Supplementary figures and tables for:

Identifying genetic variants that affect viability in large cohorts

Hakhamanesh Mostafavi[1], Tomaz Berisa[2], Felix R Day[3], John R B Perry[3], Molly Przeworski[1,4,*], Joseph K Pickrell[1,2,*]

[1] Department of Biological Sciences, Columbia University, New York, NY, USA
[2] New York Genome Center, New York, NY, USA
[3] MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Cambridge, UK
[4] Department of Systems Biology, Columbia University, New York, NY, USA
*: These authors co-supervised this project.

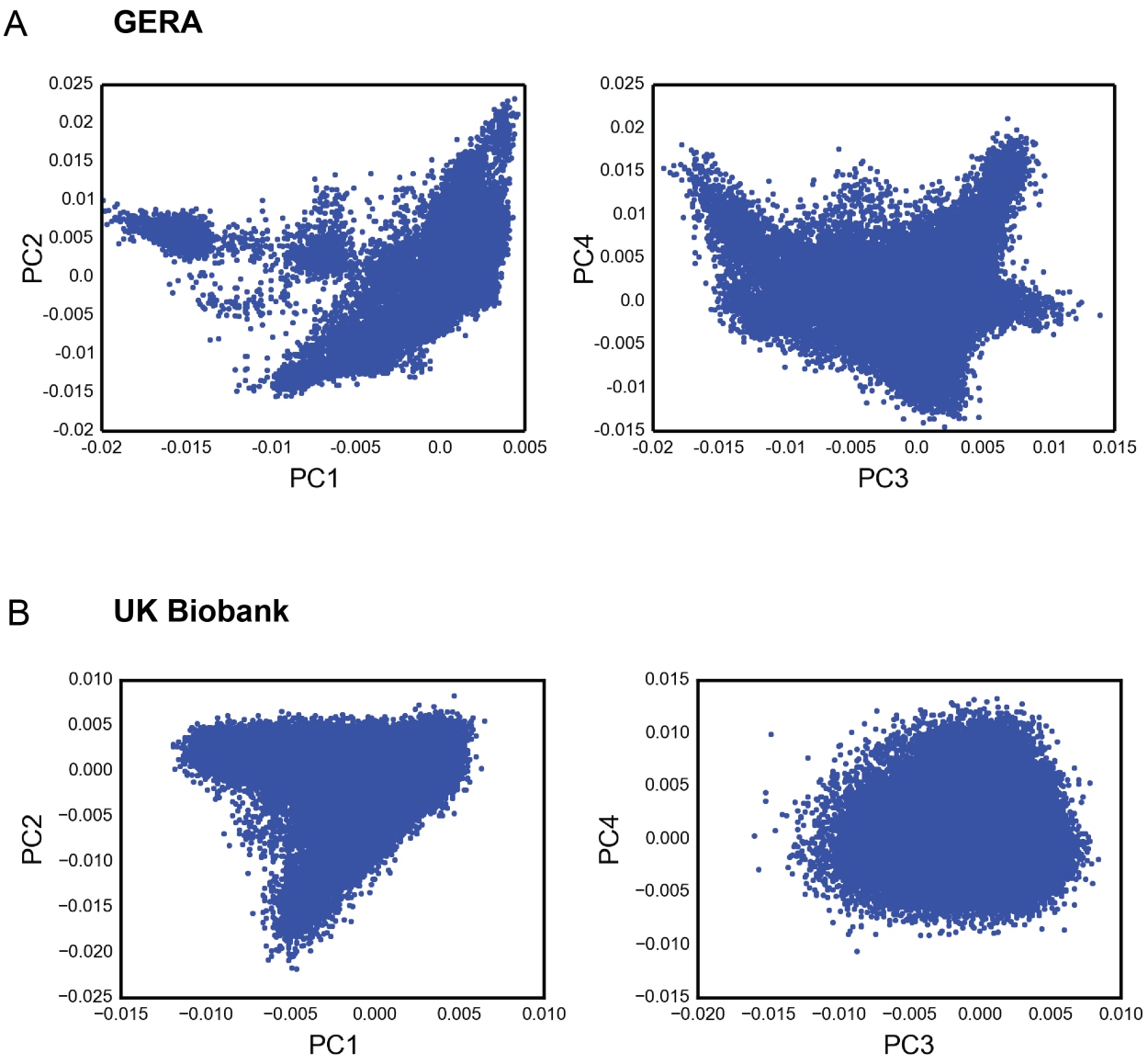

Figure S1. Results of principal component analysis (PCA). (A) PCA on 57,696 GERA individuals after quality control removing "non-European" individuals. (B) PCA on 120,286 UK Biobank participants of British ancestry. Result are in agreement with recent studies of these data [76, 81].

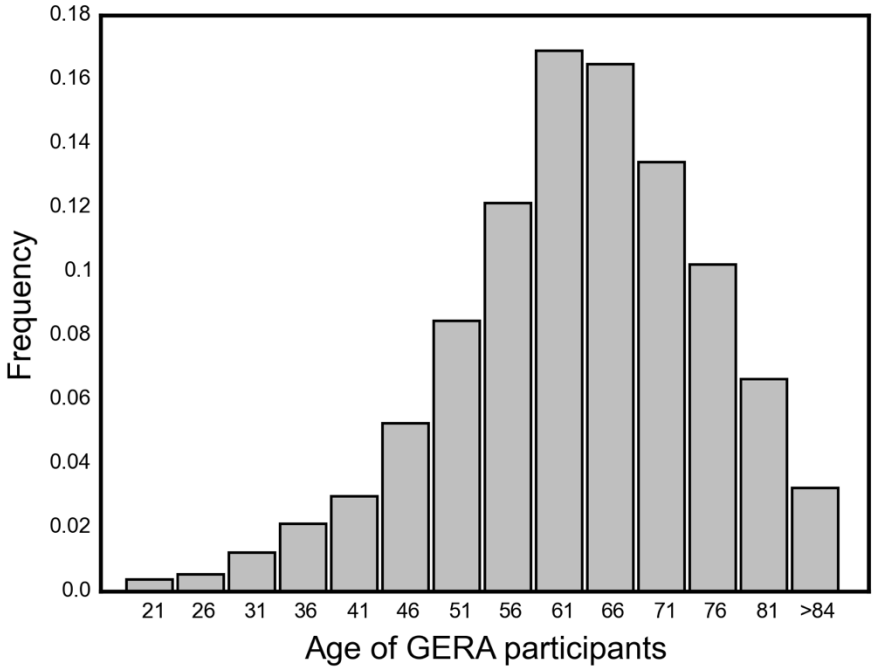

Figure S2. Age distribution of the GERA individuals at the time of the survey, year 2007. The labels on the x-axis indicate the center of 5-year interval age bins (except the last category).

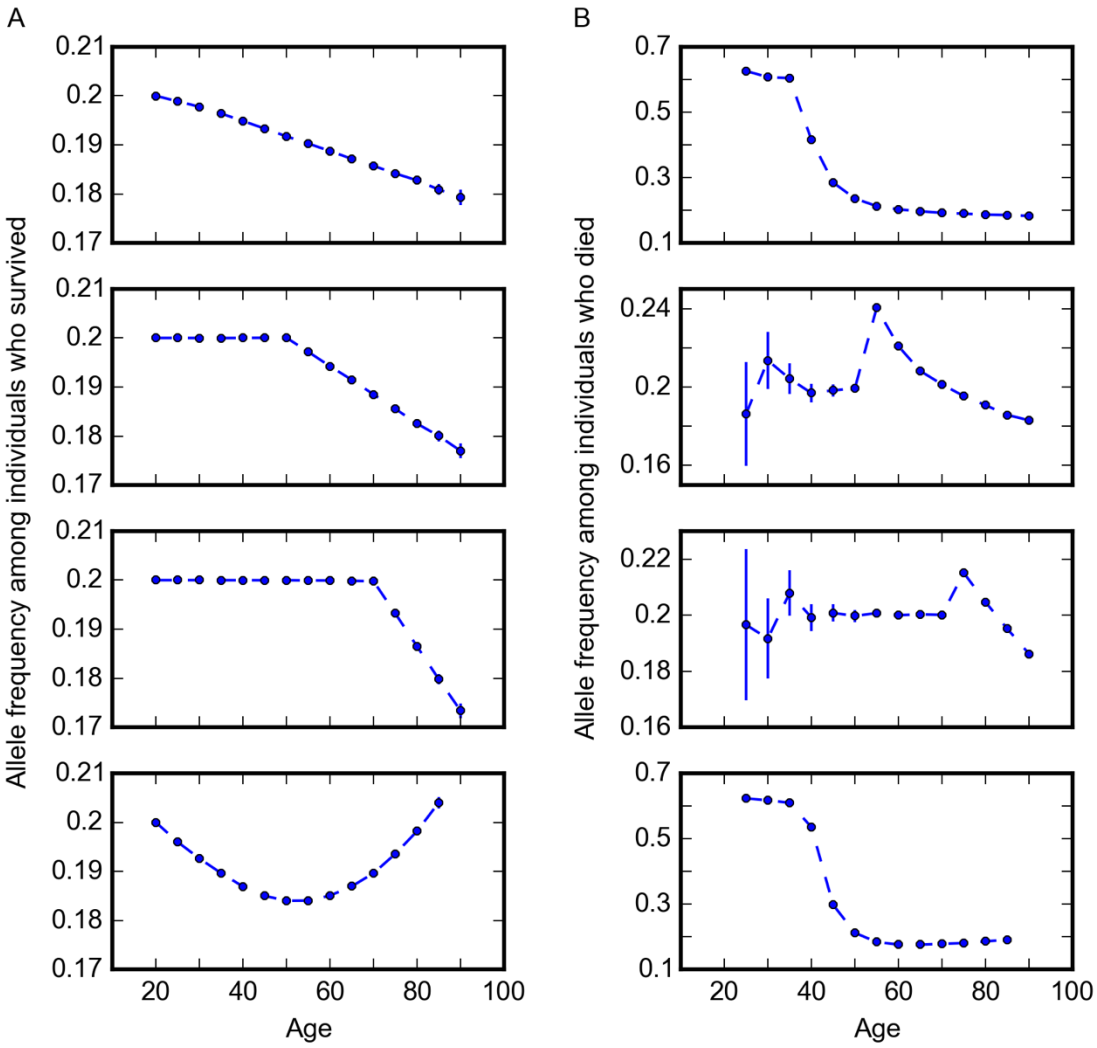

Figure S3. Comparison of trends in allele frequency with age and age at death. (A) Simulated allele frequencies among surviving individuals reproducing trends as in Figure 1A. (B) Trends in allele frequency among individuals who died, corresponding to the trends in (A). Points are allele frequency within 5-year interval age bins (mean and 95% confidence interval).

