Supplementary figures and tables for:

Identifying genetic variants that affect viability in large cohorts

Hakhamanesh Mostafavi[1], Tomaz Berisa[2], Molly Przeworski[1,2*], Joseph K. Pickrell[1,3,*]

[1] Department of Biological Sciences, Columbia University, New York, NY, USA
[2] Department of Systems Biology, Columbia University, New York, NY, USA
[3] New York Genome Center, New York, NY, USA
*: These authors co-supervised this project.

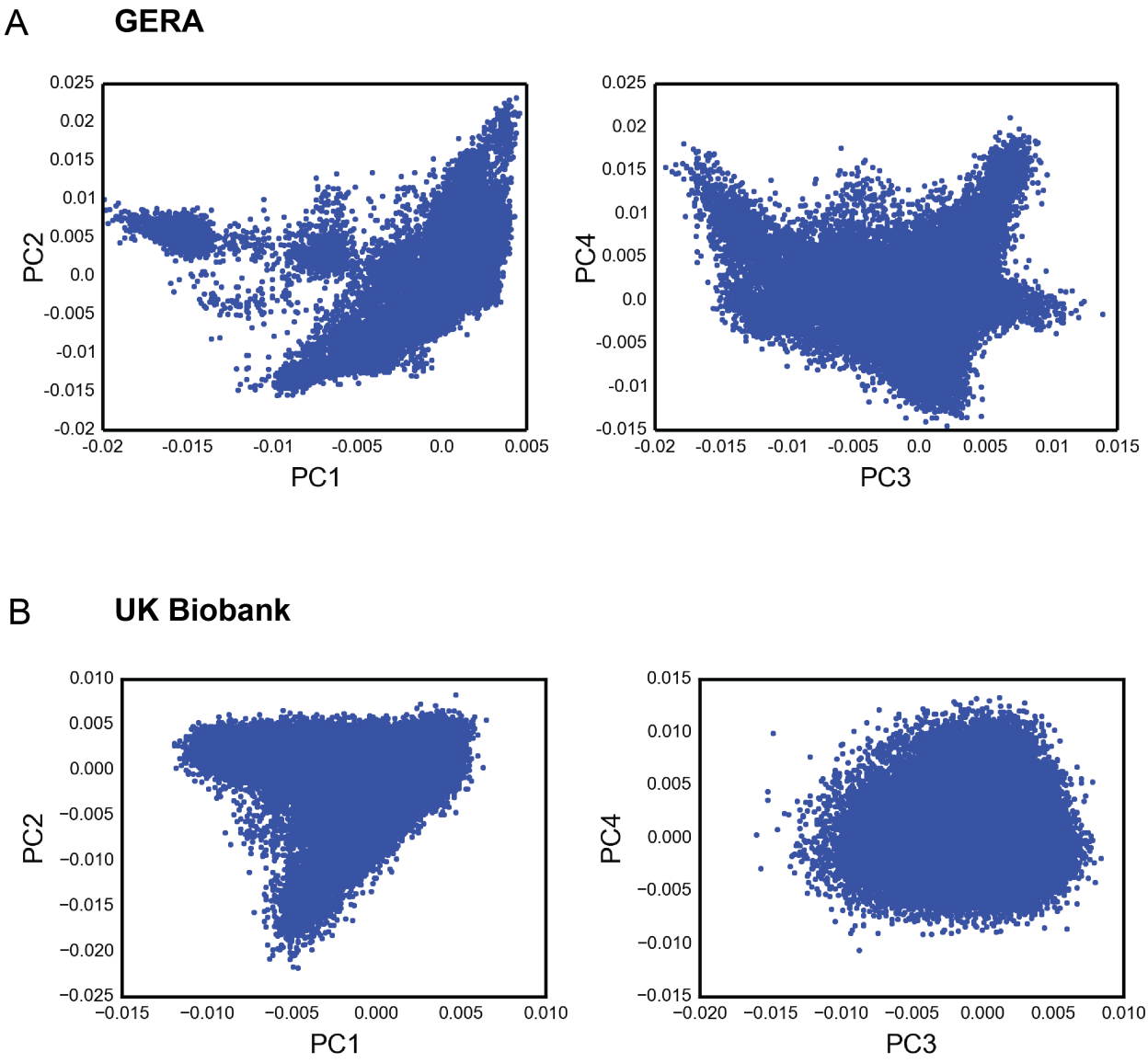


Figure S1. Results of principal component analysis (PCA). (A) PCA on 57,696 GERA individuals after quality control removing "non-European" individuals. (B) PCA on 120,286 UK Biobank participants of European ancestry. Result are in agreement with recent studies of these data (Galinsky, Bhatia, et al. 2016; Galinsky, Loh, et al. 2016).

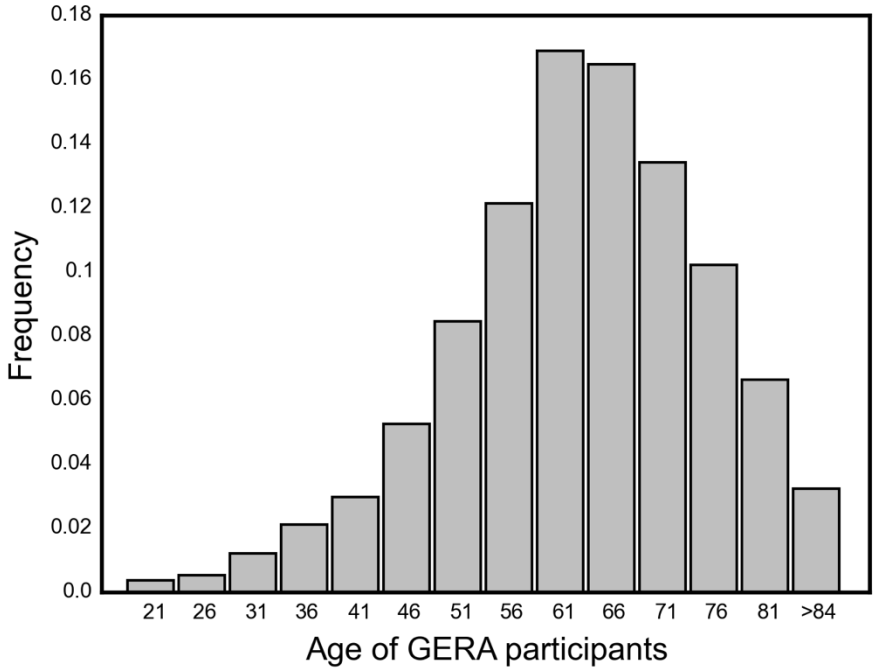


Figure S2. Age distribution of the GERA individuals. The labels on the x-axis indicate the center of 5-year interval age bins (except the last category).

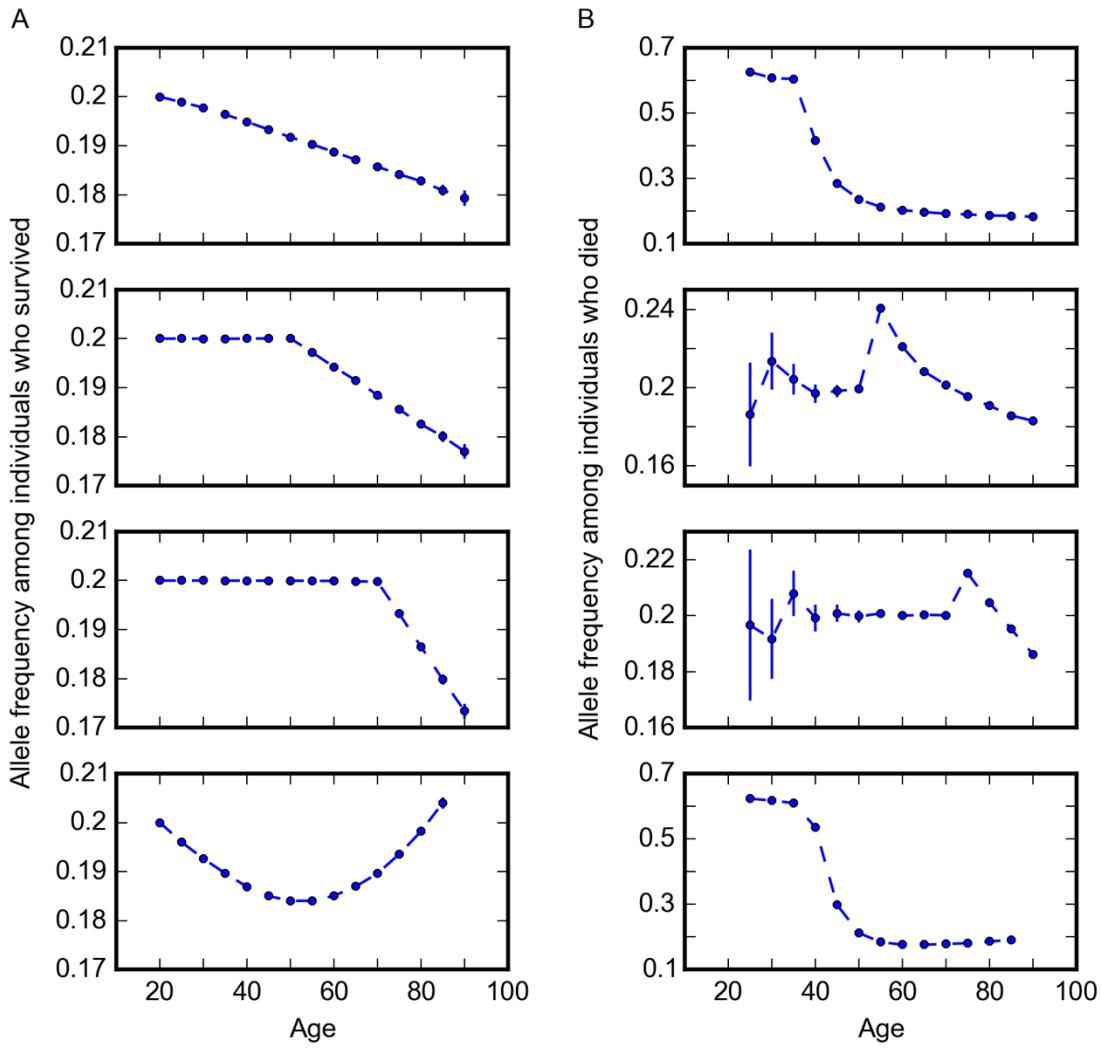


Figure S3. Comparison of trends in allele frequency with age and with age at death. (A) Simulated allele frequencies among surviving individuals reproducing trends as in Figure 1A. (B) Trends in allele frequency among individuals who died, corresponding to the trends in (A). Points are allele frequency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

