789 (2020).
69. Pinese, M. *et al.* The Medical Genome Reference Bank contains whole genome and phenotype data of 2570 healthy elderly. *Nat. Commun.* **11**, 435 (2020).
70. Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nat. Protoc.* **5**, 1564–1573 (2010).
71. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, 1–16 (2015).
72. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
73. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
74. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
75. Vergara-Lope, A. *et al.* Linkage disequilibrium maps for European and African populations constructed from whole genome sequence data. *Sci. Data* **6**, 208 (2019).
76. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
77. Choi, S. W., Mak, T.S.-H. & O'Reilly, P. F. Tutorial: A guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).

78. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
79. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience* **8**, giz082 (2019).
80. Davidson-Pilon, C. Lifelines: Survival analysis in Python. *J. Open Source Softw.* **4**, 1317 (2019).
81. Candia, J. *et al.* Assessment of variability in the SOMAscan assay. *Sci. Rep.* **7**, 14248 (2017).
82. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).

Acknowledgements

Author contributions

M.R.J. and Z.D.Z. conceived the formal analysis. M.R.J. executed the formal analysis. Z.D.Z., J.-R.L., J.M., Q.Z. and T.G. performed data curation. Z.W. performed variant imputation. M.K. performed summary statistics data munging. S.S. performed proteomic data Q.C. and contributed to editing the final draft. Z.D.Z. and N.B. conceived of the research goals and acquired funding. M.R.J. and Z.D.Z. wrote the original draft. J.V., N.B., S.M., G.A. and N.N. participated in the review and editing.

Competing interests

M.R.J is employed by and holds stock or stock options in 23andMe, Inc. J.V. is a founder of Singulomics Corp. All other authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-69069-0.

Correspondence and requests for materials should be addressed to Z.D.Z.

Reprints and permissions information is available at www.nature.com/reprints.

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