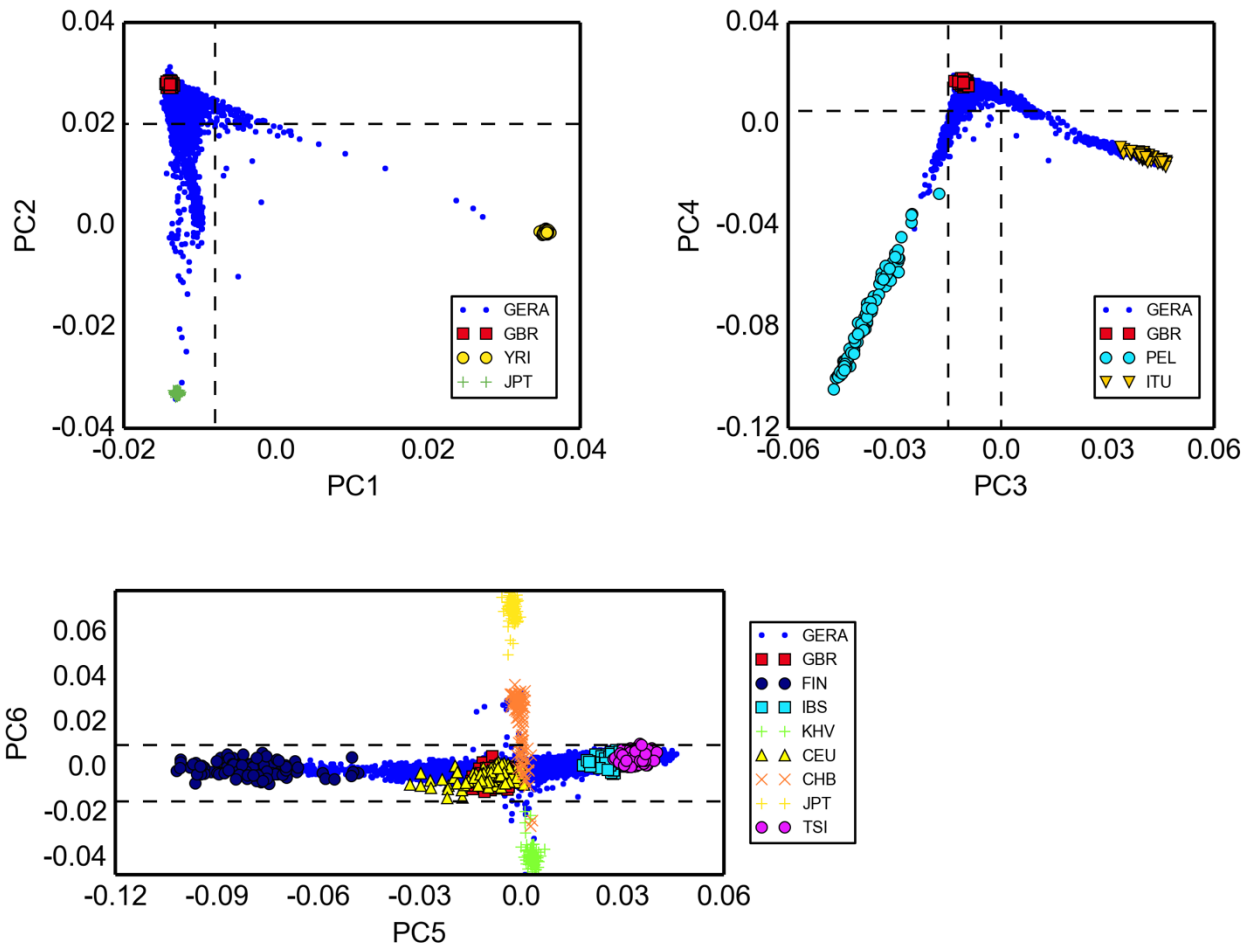

Figure S4. Validation of European ancestry in GERA. Shown are PCs inferred for all 26 populations in the 1000 Genomes Project phase 3 data. For clarity, in each plot, only a few representative populations are shown. GERA individuals (blue dots) are projected on the inferred PCs. The dashed lines correspond to the dashed lines in Figure S5, delimiting the majority of GERA individuals.

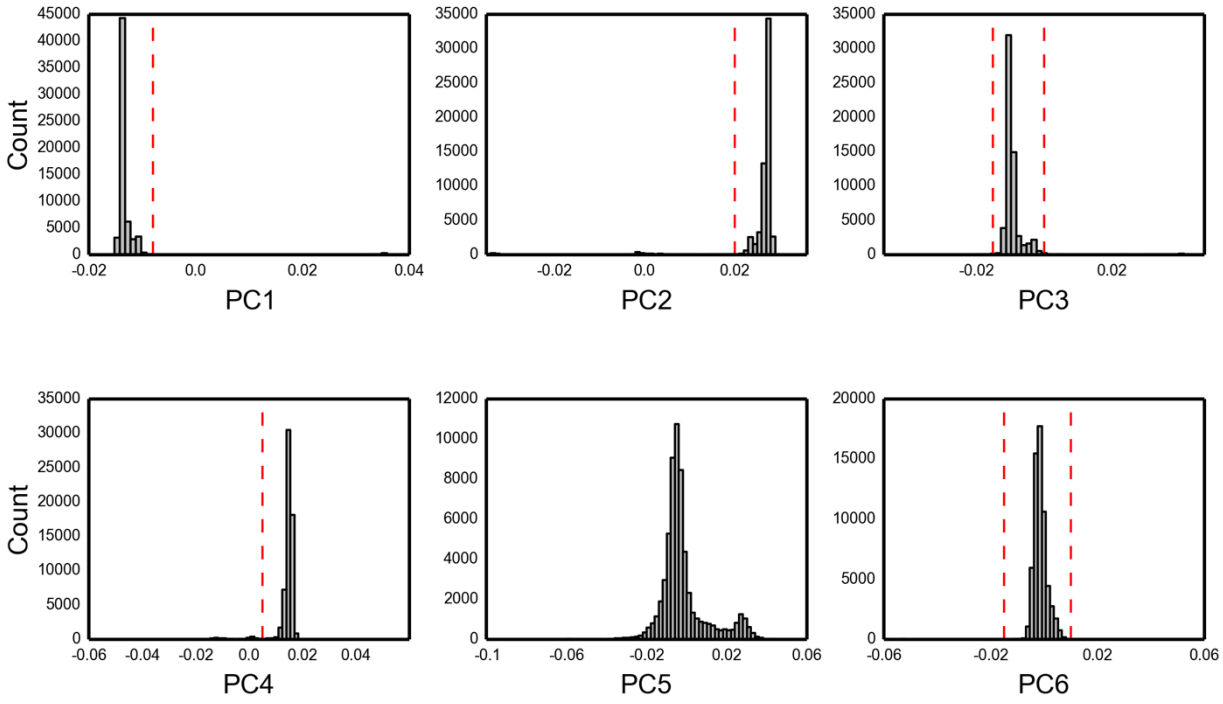

Figure S5. Distribution of GERA individuals for PCs inferred from 1000 Genome Project phase 3 data. The dashed lines enclose the majority of the data points; beyond, individuals were labeled as “non-Europeans”.

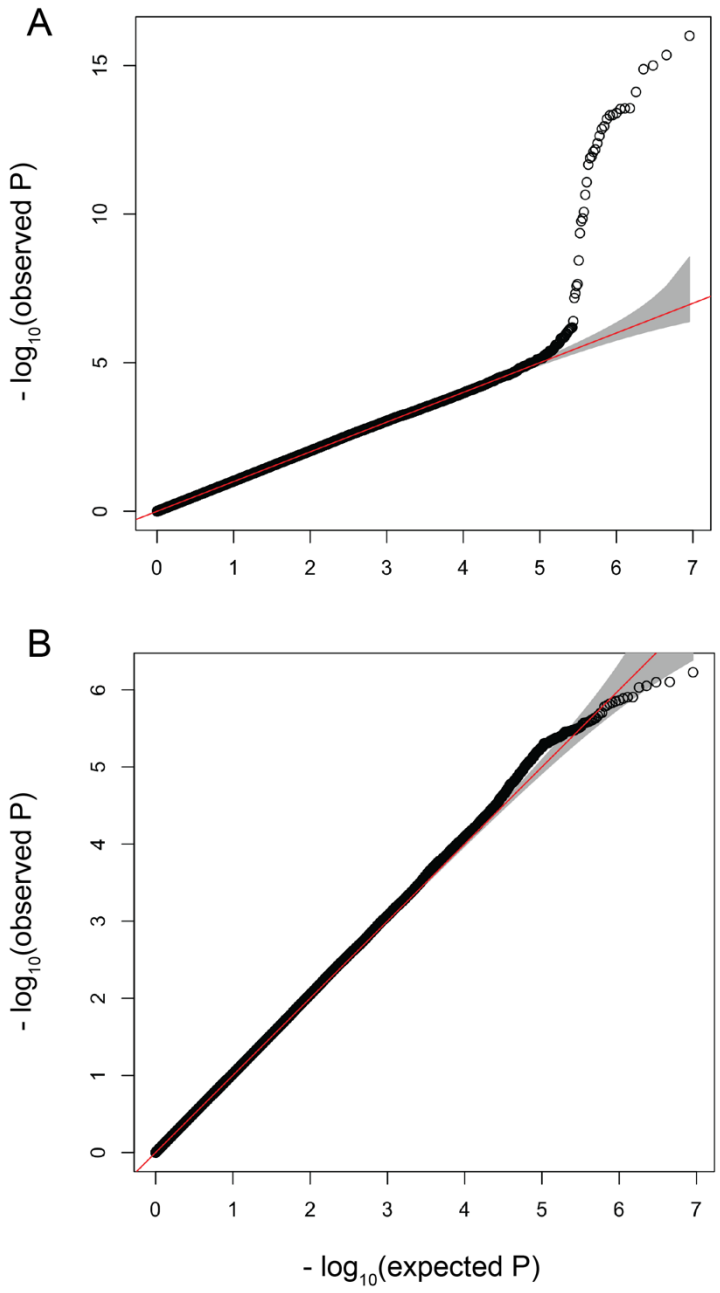

Figure S6. Quantile-quantile plots for model results for individual variants in GERA. Quantile-quantile plots for age (A), and age by sex (B) effects. The red lines indicate the distribution of the P values under the null model (of no age or age by sex effect), and the shaded bands represent the 95% confidence intervals, assuming independent SNPs.

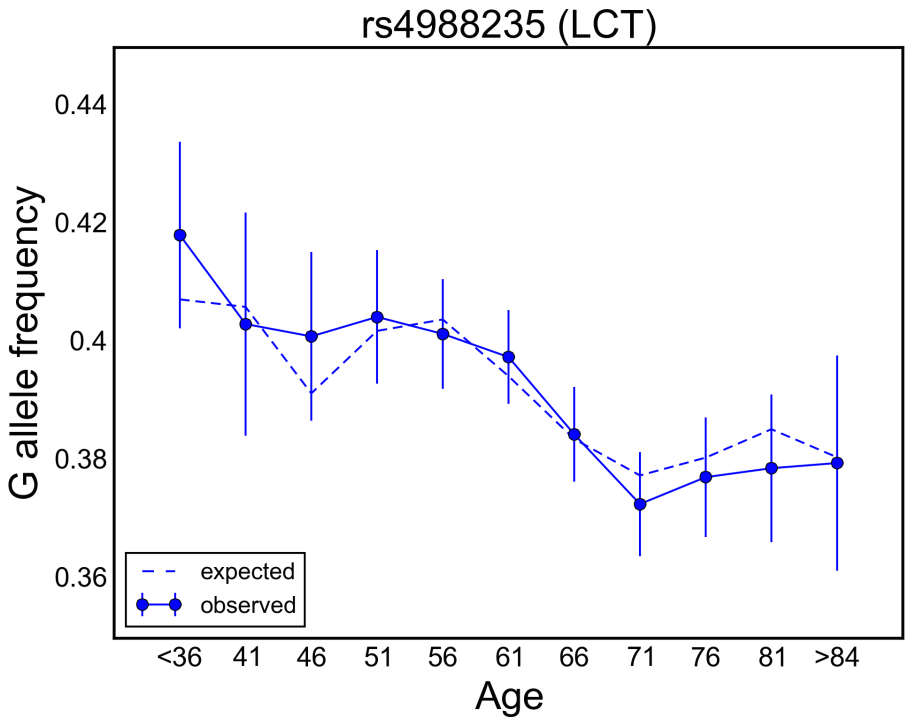

Figure S7. Frequency of the G allele of rs4988235 with age of the GERA participants. The data points are the mean frequencies within 5-year interval age bins and 95% confidence intervals. The x-axis indicates the center of the age bin. The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and, importantly, changes in ancestry.

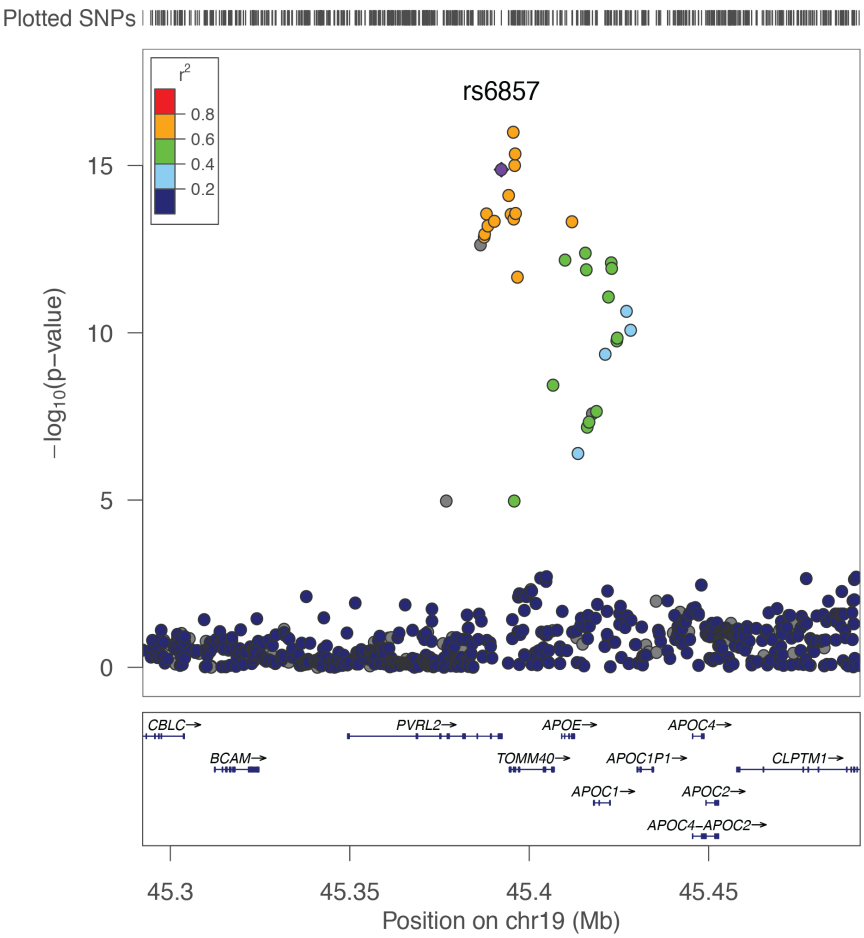

Figure S8. Regional plot for *APOE* locus. The y-axis shows P values obtained from a test of the influence of single genetic variants on age-specific mortality in GERA.