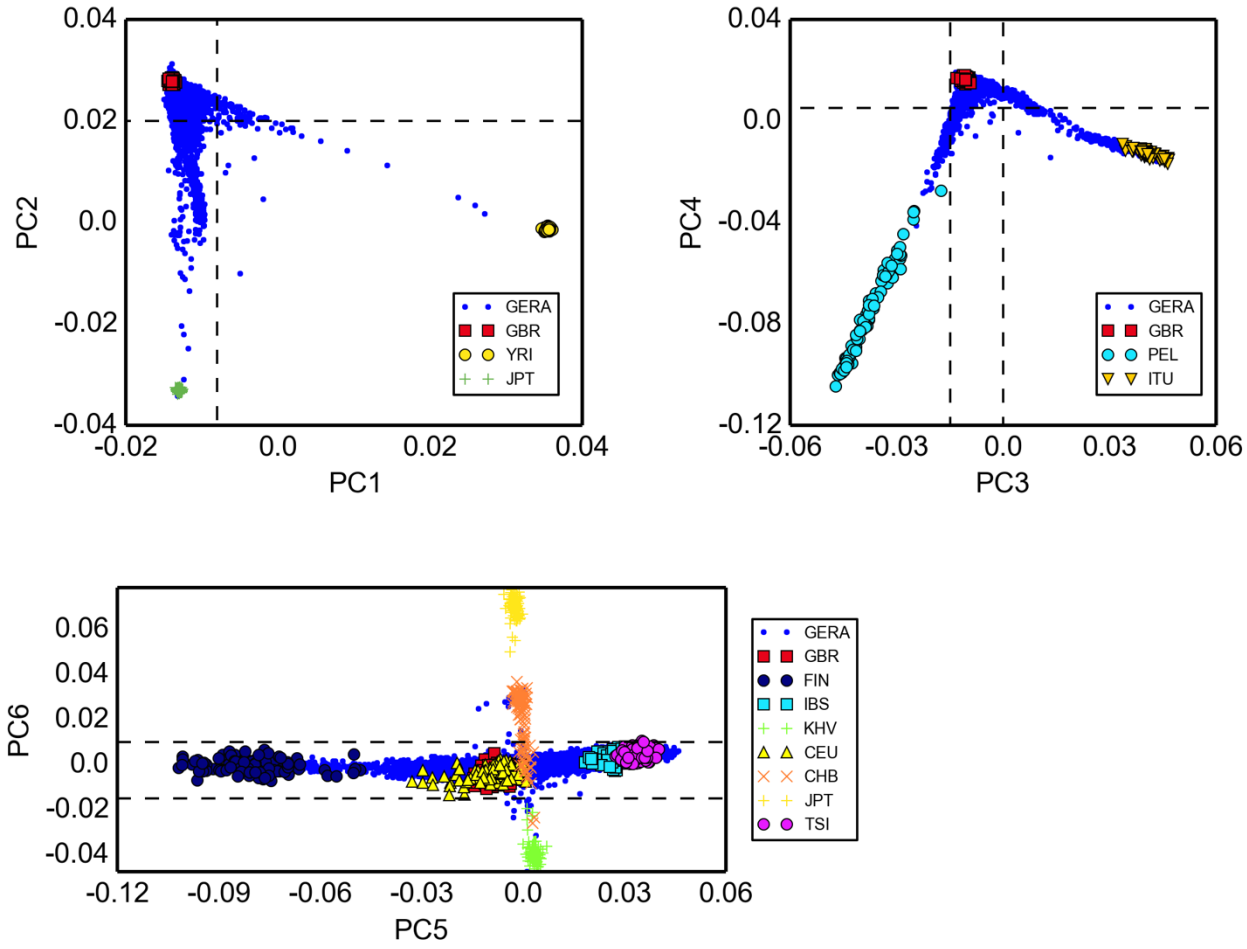


Figure S4. Validation of European ancestry in GERA. Shown are PCs inferred for all 26 populations in the 1000 Genomes Project phase 3 data. For clarity, in each plot, only few representative populations are shown. GERA individuals (blue dots) are projected on the inferred PCs. The dashed lines correspond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

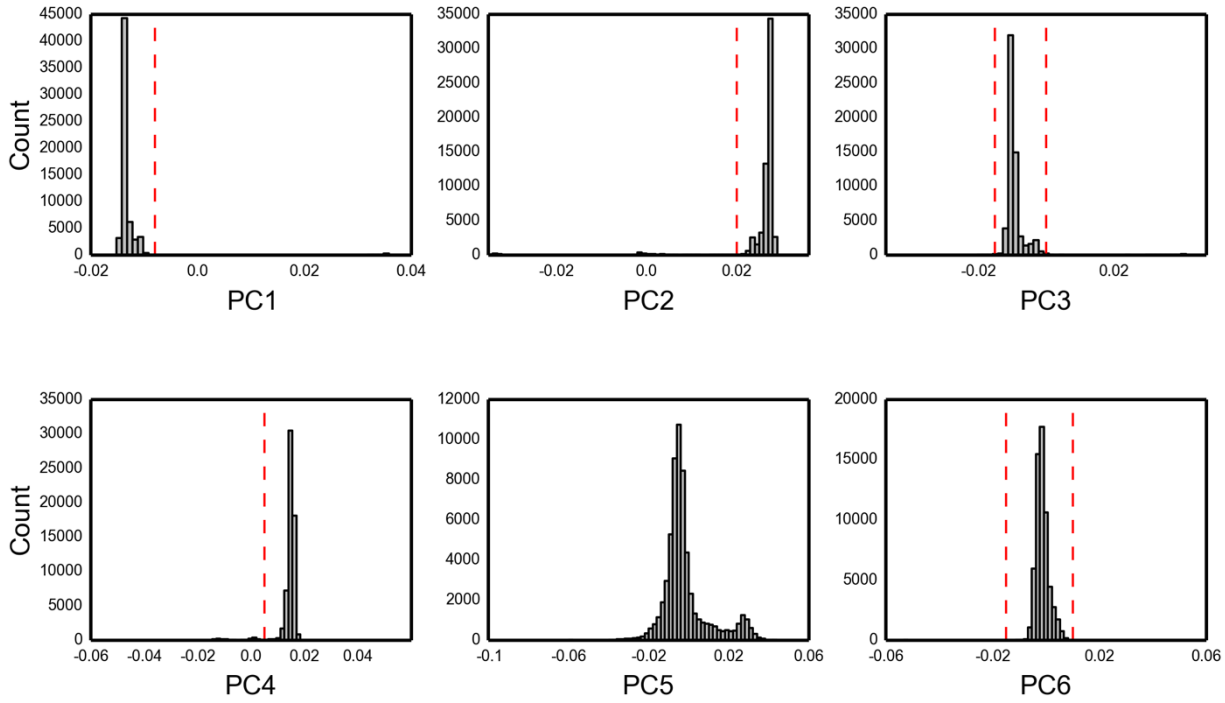


Figure S5. Distribution of GERA individuals for PCs inferred from 1000 Genome Project phase 3 data. The dashed lines enclose the majority of the data points; beyond, individuals were labeled as “non-Europeans”.

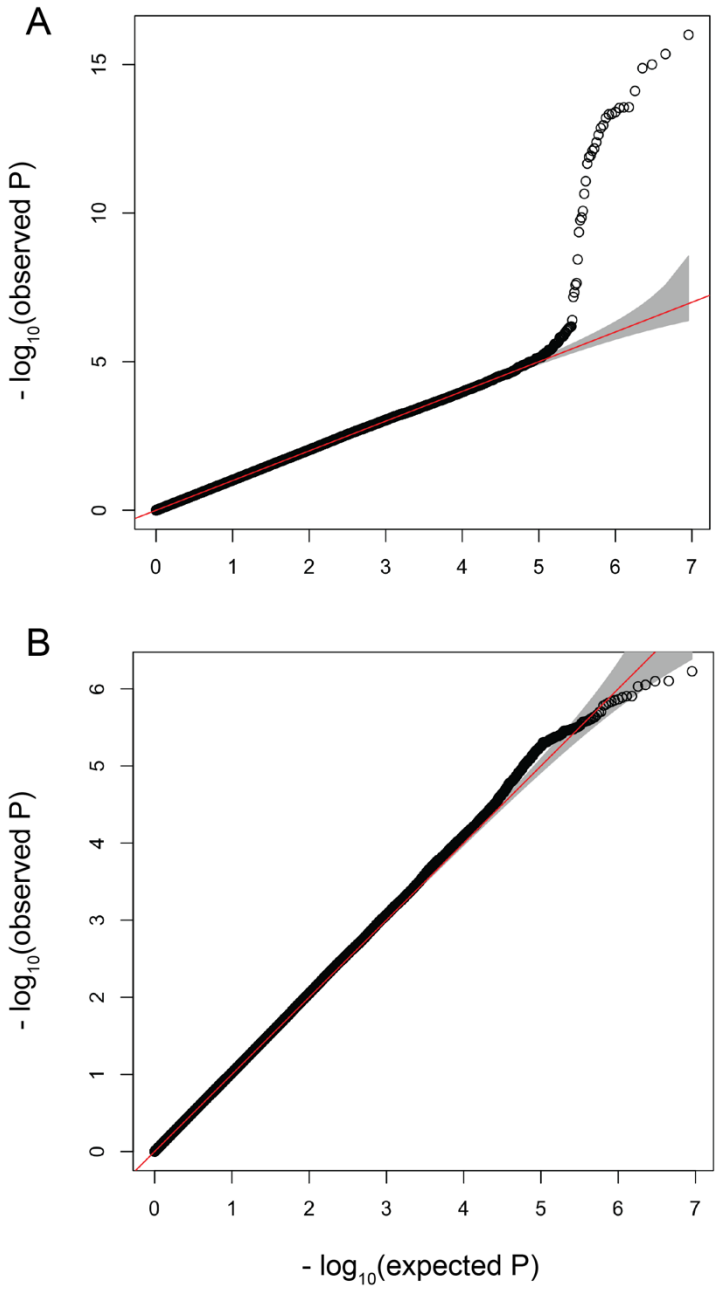


Figure S6. Quantile-quantile plots for model results for individual variants in GERA. Quantile-quantile plots for age (A), and age by sex (B) effects. The red lines indicate the distribution of the P values under the null (no age or age by sex effect), and the shaded bands represent the 95% confidence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