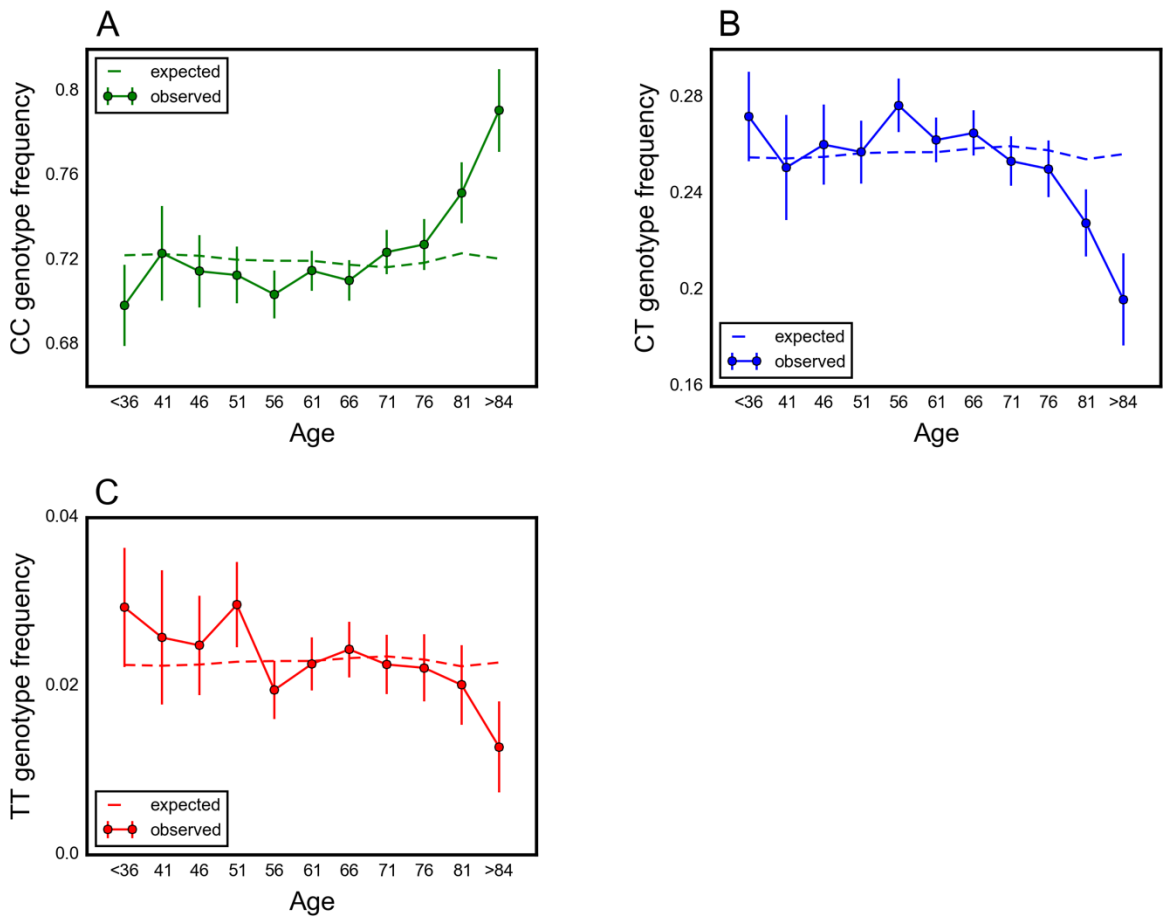

Figure S9. Frequency of rs6857 genotypes with age in GERA. Frequency of non-carriers (A), heterozygous (B) and homozygous (C) carriers of the risk allele for rs6857, tagging the ε4 allele of the *APOE* gene, across GERA age bins. Data points are frequencies within 5-year interval age bins (mean and 95% confidence interval), with the center of the bin indicated on the x-axis. Bins with ages below 36 years are merged into one bin, because of the relatively small sample sizes per bin. The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and changes in ancestry.

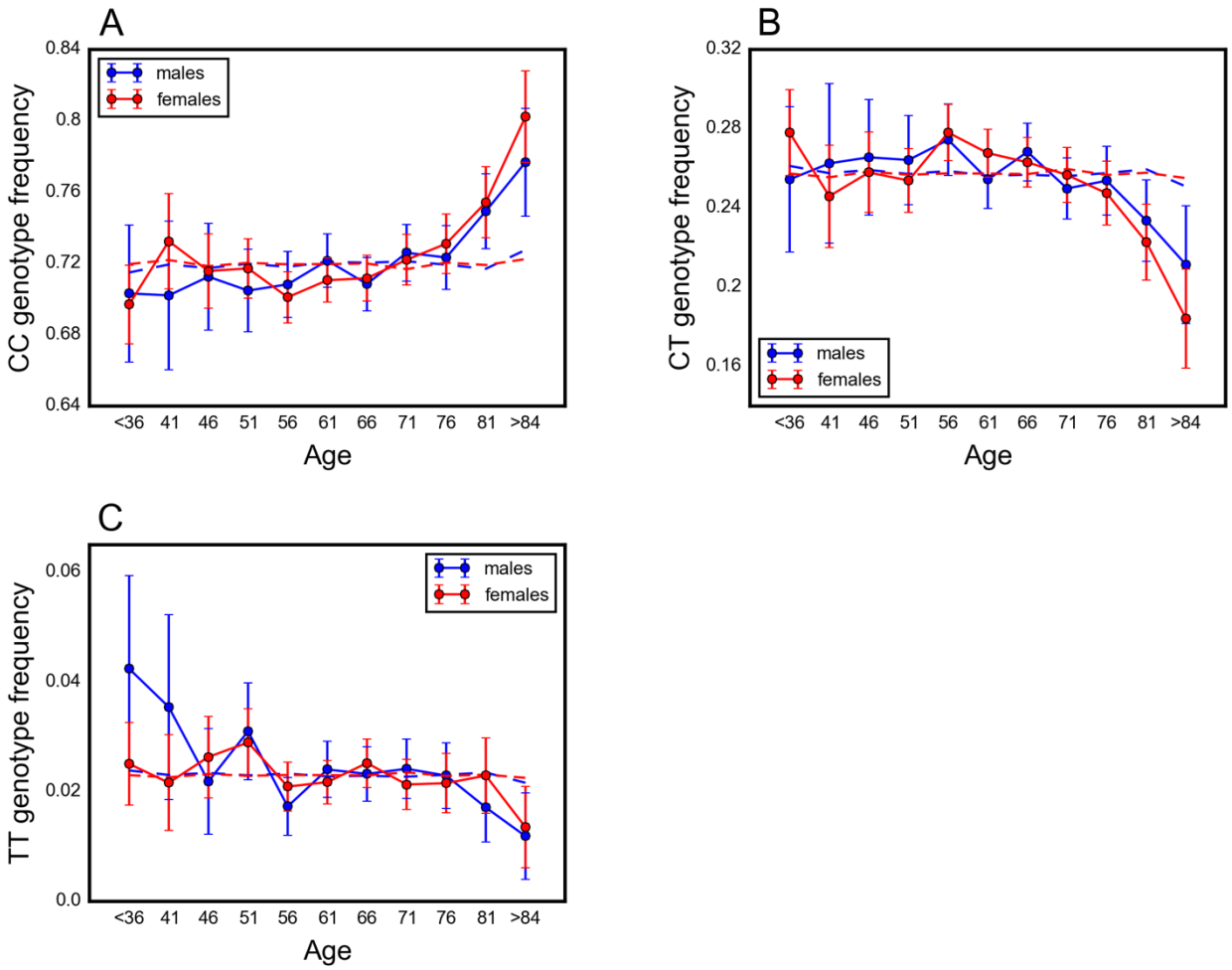

Figure S10. Frequency of rs6857 genotypes with age among males and females in GERA. Frequency of non-carriers (A), heterozygous (B) and homozygous (C) carriers of the risk allele for rs6857, tagging the ε4 allele of the *APOE* gene, across GERA age bins. Data points are frequencies within 5-year interval age bins (mean and 95% confidence interval), with the center of the bin indicated on the x-axis. Bins with ages below 36 years are merged into one bin, because of the relatively small sample sizes per bin. The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and changes in ancestry.

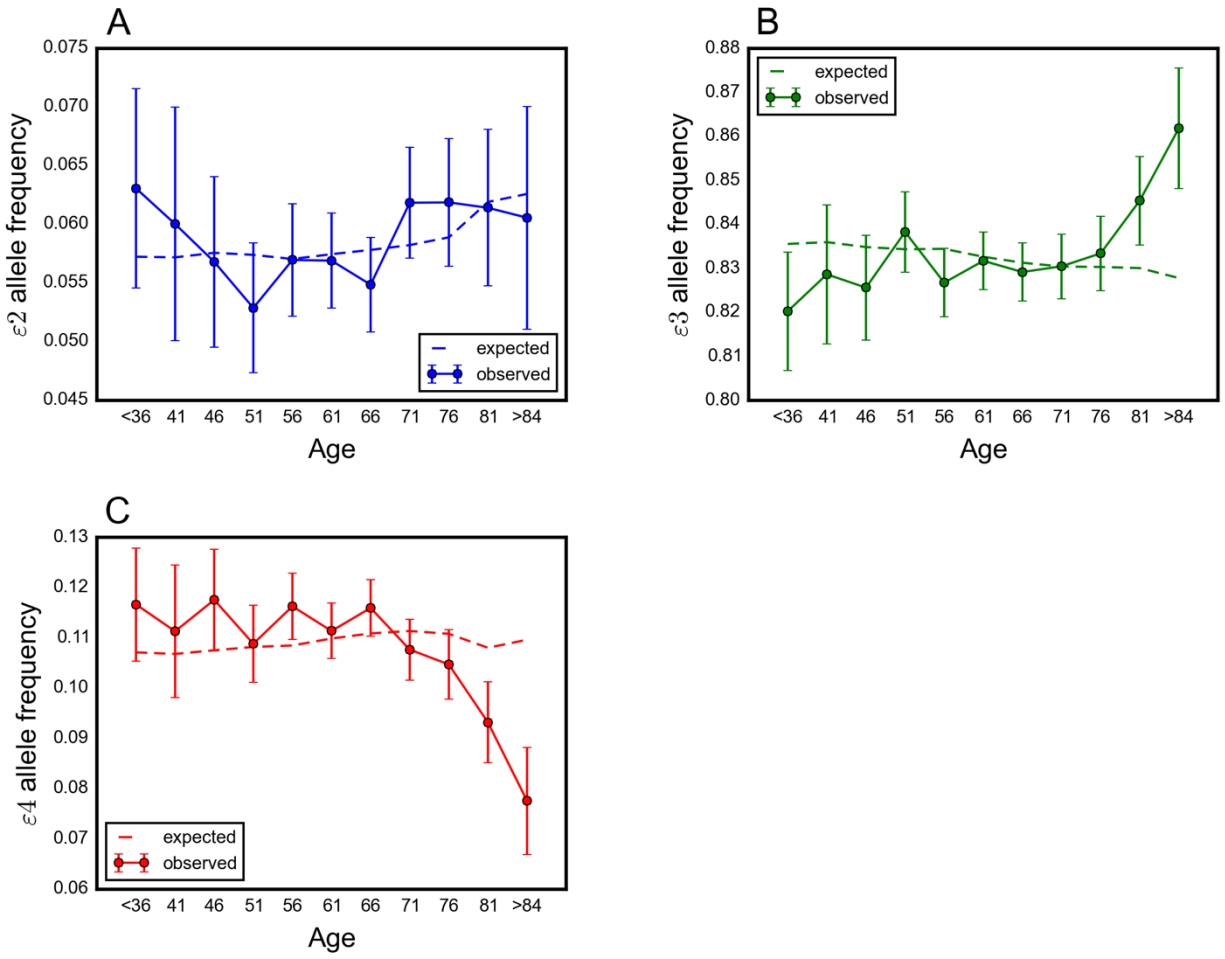

Figure S11. Frequency of the *APOE* gene alleles with age in GERA. Frequency of the ε2 (A), ε3 (B) and ε4 (C) alleles across GERA age bins. Data points are frequencies within 5-year interval age bins (mean and 95% confidence interval), with the center of the bin indicated on the x-axis. Bins with ages below 36 years are merged into one bin, because of the relatively small sample sizes per bin. The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and changes in ancestry.

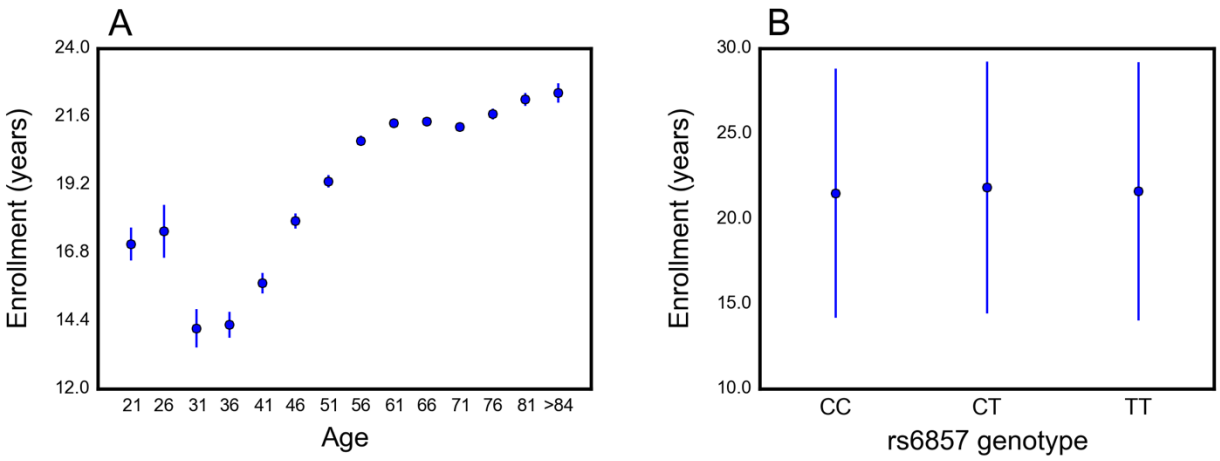

Figure S12. Enrollments of individuals in the Kaiser Permanente Medical Insurance Plan. (A) Years enrolled in the insurance plan at the time of the survey (mean and 95% confidence interval) per age bin. The x-axis indicates the center of 5-year interval age bins. (B) Years enrolled in the insurance plan (mean and 95% confidence interval) for individuals of >70 years old versus the rs6857 (*APOE*) genotype that they carry.

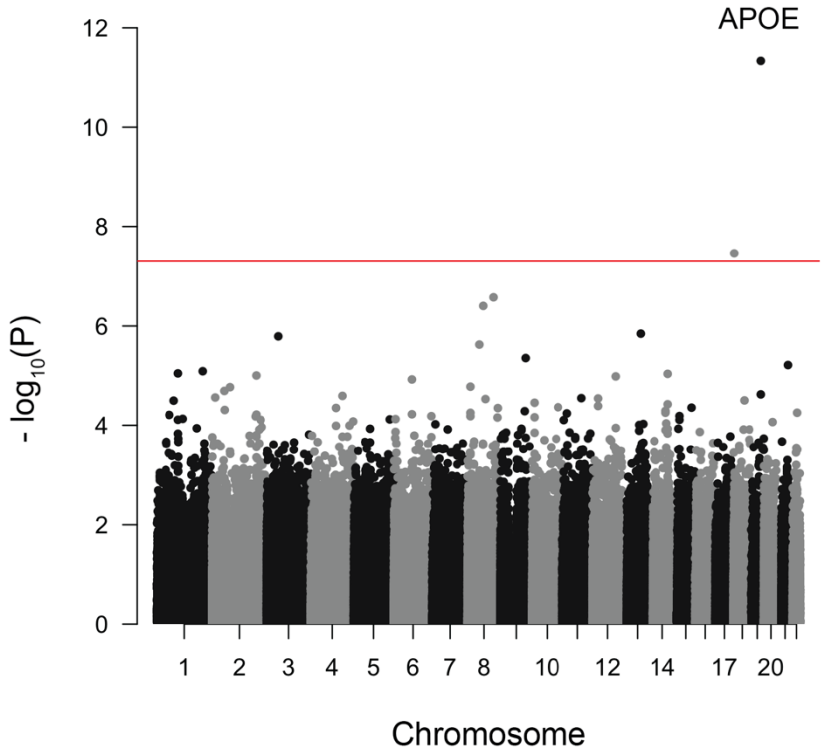

Figure S13. Testing for the influence of single genetic variants on age-specific mortality in the GERA cohort. Manhattan plot of *P* values testing for a change in allele frequency with age using the version of the model with age treated as an ordinal variable. The plot only includes the filtered genotyped SNPs in the GERA study. Red line marks the $P = 5\times10^{-8}$ threshold. The signal for variant on chromosome 18 is presumably caused by genotyping error, as other closely linked variants did not show a similar behavior, and the signal was lost when the variant was imputed using a leave-one-out approach.

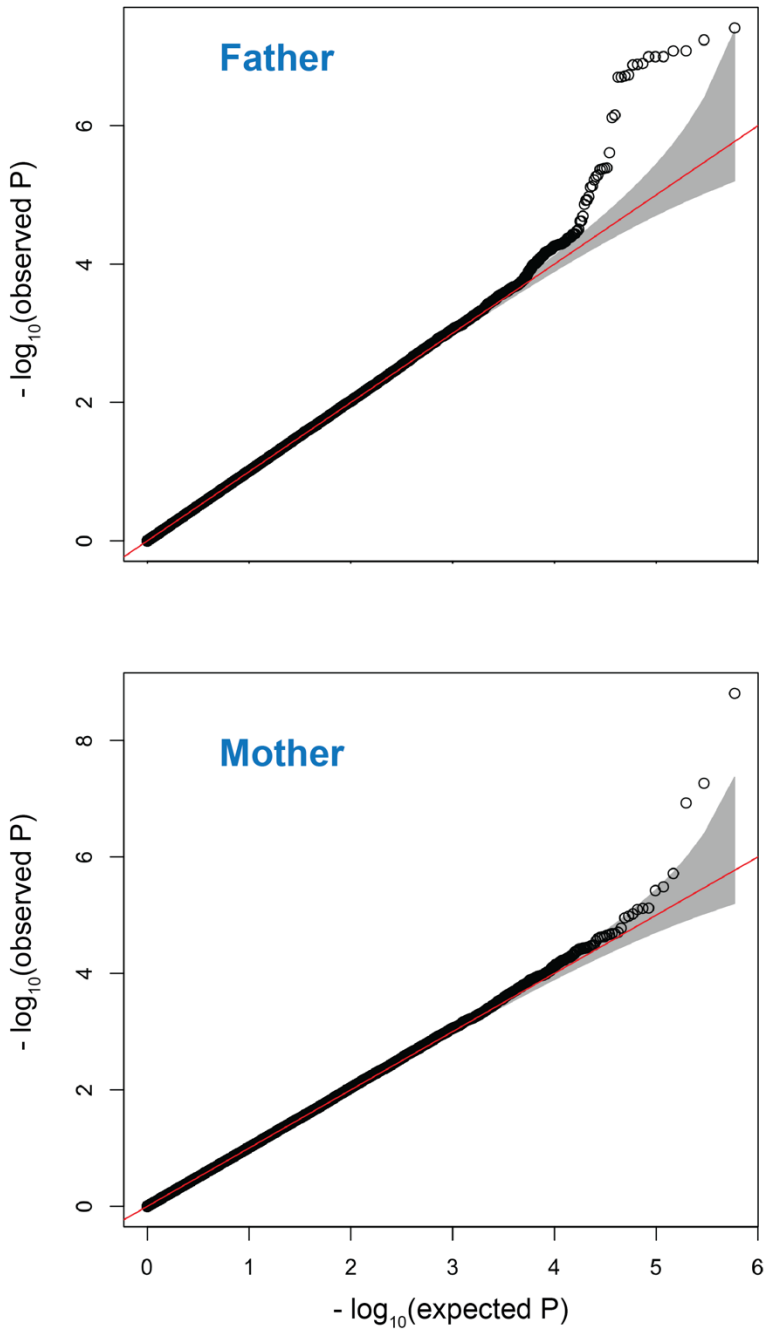

Figure S14. Quantile-quantile plots for model results for individual variants in the UK Biobank. Quantile-quantile plots for significant change in allele frequency with father's (A) and mother's (B) age at death. The red lines indicate distribution of the P values under the null (no change in frequency), and the shaded bands represent the 95% confidence intervals, assuming independent SNPs.

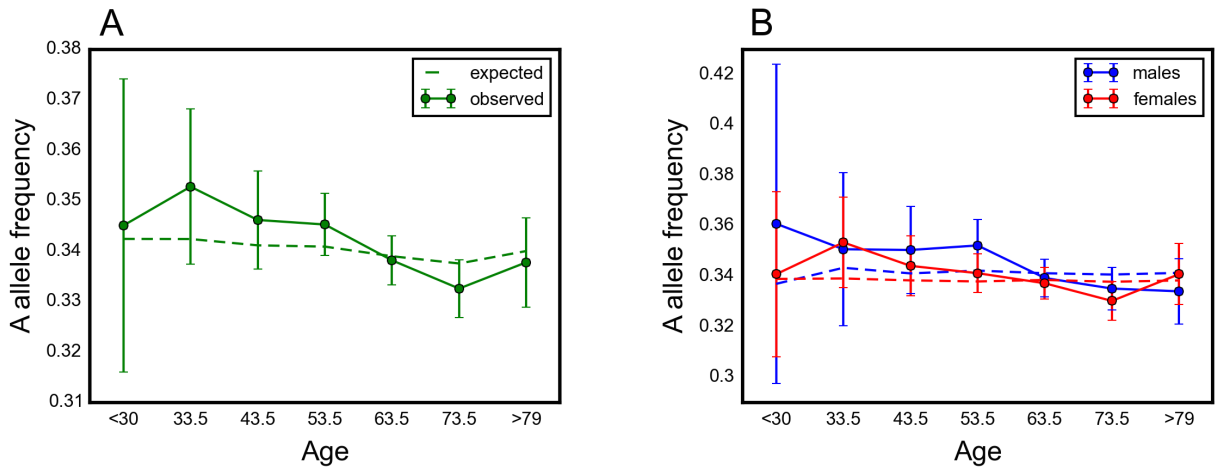

Figure S15. Effect of rs1051730 (*CHRNA3*) on survival in GERA. Allele frequency trajectory of rs1051730 with age for males and females together (A) and separately (B). The data points are the mean frequencies within 10-year interval age bins and 95% confidence intervals. The x-axis indicates the center of the age bin. The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and changes in ancestry.