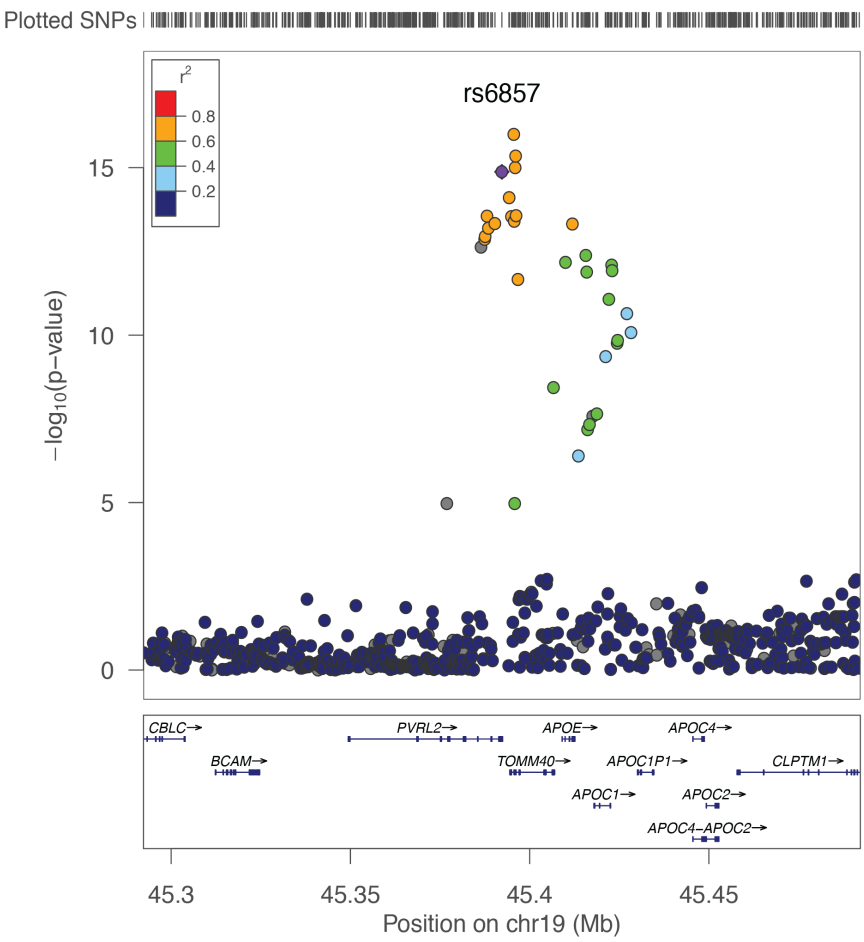


Figure S7. Regional plot for *APOE* locus. The y-axis shows P values obtained from a test of the influence of single genetic variants on age-specific mortality in GERA.

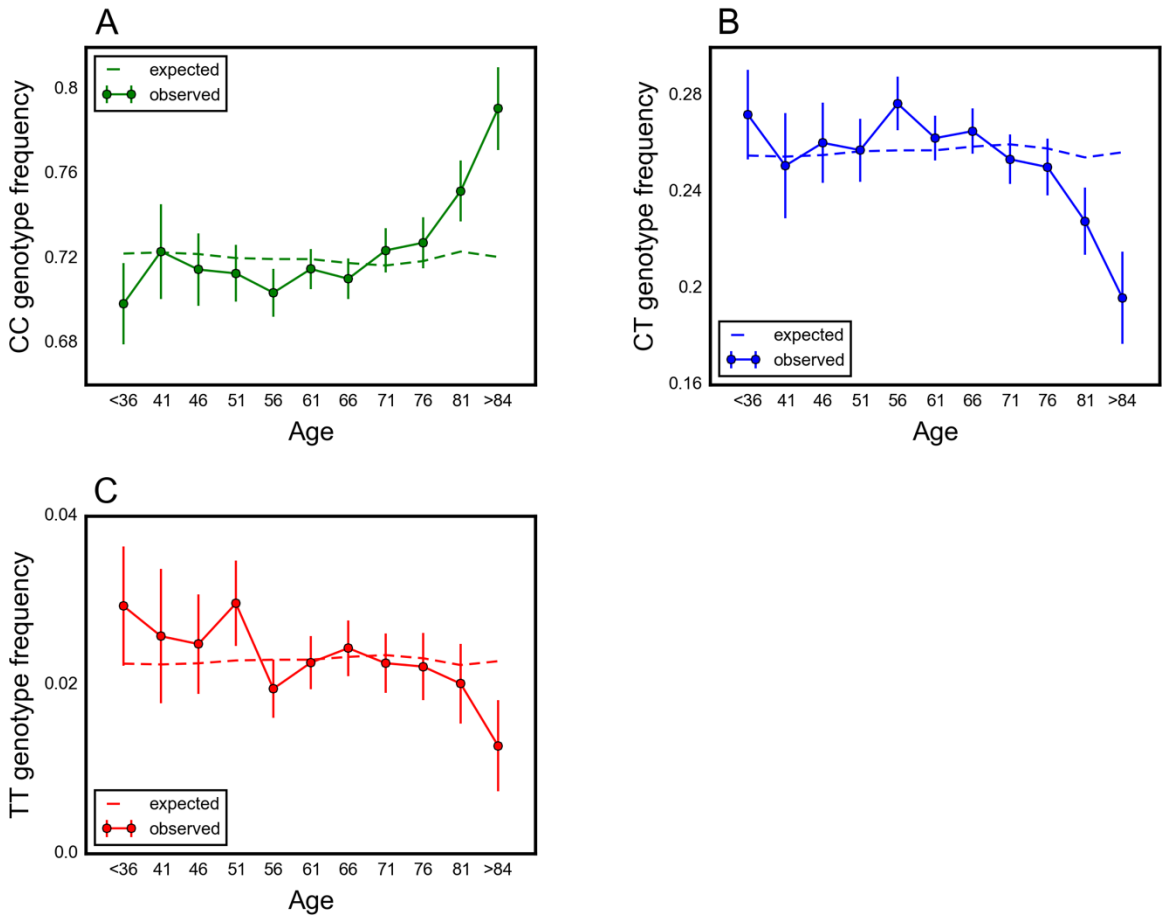


Figure S8. Frequency of rs6857 genotypes with age in GERA. Frequency of non-carriers (A), heterozygous (B) and homozygous (C) carriers of the risk allele for rs6857, tagging the ε4 allele of the APOE gene, across GERA age bins. Data points are frequency within 5-year interval age bins (mean and 95% confidence 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 baseline model, accounting for confounding batch effects and changes in ancestry.

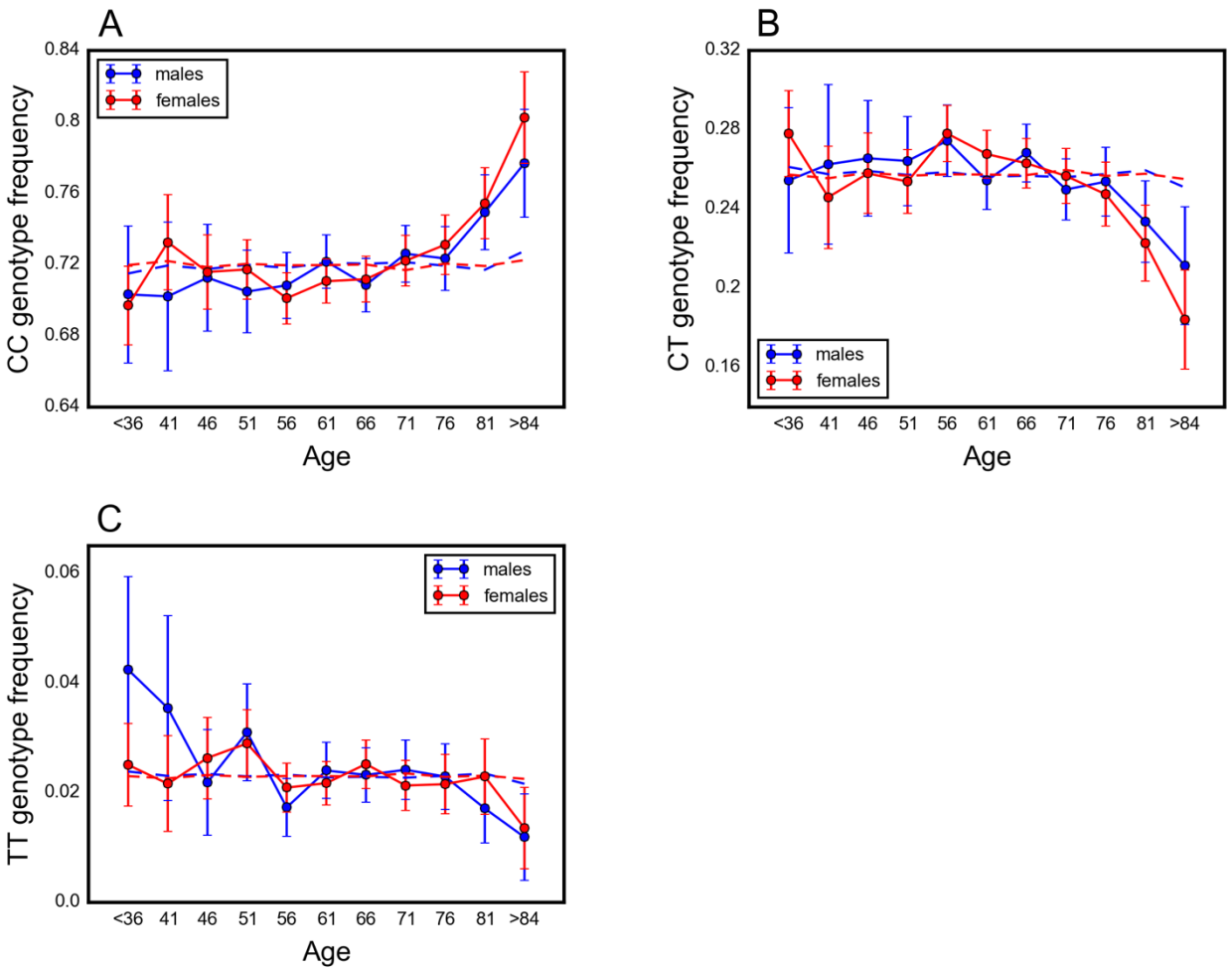


Figure S9. Frequency of rs6857 genotypes with age among males and females in GERA. Frequency of non-carriers (A), heterozygous (B) and homozygous (C) carriers of the risk allele for rs6857, tagging the ε4 allele of the *APOE* gene, across GERA age bins. Data points are frequency within 5-year interval 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 baseline model, accounting for confounding batch effects and changes in ancestry.