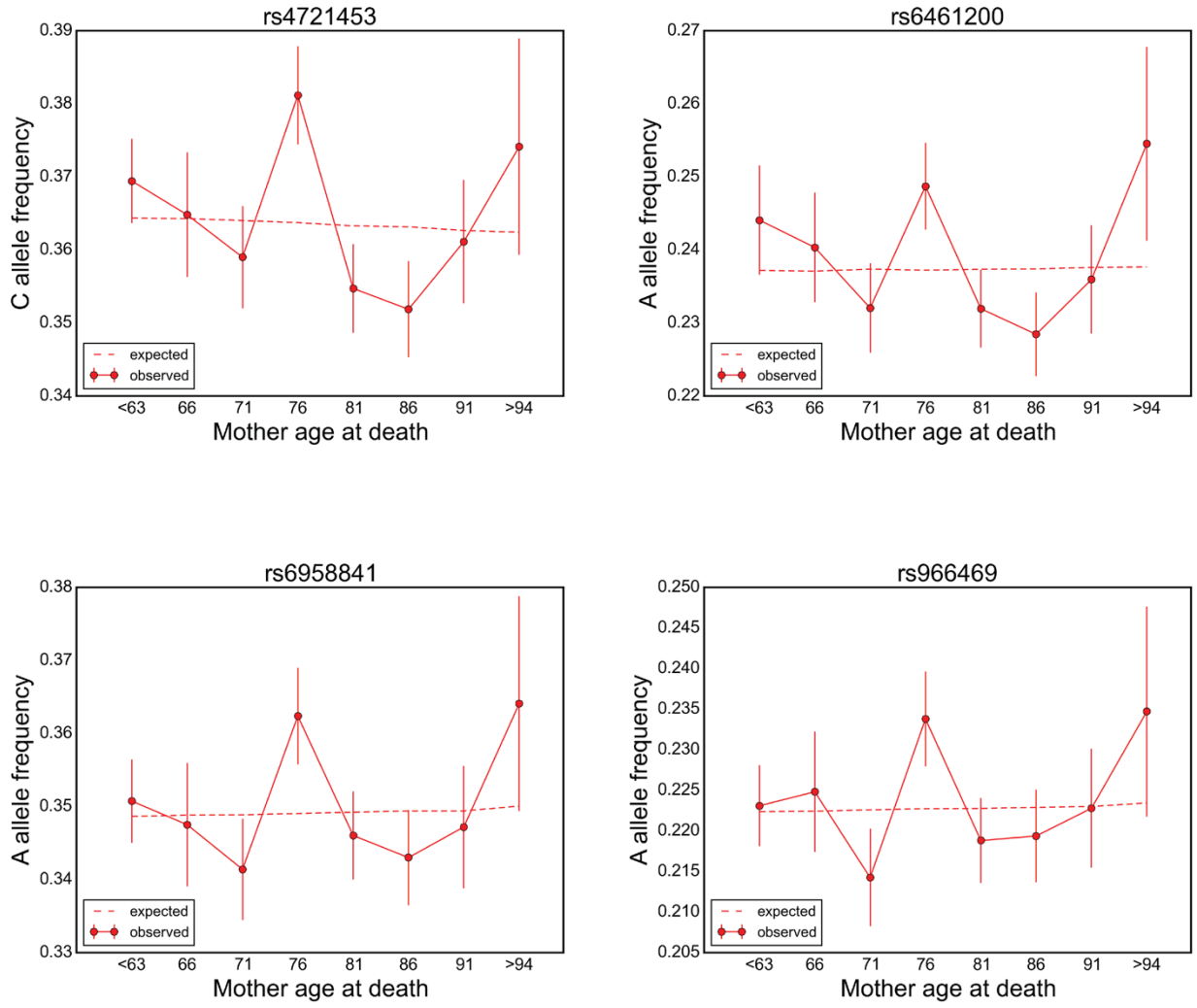

Figure S16. Allele frequencies of variants in the *MEOX2* locus with mother's age at death in the UK Biobank. Plots are for four genotyped SNPs in moderate linkage disequilibrium with $P < 10^{-4}$ for the change in allele frequency with mother's age at death. Data points are frequencies within 5-year interval age bins and 95% confidence intervals, with the center of the bin indicated on the x-axis. The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and changes in ancestry.

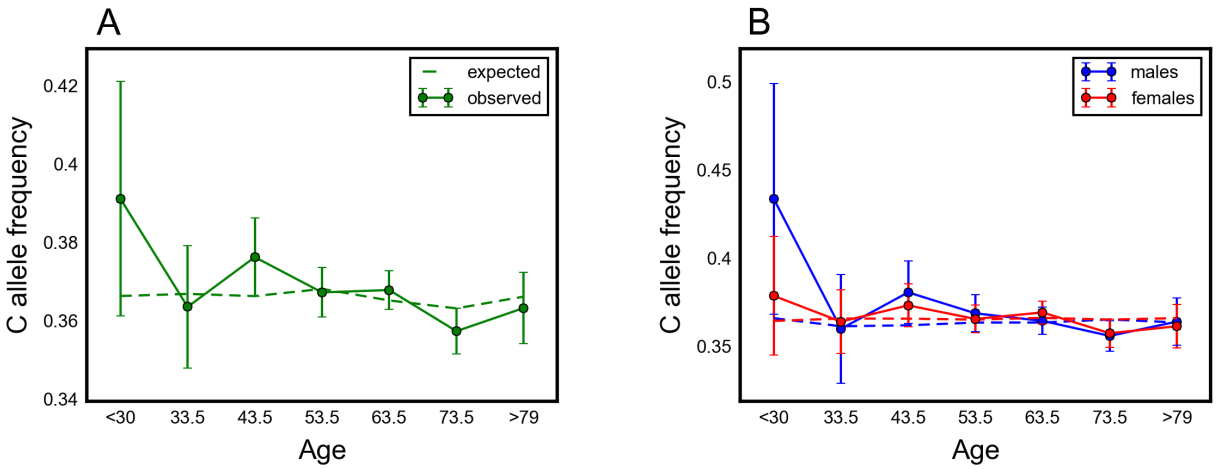

Figure S17. No significant effect of rs4721453 (near *MEOX2*) on survival in GERA ($P \sim 0.023$). Allele frequency trajectory of rs4721453 with age for males and females together (A) and separately (B). The data points are the mean frequencies within 10-year interval age bins and 95% confidence intervals. The x-axis indicates the center of the age bin. The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and changes in ancestry.

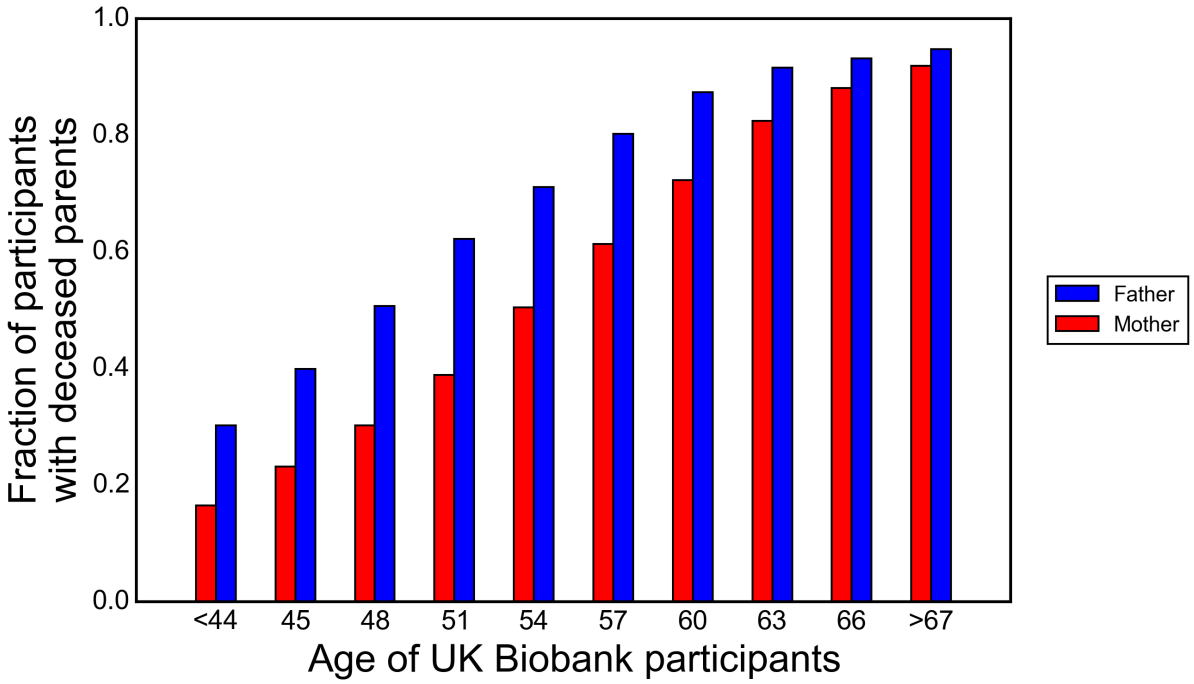

Figure S18. Ascertainment bias towards older participants introduced by using parental ages at death in the UK Biobank. Fraction of the participants in each age bin (bin size of 3 years) who reported their father's or mother's age at death.

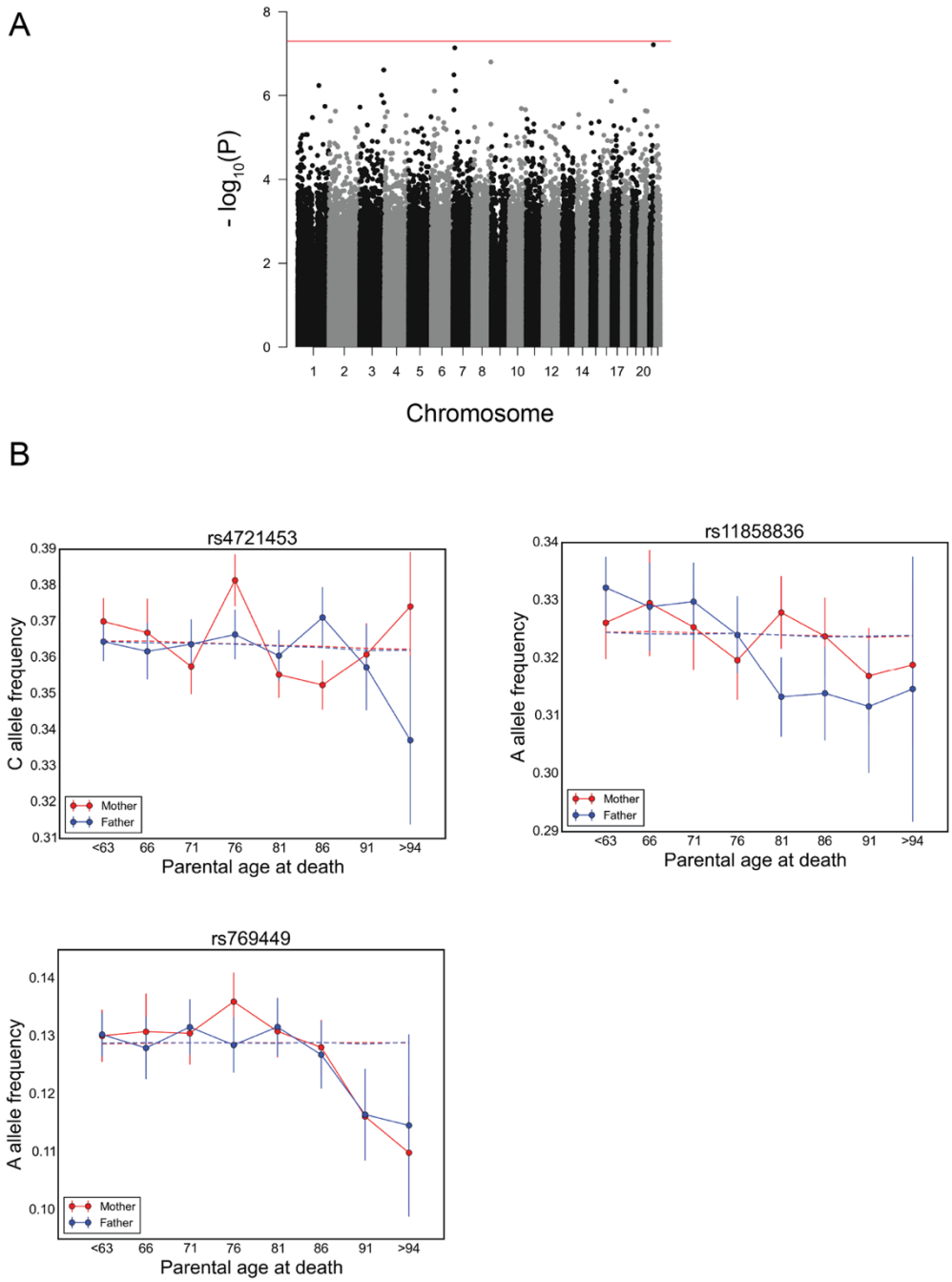

Figure S19. Testing for a significant age by sex effect of individual genetic variants in the UK Biobank. (A) Manhattan plot of *P* values, testing a difference between fathers and mothers in the change in allele frequency with parental age of death. (B) Allele frequencies as a function of father's and mother's ages at deaths, for SNPs with strongest age effects: rs4721453 (near *MEOX2*), rs11858836 (near *CHRNA3*), and rs769449 (*APOE*). The data points are the mean frequencies within 5-year interval age bins and 95% confidence intervals. The x-axis indicates the center of the age bin. The dashed line shows the expected frequency based on the null model, accounting for confounding batch effects and changes in ancestry.