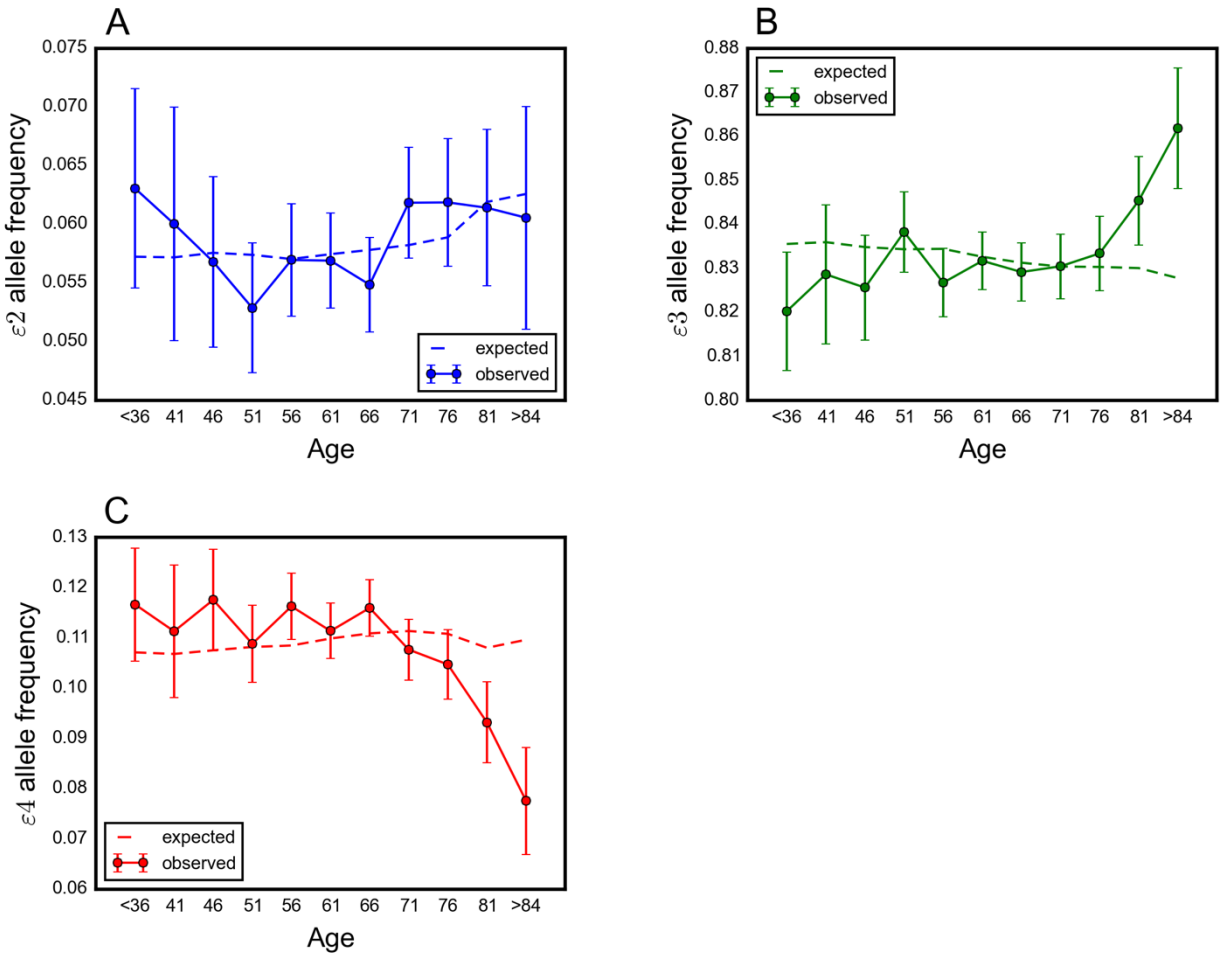


Figure S10. Frequency of the *APOE* gene alleles with age in GERA. Frequency of the ε2 (A), ε3 (B) and ε4 (C) across GERA age bins. Data points are frequency within 5-year interval age bins (mean and 95% confidence interval), with the center of the bin indicated on the x-axis. Bins with ages below 36 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 baseline model, accounting for confounding batch effects and changes in ancestry.

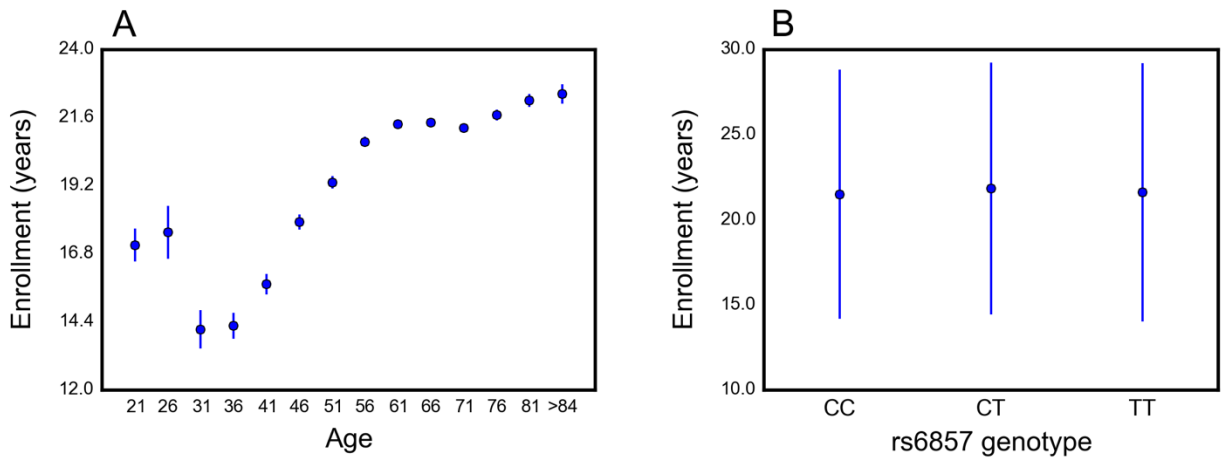


Figure S11. Enrollments of individuals in the Kaiser Permanente Medical Insurance Plan. (A) Years enrolled in the insurance plan (mean and 95% confidence interval) per age bin. The x-axis indicates the center of 5-year interval age bins. (B) Years enrolled in the insurance plan (mean and 95% confidence interval) for individuals of >70 years old versus the rs6857 genotype that they carry.

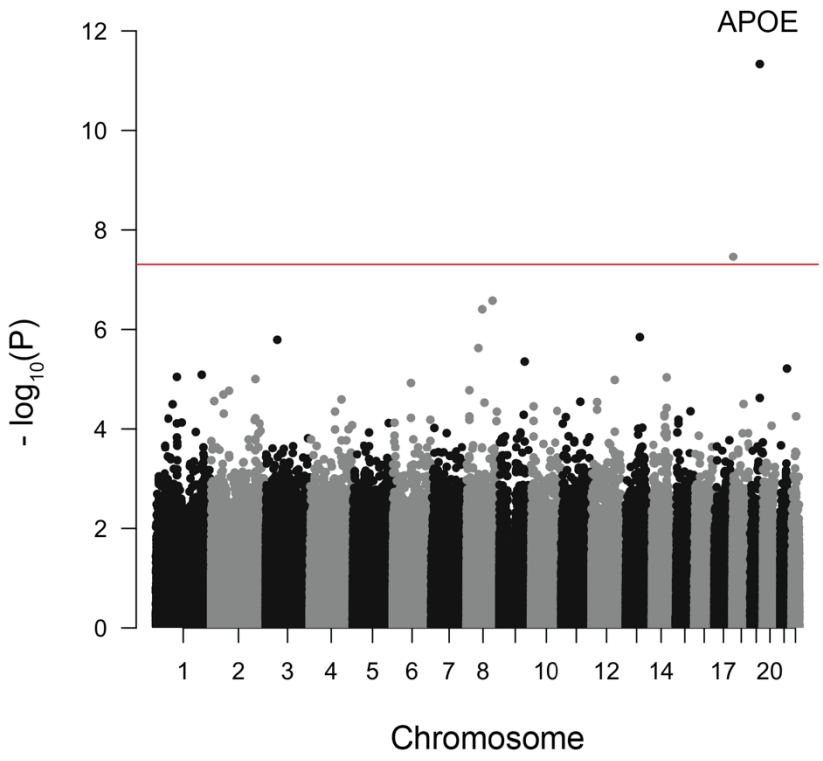


Figure S12. Testing for the influence of single genetic variants on age-specific mortality in the GERA cohort. A. Manhattan plot for change in allele frequency with age P values using the version of the model with age treated as an ordinal variable. The plot only includes the filtered genotyped SNPs in the GERA study. Red line marks the $P = 5{\times}10^{-8}$ 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 was lost when the variant was imputed using a leave-one-out approach.

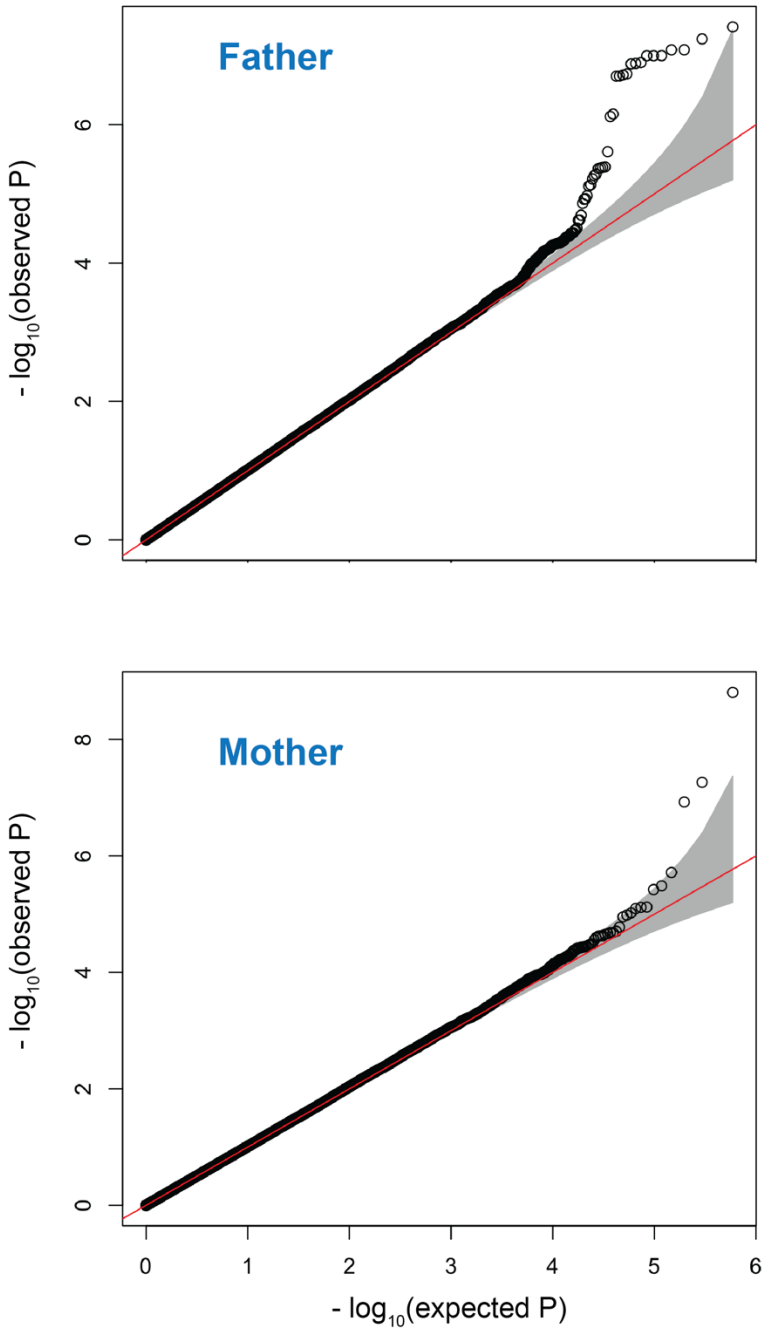


Figure S13. Quantile-quantile plots for model results for individual variants in the UK Biobank. Quantile-quantile plots for significant change in allele frequency with father's (A) and mother's (B) age at death. The red lines indicate distribution of the P values under the null (no change in frequency), and the shaded bands represent the 95% confidence intervals.