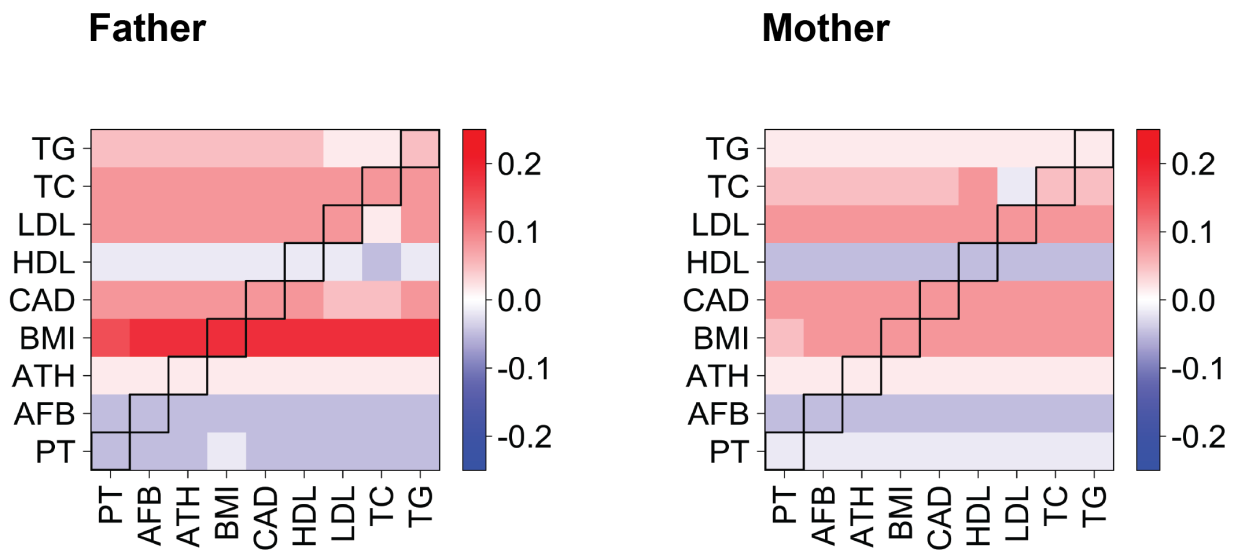

Figure S20. Heat map showing interdependence between the age effects of pairs of trait-associated variants in the UK Biobank. Each square [*i*,*j*] shows the effect size (log[hazard ratio]) of the polygenic score for trait *i* on father's (left) or mother's (right) survival in the Cox model, after accounting for the effect of the polygenic score of trait *j* (i.e., incorporating the polygenic score for trait *j* as a covariate in the null model; see Materials and Methods). Squares on the diagonal (marked by black rectangles) show the effect size of the polygenic score without accounting for the score for other traits.

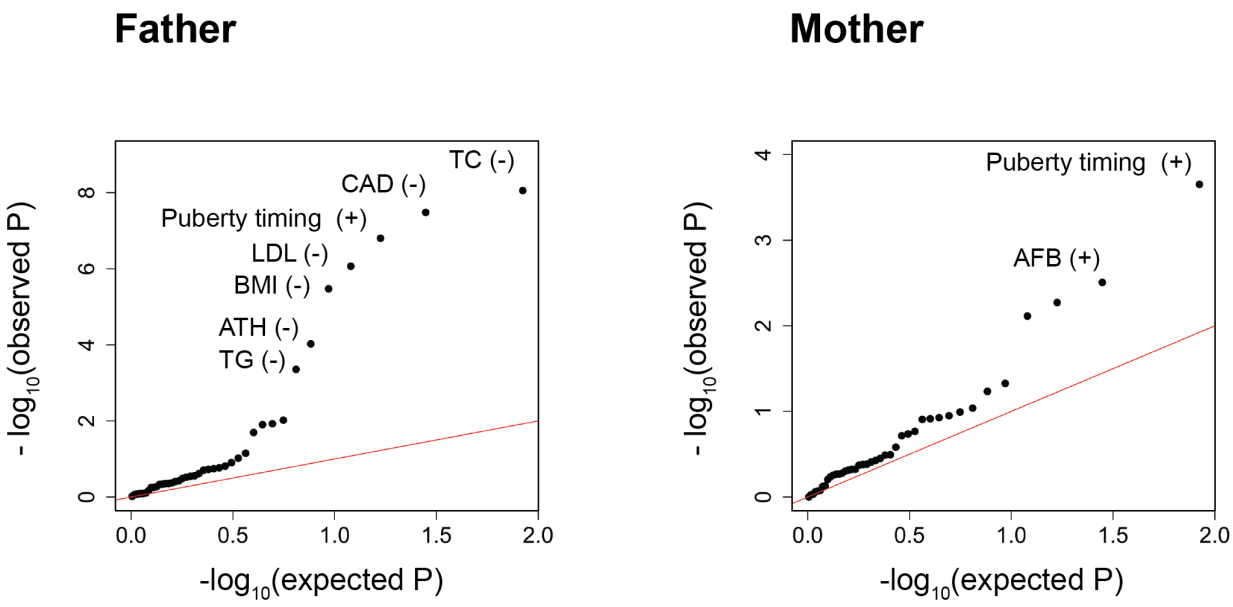

Figure S21. Testing for age effect of sets of trait-associated variants in the UK Biobank, treating age variables as ordinal. Quantile-quantile plots for changes in polygenic score of 42 traits (see Table S1) with father's (A) or mother's (B) age at death, after accounting for confounding batch effects, changes in ancestry, and the participant's age, sex, birth year, and the Townsend index (a measure of socioeconomic status). The red lines indicate the distribution of the P values under the null model. Signs '+' and '-' indicate protective and deleterious effects associated with higher values of polygenic scores, respectively.

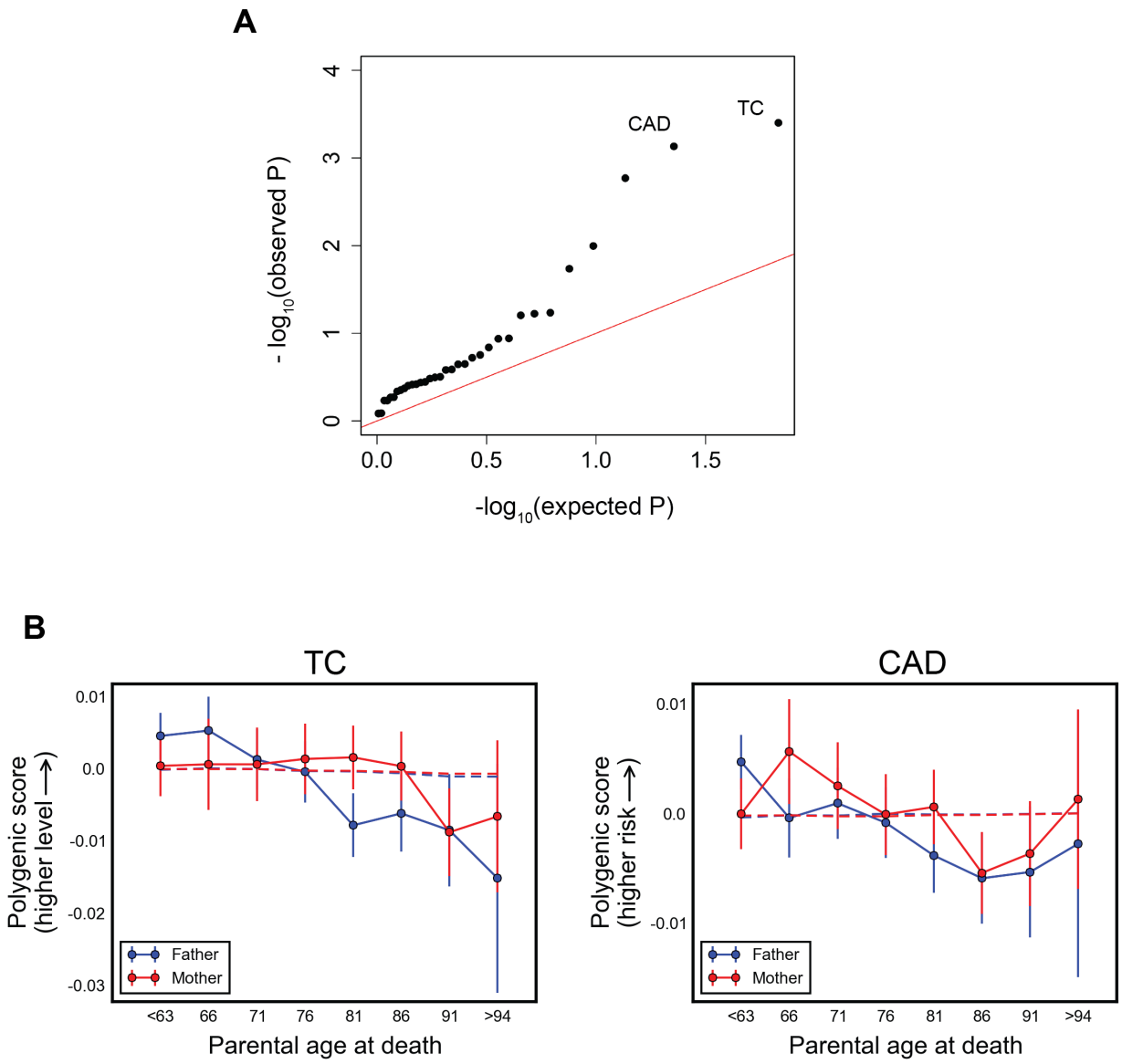

Figure S22. Testing for age by sex effect of set of trait-associated variants in the UK Biobank. (A) Quantile-quantile plot for changes in polygenic score of 42 traits (see Table S1) with parental age at deaths that are different between fathers and mothers of the UK Biobank participants. The red lines indicate the distribution of the P values under the null. (B) The trend in polygenic score with parental age at deaths for total cholesterol and coronary artery disease, which show significant age by sex effects. The data points are the mean polygenic scores within 5-year interval age bins and 95% confidence intervals. The x-axis indicates the center of the age bin. The dashed line shows the expected polygenic score based on the null model, accounting for confounding batch effects, changes in ancestry, and the participant's age, sex, birth year, and the Townsend index (a measure of socioeconomic status).

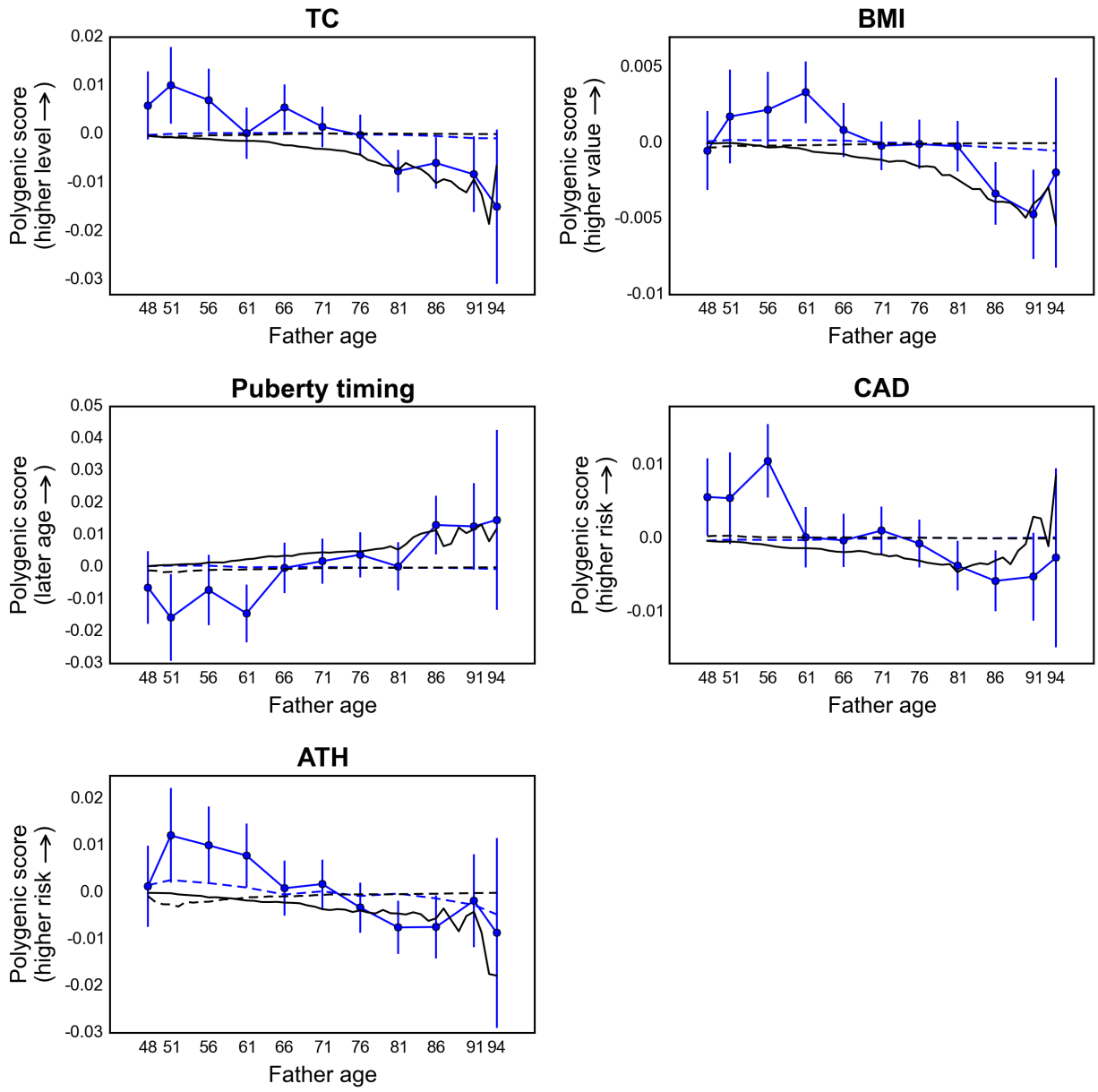

Figure S23. Trajectories of polygenic scores with paternal age for traits associated with parental survival in the UK Biobank. Each plot shows (in blue) the mean polygenic score (and 95% confidence interval) among the fathers who died in a 5-year interval centered around the plotted discs, and (in black) the mean polygenic score among fathers alive up to a given age, i.e., all fathers with age or age at death (if deceased) exceeding a given age. The dashed lines show the expected changes in polygenic scores based on the null model. If there is no effect of the score on survival at a given time (age), then the score among those who died (blue disc) should be the same as the score among those who were alive at the previous time interval. Thus, the divergence between the blue and the black lines in any time interval is an indicator of the effect of the score on survival (and its direction) within that interval. The precise effect, however, also depends on the total hazard rate of the sample, which varies by age.