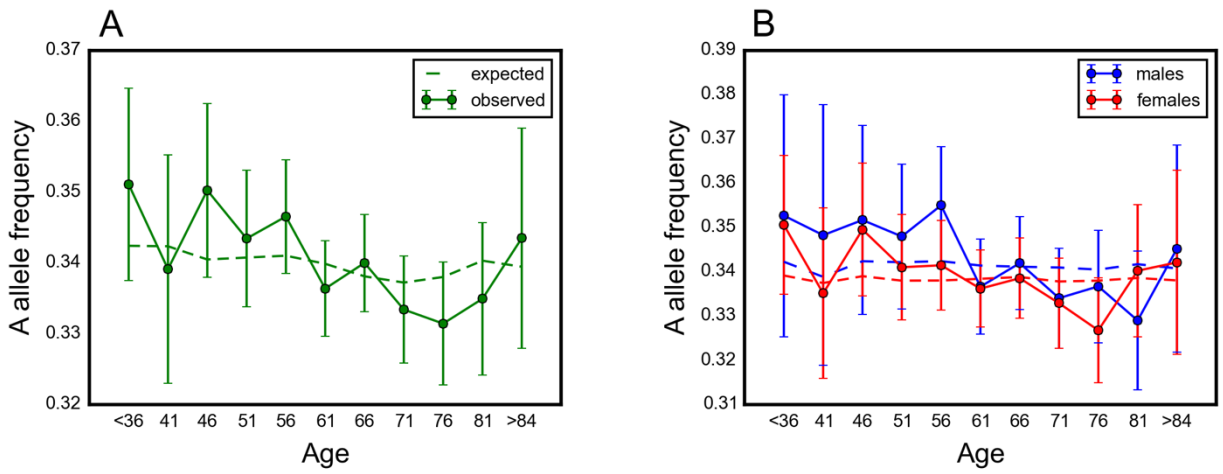


Figure S14. No significant effect of rs1051730 on survival in GERA. Allele frequency trajectory of rs1051730 with age for males and females together (A) and separately (B). The data points are the mean frequency within 5-year interval age bins and 95% confidence interval. The x-axis indicates the center of the age bin. The dashed line shows the expected frequency based on the baseline model, accounting for confounding batch effects and changes in ancestry.

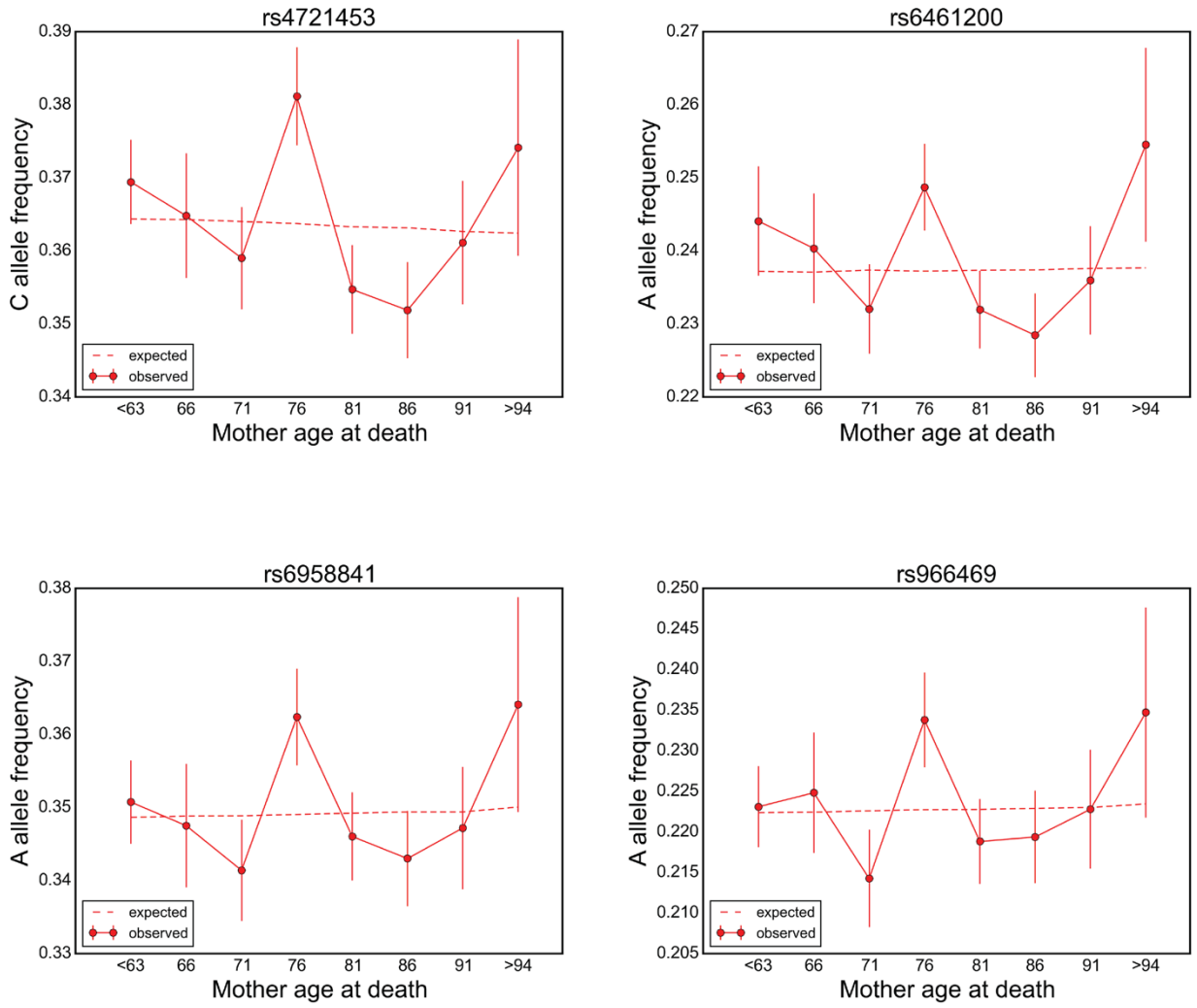


Figure S15. Allele frequency of variants in the *MEOX2* locus with mother's age at death in the UK Biobank. Plots are for four genotyped SNPs in moderate linkage disequilibrium with $P < 10^{-4}$ for the change in allele frequency with mother's age at death. Data points are frequency within 5-year interval age bins 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 baseline model, accounting for confounding batch effects and changes in ancestry.

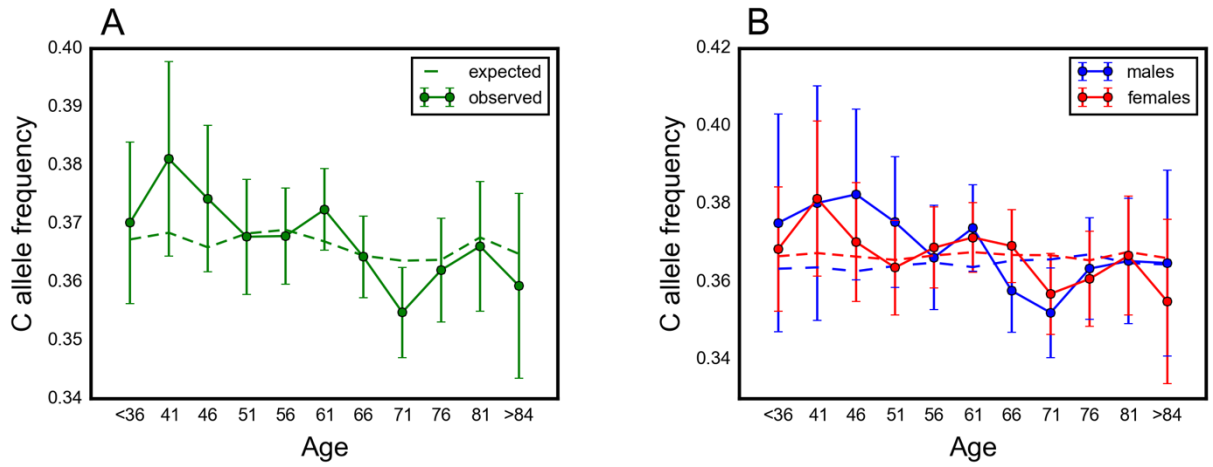
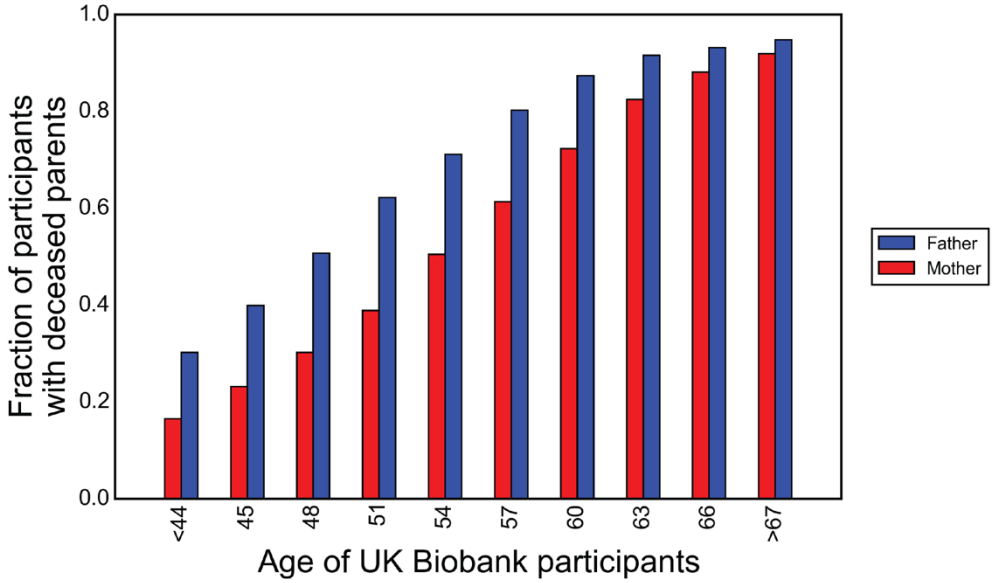
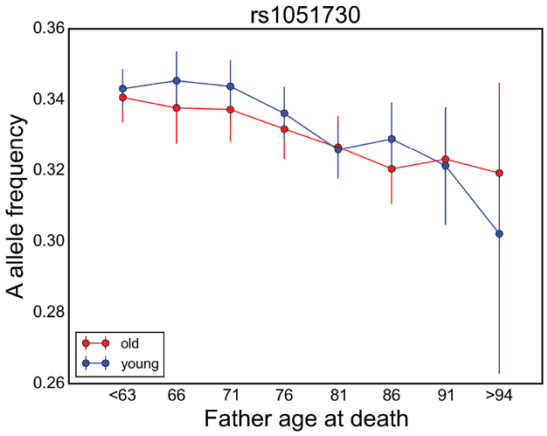


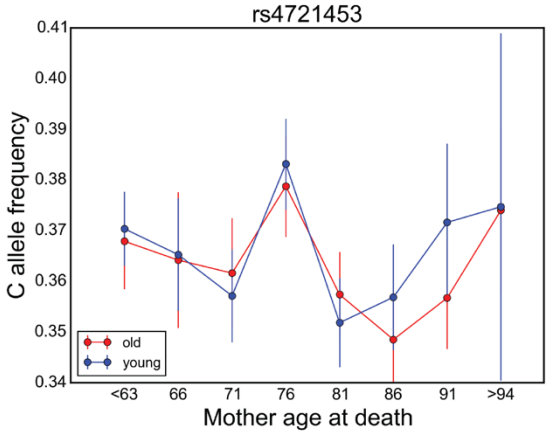
Figure S16. No significant effect of rs4721453 on survival in GERA. Allele frequency trajectory of rs4721453 with age for males and females together (A) and separately (B). The data points are the mean frequency within 5-year interval age bins and 95% confidence interval. The x-axis indicates the center of the age bin. The dashed line shows the expected frequency based on the baseline model, accounting for confounding batch effects and changes in ancestry.