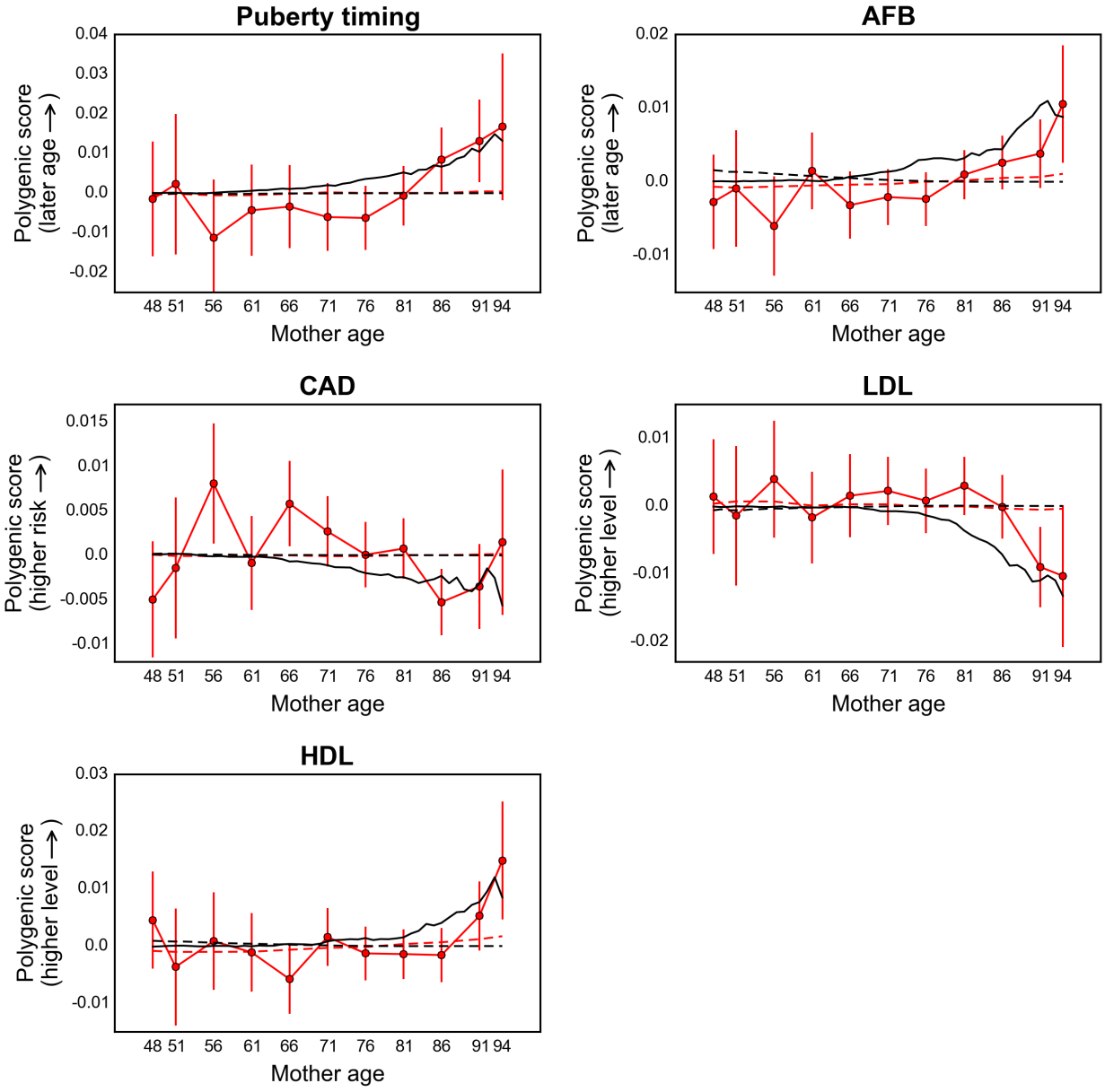

Figure S24. Trajectories of polygenic scores with maternal age for traits associated with maternal survival in the UK Biobank. Same as Figure S23, but plotted for mothers (with red instead of blue).

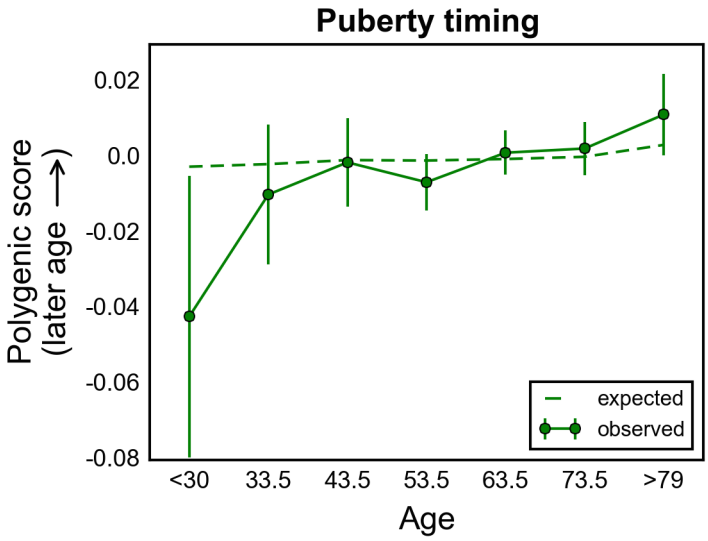

Figure S25. Protective effect of later predicted puberty timing on survival in GERA ($P \sim 0.0067$). Polygenic score for puberty timing with age of the participants. The data points are the mean scores within 10-year interval age bins and 95% confidence intervals. The x-axis indicates the center of the age bin. The dashed line shows the expected score based on the null model, accounting for confounding batch effects and changes in ancestry.

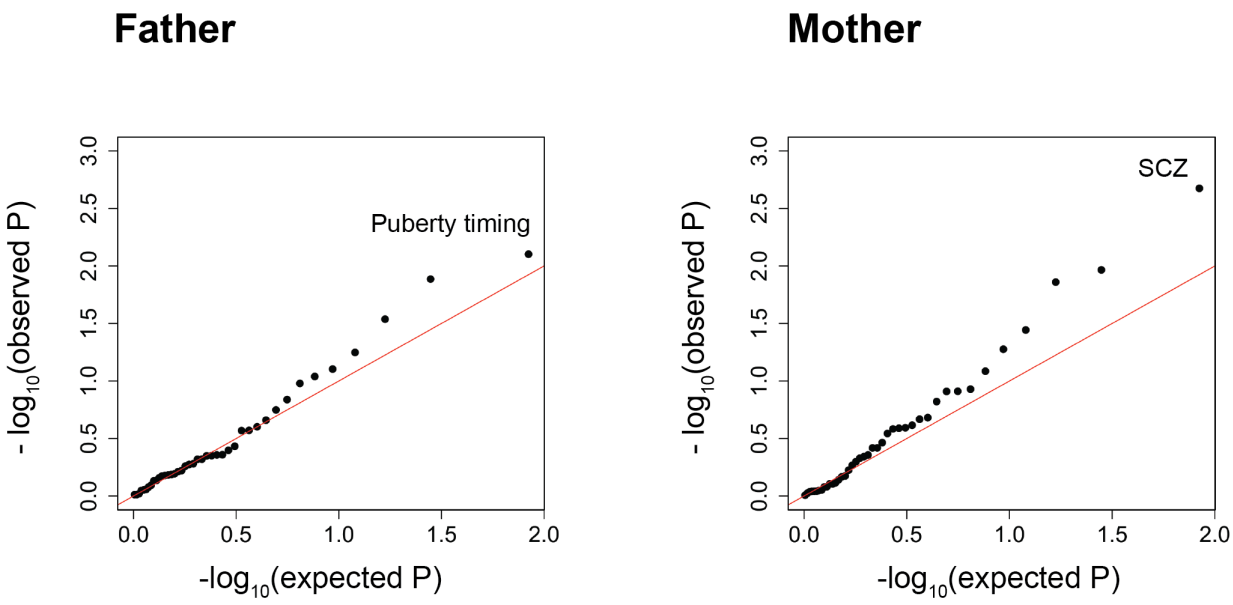

Figure S26. Testing for stabilizing selection on traits in the UK Biobank. Quantile-quantile plots testing for a change in the squared difference of polygenic score from the mean with fathers' (A) and mothers' (B) age at death. 42 traits were tested (see Table S1). The red line indicates the distribution of the P values under the null model.

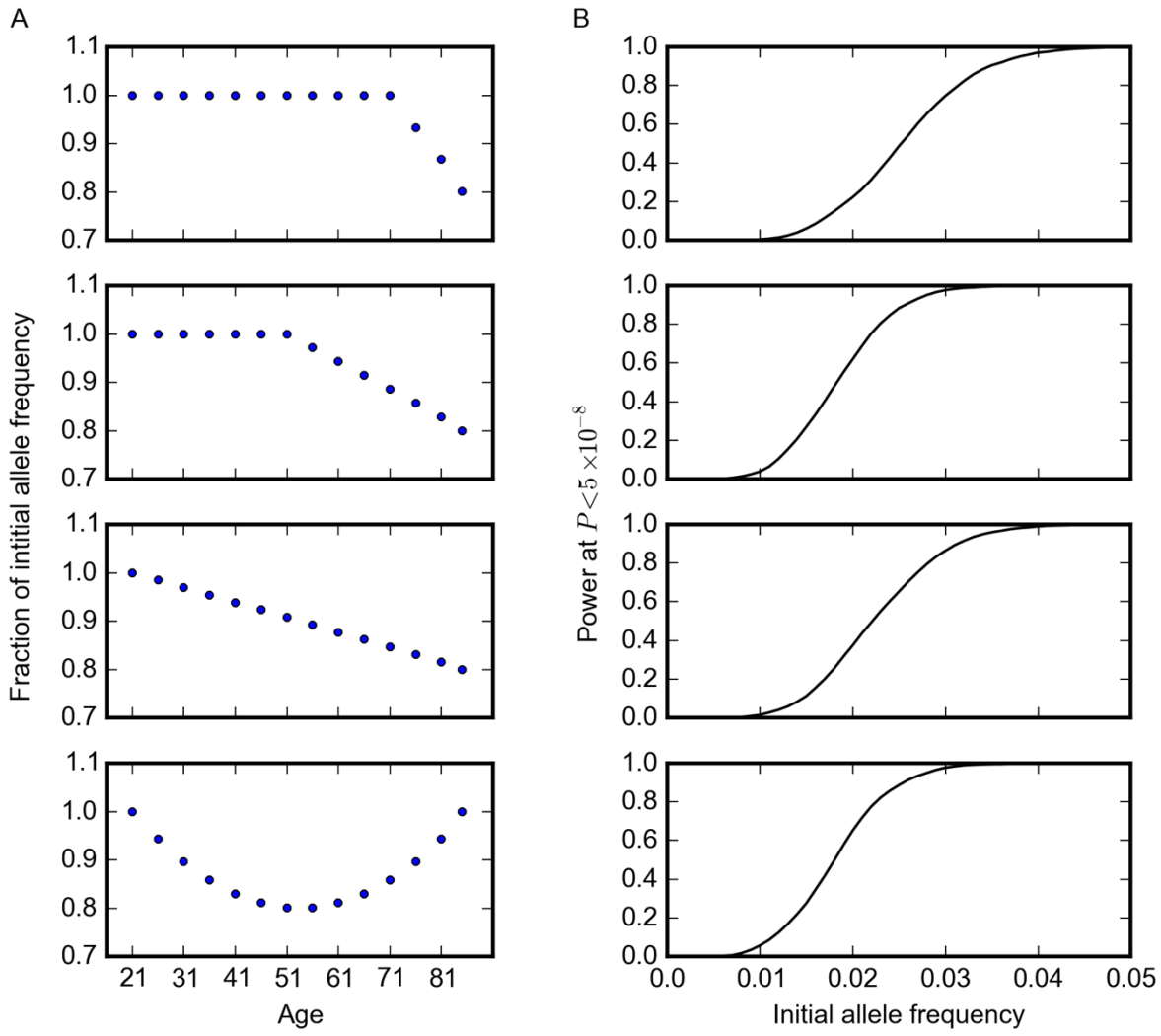

Figure S27. Power of the model to detect changes in allele frequency with age. Same as Figure 1, but with 500,000 samples evenly distributed among age categories, and only showing the results using models with age treated as a categorical variable. As can be seen, there should be substantial power to detect such effects even for relatively rare variants (i.e., at a couple of percent frequency in the population).

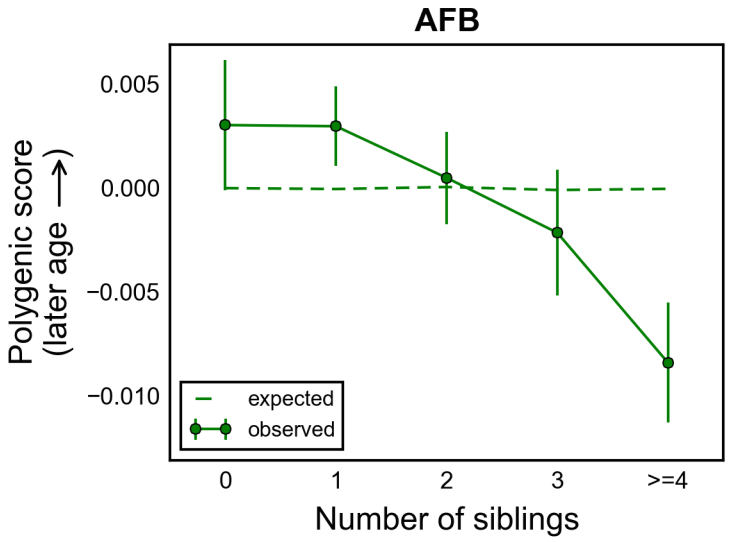

Figure S28. Association between variants influencing age at first birth and apparent fertility in the UK Biobank ($P \sim 4.2 \times 10^{-8}$). Polygenic score versus the number of siblings for 112,130 participants with mother's age ≥ 50 years. Data points are mean scores (and 95% confidence intervals). The polygenic score was regressed on the number of siblings, accounting for the confounding batch effects, changes in ancestry, and the participant's age, sex, birth year, and the Townsend index (a measure of socioeconomic status). The dashed line shows the expected score based on the null model.

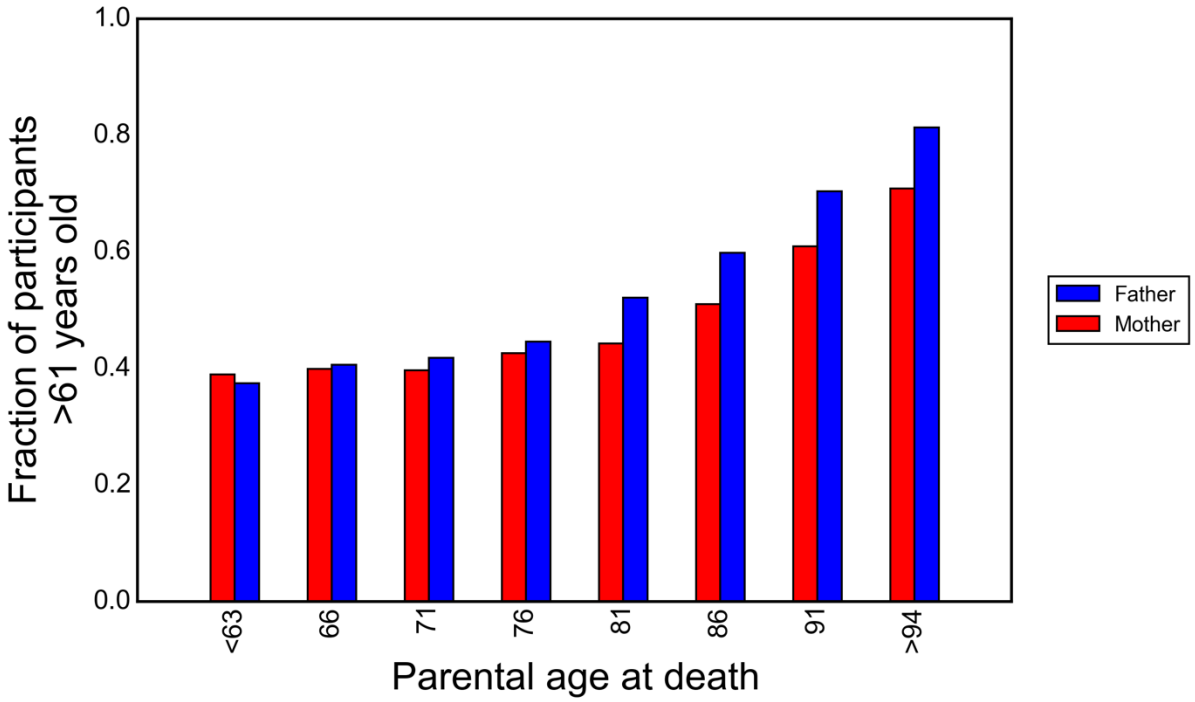

Figure S29. Ascertainment bias towards older participants for older parental age at death categories. Fraction of the participants > 61 years old (last three age categories in panel A of Figure S18) in each parental age bin. Assuming parents of older participants on average belong to earlier generations, older age at death categories will contain parents born earlier.

Table S1. List of phenotypes and abbreviations. The numbers of loci passing quality control measures are shown for each dataset.

Phenotype	Abbreviation	# loci in UK Biobank [a]	# loci in UK Biobank [b]	# loci in GERA
Age at first birth	AFB	10	10	10
Age at natural menopause	ANM	53		
Age at voice drop	AVD	5		
Alzheimer's disease [c]	AD	4		
Any allergies	ALL	35		
Asthma	ATH	33	33	32
Beighton hypermobility	BHM	17		
Body mass index	BMI	30	30	30
Bone mineral density (femoral neck)	FNBMD	18		
Bone mineral density (lumbar spine)	LSBMD	20		
Breast size	CUP	14		
Childhood ear infections	CEI	13		
Chin dimples	DIMP	52		
Coronary artery disease	CAD	11	11	10
Crohn's disease	CD	58		
Fasting glucose	FG	15		
Height	HEIGHT	561		
Hemoglobin	HB	15		
High-density lipoproteins	HDL	45	45	46
Hypothyroidism	HTHY	26		
Low-density lipoproteins	LDL	39	40	39
Male pattern baldness	MPB	44		
Mean cell hemoglobin concentration	MCHC	15		
Mean platelet volume	MPV	29		
Mean red cell volume	MCV	41		
Migraine	MIGR	29		
Nearsightedness	NST	159		
Nose size	NOSE	10		
Packed red cell volume	PCV	12		
Parkinson's disease	PD	23		
Photic sneeze reflex	PS	60		
Platelet count	PLT	49		
Puberty timing [d]	PT	359	358	359
Red blood cell count	RBC	22		
Rheumatoid arthritis	RA	68		
Schizophrenia	SCZ	191		
Tonsillectomy	TS	38		
Total cholesterol	TC	49	51	50
Triglycerides	TG	28		
Type 2 diabetes	T2D	11		
Unibrow	UB	53		
Waist-hip ratio	WHR	12		

a: Among participants of British genetic ancestry

b: Among participants of non-British genetic ancestry

c: For AD the *APOE* locus was excluded.

d: Age at menarche associated variants were used to proximate puberty timing scores in both males and females because of the strong genetic correlation between the timing of puberty in males and females [57].

Table S2. Results of the Cox model for association of polygenic scores for 42 traits with survival of parents of the UK Biobank participants.

Table S3. Age-dependency of hazard ratios for the top associations with parental survival in the UK Biobank under the Cox model.

Trait	Age range	Father			Mother		
		Effect size (s.e.)	HR*	*P* value	Effect size (s.e.)	HR*	*P* value
Puberty timing	>75	-0.0167 (0.0129)	0.98	0.20	-0.0174 (0.0123)	0.98	0.16
	≤75	-0.0486 (0.0102)	0.95	2.1×10^{-6}	-0.0390 (0.0130)	0.96	0.0027
AFB	>75	-0.0453 (0.0291)	0.96	0.12	-0.0399 (0.0276)	0.96	0.15
	≤75	-0.0370 (0.0229)	0.96	0.11	-0.0903 (0.0290)	0.91	0.0019
ATH	>75	-0.0122 (0.0176)	0.99	0.49	0.0130 (0.0168)	1.01	0.44
	≤75	0.0524 (0.0138)	1.05	1.5×10^{-4}	0.0173 (0.0176)	1.02	0.32
BMI	>75	0.1658 (0.0575)	1.18	0.0040	0.0659 (0.0545)	1.07	0.23
	≤75	0.2182 (0.0451)	1.24	1.3×10^{-6}	0.0961 (0.0572)	1.1	0.093
CAD	>75	-0.0111 (0.0286)	0.99	0.70	0.0647 (0.0270)	1.07	0.017
	≤75	0.1337 (0.0224)	1.14	2.6×10^{-9}	0.1150 (0.0284)	1.12	5.2×10^{-5}
HDL	>75	-0.0304 (0.0225)	0.97	0.18	-0.0807 (0.0214)	0.92	1.6×10^{-4}
	≤75	-0.0360 (0.0176)	0.96	0.041	-0.0394 (0.0224)	0.96	0.078
LDL	>75	0.0431 (0.0227)	1.04	0.058	0.1183 (0.0214)	1.12	3.3×10^{-8}
	≤75	0.1030 (0.0177)	1.11	6.2×10^{-9}	0.0487 (0.0225)	1.05	0.03
TC	>75	0.0490 (0.0222)	1.05	0.027	0.1014 (0.0210)	1.11	1.4×10^{-6}
	≤75	0.1144 (0.0173)	1.12	4.2×10^{-11}	0.0319 (0.0220)	1.03	0.15

*Hazard ratio.

Table S4. Replication of associations in the discovery panel (UK Biobank individuals of British ancestry) in the UK Biobank participants of non-British ancestry.

Trait	Father			Mother			Meta-analysis [a]		
	Effect size (s.e.)	HR	*P* value	Effect size (s.e.)	HR	*P* value	Effect size (s.e.)	HR	*P* value
Puberty timing	-0.0497 (0.0164)	0.95	0.0024	-0.0276 (0.0186)	0.97	0.14	-0.0401 (0.0123)	0.96	0.0011
AFB	-0.0167 (0.0359)	0.98	0.64	-0.0277 (0.0409)	0.97	0.50	-0.0215 (0.0270)	0.98	0.42
ATH	0.0296 (0.0204)	1.03	0.15	0.0253 (0.0228)	1.02	0.27	0.0277 (0.0152)	1.03	0.068
BMI	0.2240 (0.0719)	1.25	0.0018	0.2137 (0.0814)	1.24	0.0087	0.2194 (0.0539)	1.24	4.7×10^{-5}
CAD	0.0466 (0.0354)	1.05	0.19	0.0578 (0.0399)	1.06	0.15	0.0516 (0.0265)	1.05	0.051
HDL	-0.0168 (0.0276)	0.98	0.54	-0.0442 (0.0313)	0.96	0.16	-0.0288 (0.0207)	0.97	0.16
LDL	0.0711 (0.0282)	1.07	0.012	0.0865 (0.0320)	1.09	0.0068	0.0778 (0.0211)	1.08	2.3×10^{-4}
TC	0.0460 (0.0273)	1.05	0.092	0.0738 (0.0310)	1.08	0.017	0.0581 (0.0205)	1.06	0.0045

a: Combined results for fathers and mothers using inverse-variance meta-analysis on the effect sizes.

Table S5. Testing for change in polygenic scores with age of the GERA participants.

Trait	Beta (s.e.) ($\times 10^4$) [a]	*P* value
Puberty timing	3.7 (1.4)	0.0067
AFB	-0.69 (0.60)	0.25
ATH	-0.41 (1.0)	0.70
BMI	-0.26 (0.32)	0.42
CAD	-1.1 (0.60)	0.062
HDL	0.042 (0.81)	0.96
LDL	-0.74 (0.74)	0.32
TC	-1.1 (0.77)	0.15

a: Linear regression slope coefficient (polygenic score per year).