A


Fraction of participants with deceased parents
1.0
0.8
0.6
0.4
0.2
0.0
<44
45
48
51
54
57
60
63
66
>67
Age of UK Biobank participants
Father
Mother

B


rs1051730
A allele frequency
0.36
0.34
0.32
0.30
0.28
0.26
old
young
<63
66
71
76
81
86
91
>94
Father age at death


rs4721453
C allele frequency
0.41
0.40
0.39
0.38
0.37
0.36
0.35
0.34
old
young
<63
66
71
76
81
86
91
>94
Mother age at death

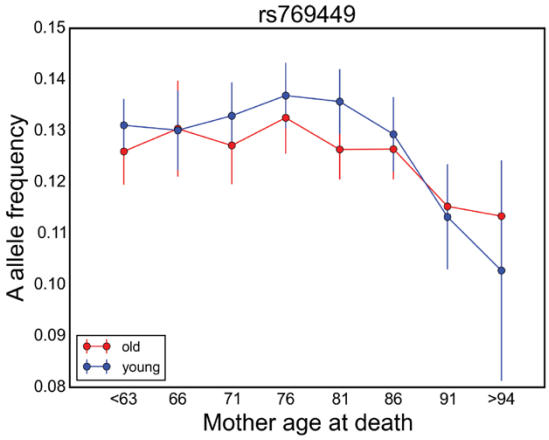

rs769449
A allele frequency
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
old
young
<63
66
71
76
81
86
91
>94
Mother age at death

Figure S17. The top signals for change in allele frequency with age at death of the UK Biobank participants are not affected by ascertainment bias towards older participants. (A) Fraction of the participants in each age bin (bin size of 3 years) who reported father's or mother's age at death. (B) Allele frequency of SNPs with strongest age effects as a function of parental age at deaths, conditioned on the age of the participants. Old and young labels refer to individuals with age ≥ 62 (last three age categories in panel A) and ≤ 61, respectively. The data points are the mean frequency within 5-year interval age bins and 95% confidence interval. The x-axis indicates the center of the age bin. The dashed line shows the expected frequency based on the baseline model, accounting for confounding batch effects and changes in ancestry.

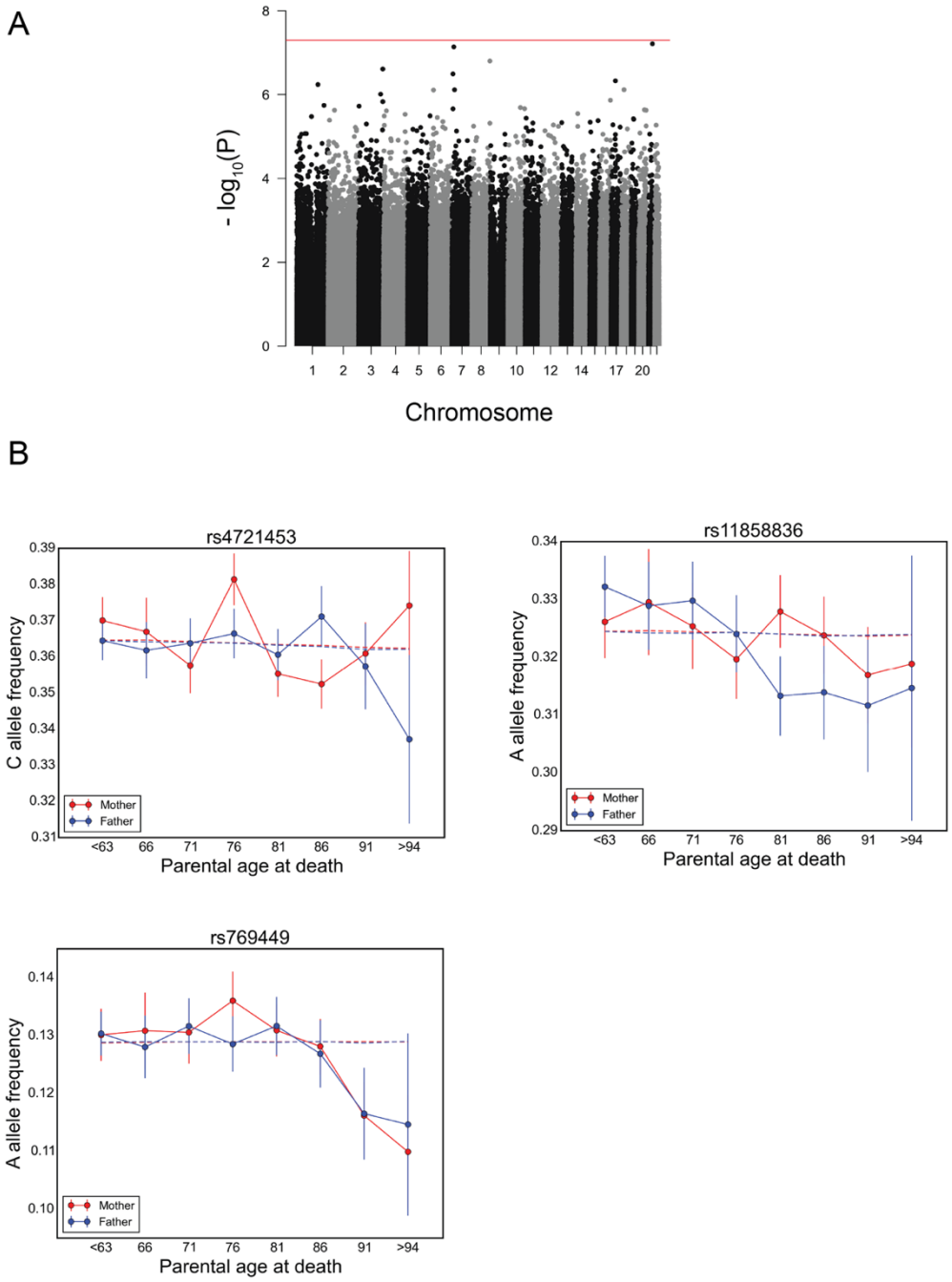


Figure S18. Testing for significant age by sex effect of individual genetic variants in the UK Biobank. (A) Manhattan plot for change in allele frequency with parental age at deaths that differ between fathers and mothers of the UK Biobank participants. (B) Allele frequency of SNPs with significant age effects as a function of father's and mother's age at deaths. The data points are the mean frequency within 5-year interval age bins and 95% confidence interval. The x-axis indicates the center of the age bin. The dashed line shows the expected frequency based on the baseline model, accounting for confounding batch effects and changes in ancestry.

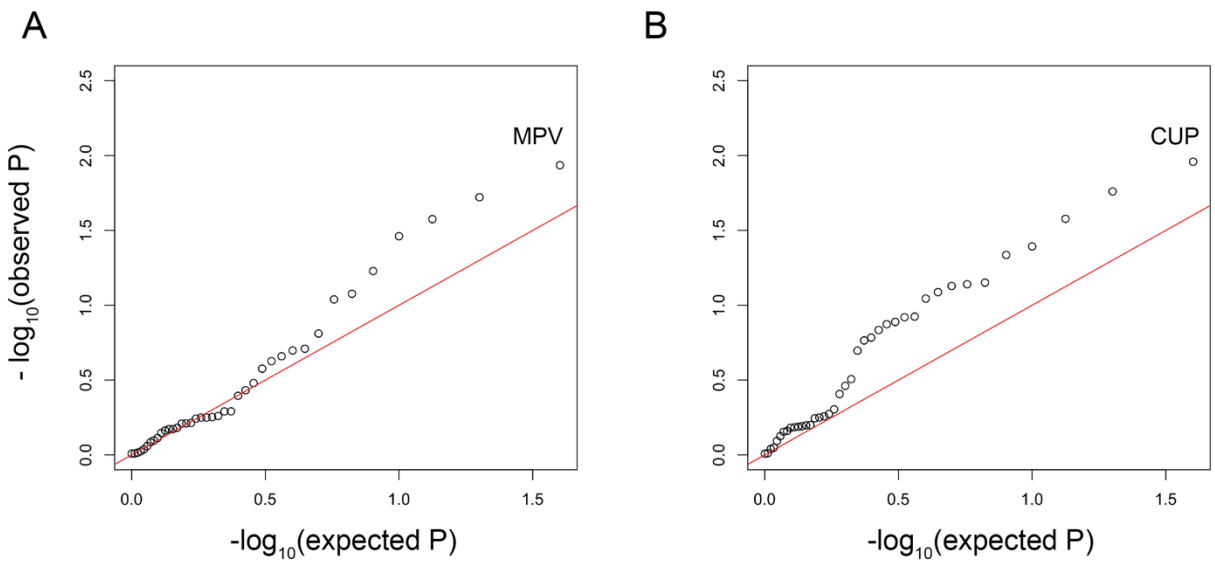


Figure S19. Testing for age by sex effect of set of trait-associated variants on age-specific mortality in the GERA cohort. Quantile-quantile plots for changes in polygenic score of 40 traits (see Table S1) with age that are different between males and females in the GERA cohort, treating age as a categorical (A) or an ordinal (B) variable. The red line indicates the distribution of the P values under the null (of no change in polygenic score).

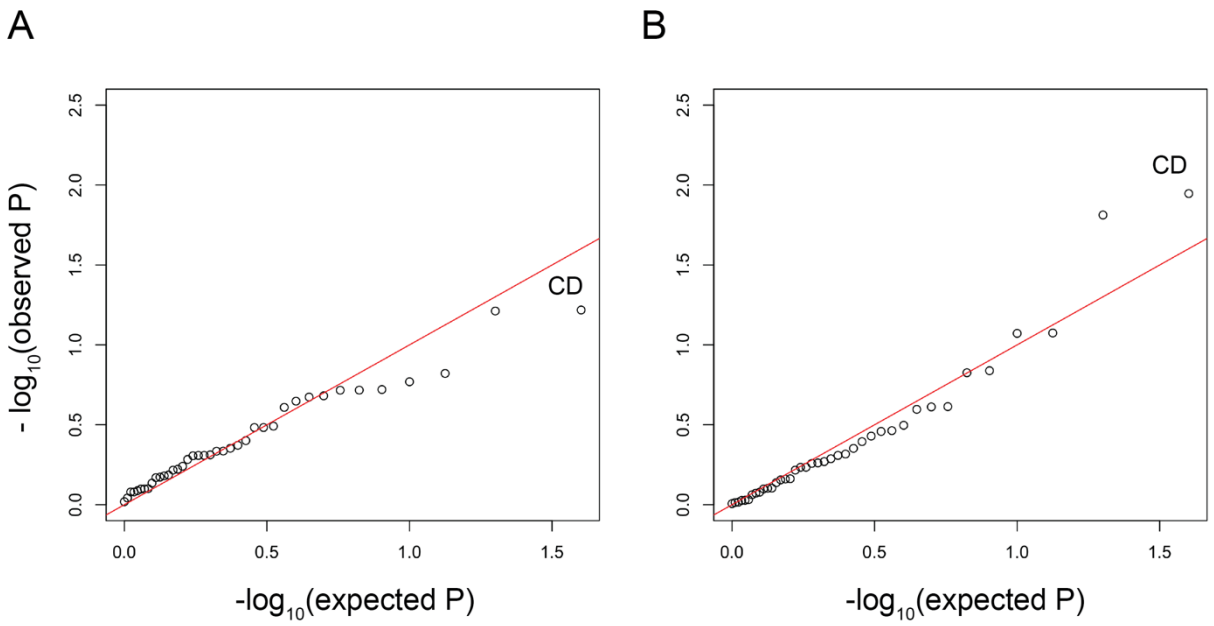


Figure S20. Testing for stabilizing selection on traits in the GERA cohort. Quantile-quantile plots for change in the squared difference of polygenic score from the population mean for 40 traits (see Table S1) with age in the GERA cohort, treating age as a categorical (A) or an ordinal (B) variable. The red line indicates distribution of the P values under the null.

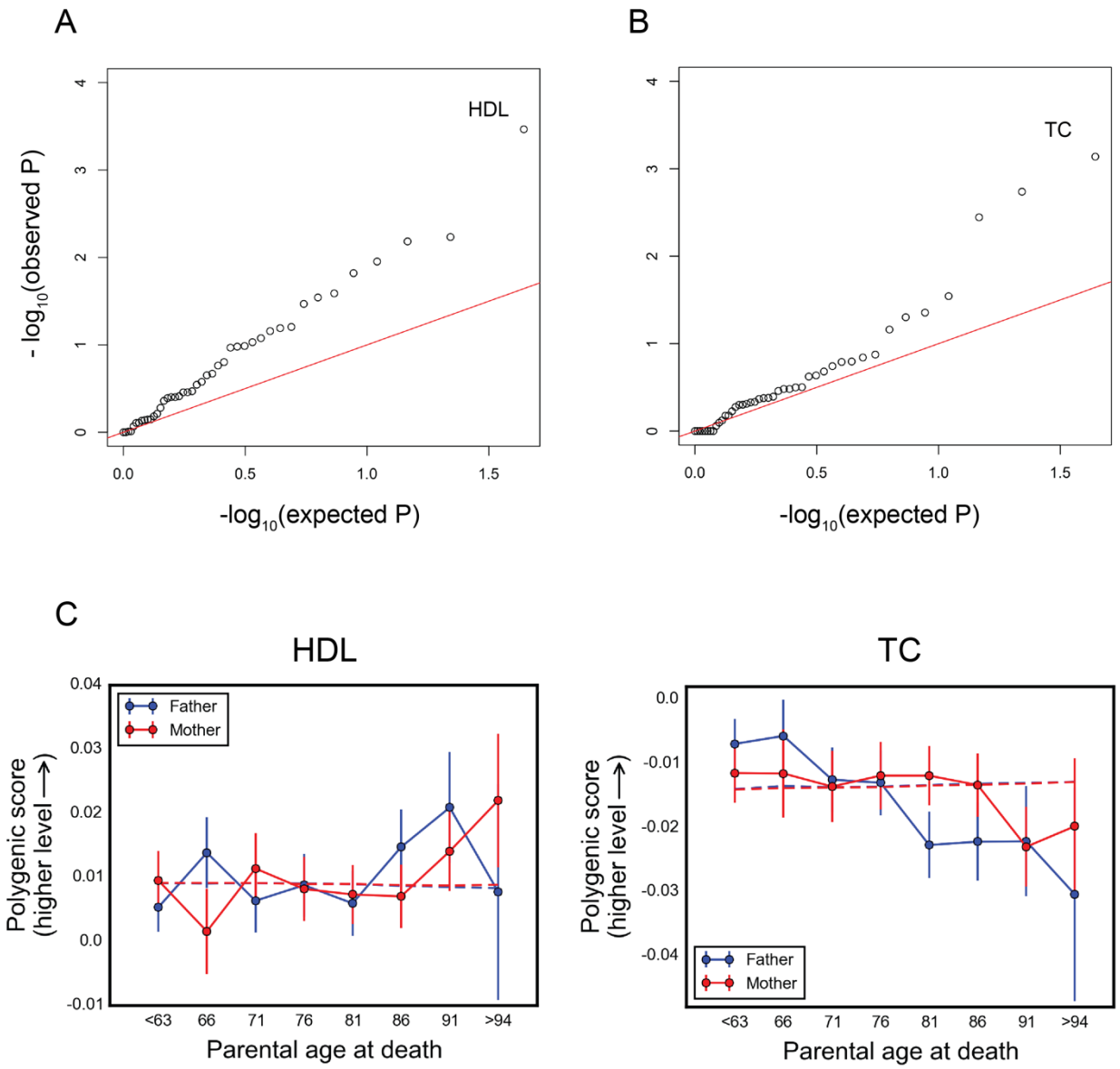


Figure S21. Testing for age by sex effect of set of trait-associated variants on age-specific mortality in the UK Biobank. Quantile-quantile plots for changes in polygenic score of 40 traits (see Table S1) with parental age at death that are different between fathers and mothers of the UK Biobank participants, treating age variables as categorical (A) or ordinal (B). The red lines indicate distribution of the P values under the null. (C) Trends in polygenic score with parental age at deaths for traits showing significant age by sex effect in the UK Biobank. The data points are the mean polygenic score within 5-year interval age bins and 95% confidence interval for high-density lipoproteins (left) and total cholesterol (right). 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 and changes in ancestry.

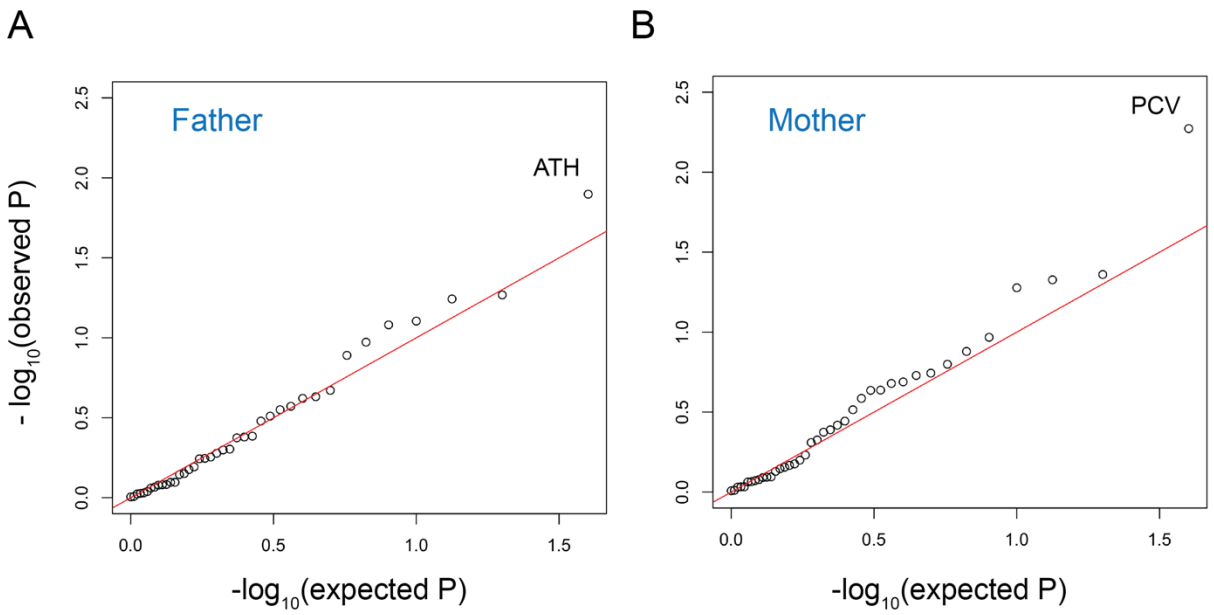


Figure S22. Testing for stabilizing selection on traits in the UK Biobank. Quantile-quantile plots for change in the squared difference of polygenic score from the population mean for 40 traits (see Table S1) with age at death of fathers (A) and mothers (B) of the UK Biobank participants. The red line indicates the distribution of the P values under the null.

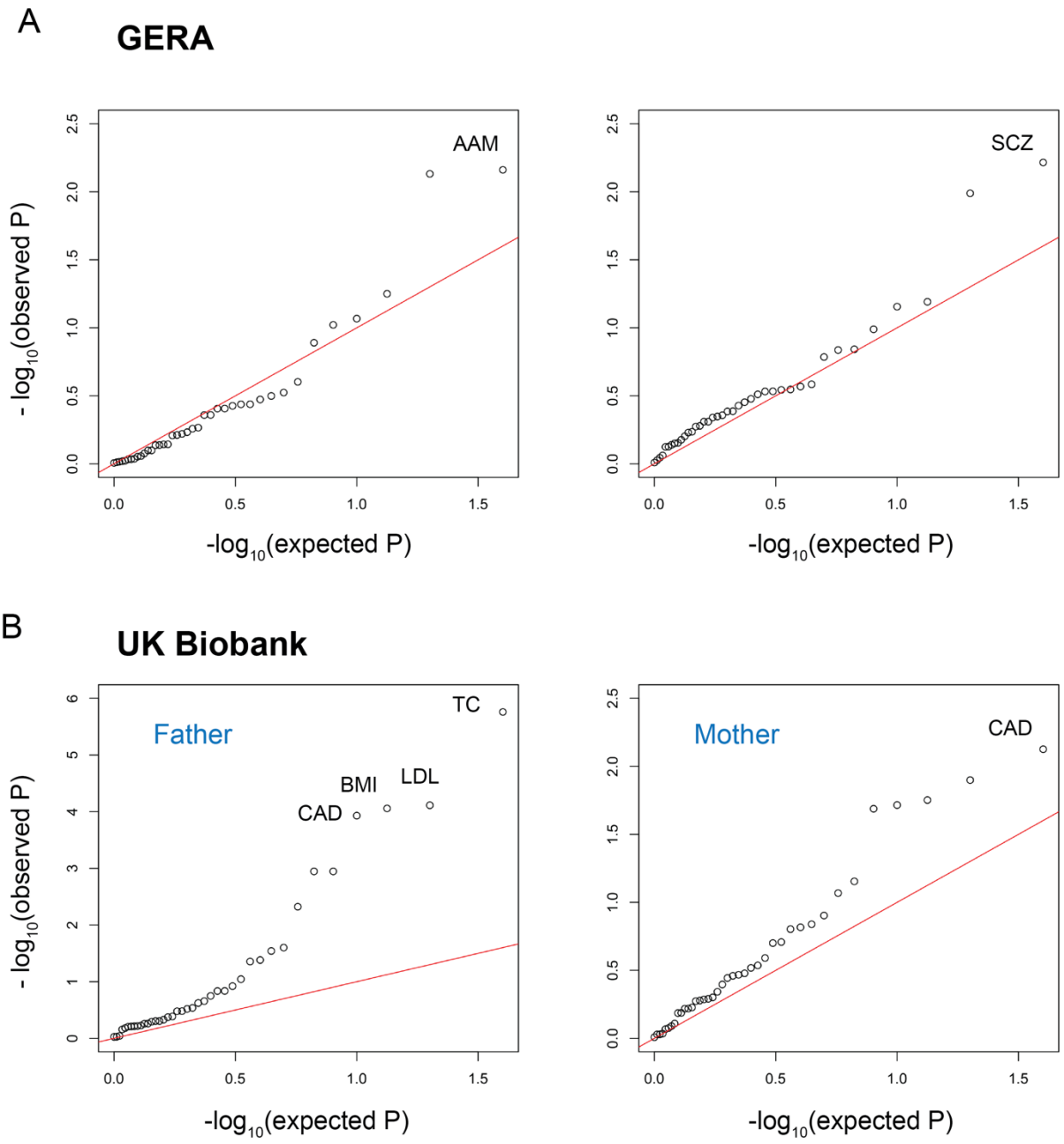


Figure S23. Testing for the influence of set of trait-associated variants on age-specific mortality using variants passing quality control steps in both GERA and UK Biobank datasets. Quantile-quantile plots for change in the polygenic score of 40 traits (see Table S1) with age in the GERA cohort (A), treating age as a categorical (left) or an ordinal (right) variable, and with age at death of fathers and mothers of the UK Biobank participants (B). The red line indicates distribution of the P values under the null (no change in polygenic score), and the shaded band represents the 95% confidence intervals.

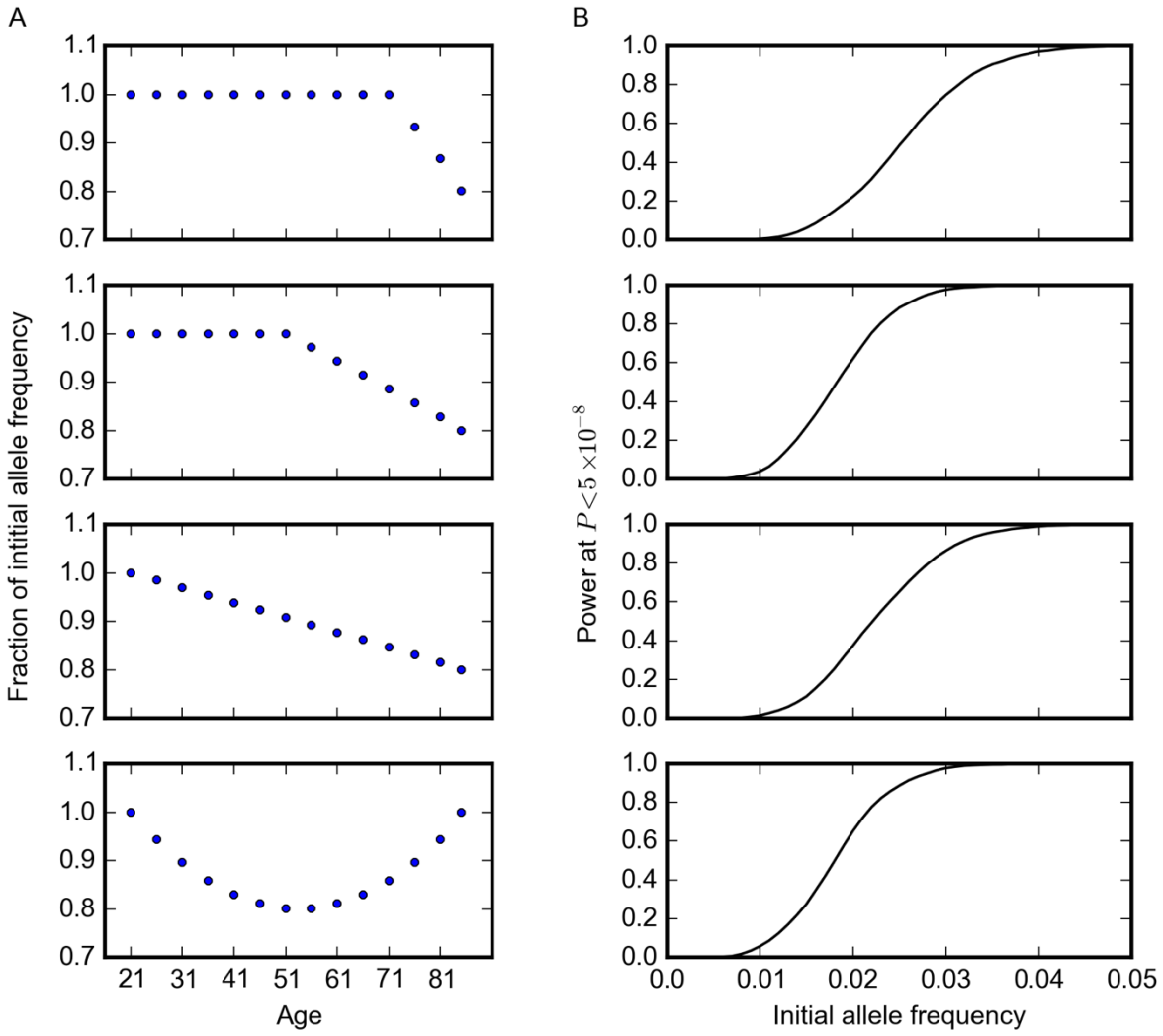


Figure S24. Power of the model to detect changes in allele frequency with age. Same as Figure 1, but with 500,000 samples evenly distributed among age categories, and only showing the results using models with age treated as a categorical variable.

Phenotype	Abbreviation	# loci in GERA	# loci in UK Biobank
Age at menarche	AAM	88	83
Age at voice drop	AVD	5	5
Alzheimer's disease	AD	9	4
Any allergies	ALL	38	35
Asthma	ATH	32	33
Beighton hypermobility	BHM	17	17
Body mass index	BMI	30	30
Bone mineral density (femoral neck)	FNBMD	19	18
Bone mineral density (lumbar spine)	LSBMD	20	20
Breast size	CUP	14	14
Childhood ear infections	CEI	13	13
Chin dimples	DIMP	54	52
Coronary artery disease	CAD	10	11
Crohn's disease	CD	57	58
Fasting glucose	FG	15	15
Height	HEIGHT	559	561
Hemoglobin	HB	15	15
High-density lipoproteins	HDL	46	45
Hypothyroidism	HTHY	27	26
Low-density lipoproteins	LDL	39	39
Male pattern baldness	MPB	47	44
Mean cell hemoglobin concentration	MCHC	14	15
Mean platelet volume	MPV	27	29
Mean red cell volume	MCV	40	41
Migraine	MIGR	30	29
Nearsightedness	NST	166	159
Nose size	NOSE	11	10
Packed red cell volume	PCV	12	12
Parkinson's disease	PD	37	23
Photic sneeze reflex	PS	61	60
Platelet count	PLT	48	49
Red blood cell count	RBC	21	22
Rheumatoid arthritis	RA	68	68
Schizophrenia	SCZ	195	191
Tonsillectomy	TS	38	38
Total cholesterol	TC	50	49
Triglycerides	TG	30	28
Type 2 diabetes	T2D	11	11
Unibrow	UB	54	53
Waist-hip ratio	WHR	13	12

Table S1. List of phenotypes. All phenotypes and associated variants from Pickrell et al. (Pickrell et al. 2016). The numbers of SNPs passing quality control measures are shown for each dataset.

Phenotype	*P* value GERA	*P* value UK Biobank (father)	*P* value UK Biobank (mother)
AAM	0.0034	0.24	0.65
AD	0.15	0.73	0.99
ALL	0.78	0.4	0.3
ATH	0.12	0.0022	0.92
AVD	0.36	0.22	0.35
BHM	0.94	0.73	0.21
BMI	0.13	8.8×10^{-5}	0.16
CAD	0.62	7.5×10^{-6}	0.0071
CD	0.87	0.59	0.27
CEI	0.45	0.88	0.18
CUP	0.54	0.3	0.52
DIMP	0.35	0.61	0.15
FG	0.32	0.041	0.12
FNBMD	0.76	0.55	0.07
HB	0.6	0.088	0.38
HDL	0.59	0.0011	0.019
HEIGHT	0.41	0.33	0.44
HTHY	0.95	0.26	0.59
LDL	0.48	1.5×10^{-5}	0.011
LSBMD	0.6	0.29	0.65
MCHC	0.55	0.81	0.29
MCV	0.88	0.51	0.76
MIGR	0.009	0.89	0.44
MPB	0.37	0.12	0.02
MPV	0.3	0.49	0.92
NOSE	0.98	0.94	0.29
NST	0.88	0.55	0.89
PCV	0.73	0.24	0.31
PD	0.13	0.18	0.62
PLT	0.9	0.52	0.84
PS	0.96	0.02	0.013
RA	0.92	0.12	0.1
RBC	0.78	0.88	0.062
SCZ	0.1	0.57	0.3
T2D	0.97	0.044	0.78
TC	0.52	2.7×10^{-7}	0.095
TG	0.58	0.0048	0.5
TS	0.74	0.07	0.3
UB	0.39	0.81	0.86
WHR	0.97	0.49	0.6

Table S2. Model results on quantitative traits. Shown are *P* values for change in polygenic score with age in GERA and with age at death of parents of the UK Biobank participants, with models treating all age variables as categorical.