

1 **Genetic associations with human longevity are enriched for**
2 **oncogenic genes**

3
4

5 Junyoung Park, PhD¹, Andrés Peña-Tauber, BA¹, Lia Talozzi, PhD¹, Michael D. Greicius,
6 MD^{1*}, Yann Le Guen, PhD^{2*}

7
8

9 **Affiliations**

10 ¹Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA,
11 94305, USA.

12 ²Quantitative Sciences Unit, Department of Medicine, Stanford University, Stanford, CA,
13 94304, USA.

14
15

16 **Corresponding authors:**

17 Junyoung Park
18 Stanford Neuroscience Health Center
19 290 Jane Stanford Way,
20 Stanford, CA 94305-5090
21 jpark01@stanford.edu
22 (650) 666-2696

23
24

25 * **These authors contributed equally**

26 **Abstract**

27

28 Human lifespan is shaped by both genetic and environmental exposures and their interaction.
29 To enable precision health, it is essential to understand how genetic variants contribute to
30 earlier death or prolonged survival. In this study, we tested the association of common genetic
31 variants and the burden of rare non-synonymous variants in a survival analysis, using age-at-
32 death (N = 35,551, median [min, max] = 72.4 [40.9, 85.2]), and last-known-age (N = 358,282,
33 median [min, max] = 71.9 [52.6, 88.7]), in European ancestry participants of the UK Biobank.
34 The associations we identified seemed predominantly driven by cancer, likely due to the age
35 range of the cohort. Common variant analysis highlighted three longevity-associated loci:
36 *APOE*, *ZSCAN23*, and *MUC5B*. We identified six genes whose burden of loss-of-function
37 variants is significantly associated with reduced lifespan: *TET2*, *ATM*, *BRCA2*, *CKMT1B*,
38 *BRCA1* and *ASXL1*. Additionally, in eight genes, the burden of pathogenic missense variants
39 was associated with reduced lifespan: *DNMT3A*, *SF3B1*, *CHL1*, *TET2*, *PTEN*, *SOX21*, *TP53*
40 and *SRSF2*. Most of these genes have previously been linked to oncogenic-related pathways
41 and some are linked to and are known to harbor somatic variants that predispose to clonal
42 hematopoiesis. A direction-agnostic (SKAT-O) approach additionally identified significant
43 associations with *C1orf52*, *TERT*, *IDH2*, and *RLIM*, highlighting a link between telomerase
44 function and longevity as well as identifying additional oncogenic genes.
45 Our results emphasize the importance of understanding genetic factors driving the most
46 prevalent causes of mortality at a population level, highlighting the potential of early genetic
47 testing to identify germline and somatic variants increasing one's susceptibility to cancer
48 and/or early death.

49 Introduction

50 Longevity is a complex trait influenced by both genetic and environmental factors and their
51 interactions [1]. According to previous studies, genetics accounts for as much as 40% of the
52 heritability of longevity [2-4]. Identifying the genetic variants that contribute to earlier death
53 or prolonged survival can highlight key biological pathways linked to lifespan and inform
54 genetic testing for general health and screening and enabling precision health. Previous
55 genome-wide association studies (GWAS) have identified over 20 associated loci including
56 *APOE* [5, 6], *CHRNA3/5* [7], *HLA-DQAI* and *LPA* [8]. Recently, a burden analysis of protein-
57 truncating variants from whole-exome sequencing (WES) data identified four additional genes
58 (*BRCA2*, *BRCA1*, *ATM*, and *TET2*) linked to reduced lifespan [9]. However, most previous
59 research on lifespan genetics has predominantly used proxy data, such as parents' age at death,
60 due to a lack of proband lifespan data. While proxy-based GWAS have been necessary for
61 large cohorts of primarily middle-aged individuals with limited mortality data, this approach
62 restricts the accuracy and scope of findings, as it may fail to comprehensively capture the
63 genetic influences that directly impact an individual's lifespan [10]. On the other hand, some
64 studies have employed logistic regression models on cases of extreme longevity and younger
65 controls [11-13]. This approach may offer new insights by focusing on exceptionally long-
66 lived individuals, yet they can be limited and costly. Moreover, replication of borderline
67 significant variants remains an issue due to varying case definitions across studies, with some
68 defining cases as individuals who survive to ages beyond 90 or 100 years or using the 90th or
69 99th survival percentiles as the age cutoff.

70 In this study, we carried out a genetic analysis of direct mortality data in the UK Biobank, the
71 genetic database with the largest number of reported deaths (35,551 subjects) and aged
72 individuals (344,237 subjects over 60 years old). To assess the association of genetic variants
73 with longevity in a survival analysis, we performed GWAS of common variants imputed from
74 microarray data as well as burden/sequence kernel association test-optimized (SKAT-O)
75 association of rare non-synonymous variants from WES data.

77 Results

79 Genome-wide association analyses in imputed array data

80 Our GWAS assessed 6,127,227 common variants (minor allele frequency (MAF) $\geq 1\%$) using
81 Martingale residuals on 393,833 individuals including 35,551 deceased subjects (mean age at
82 death: 71.2 years) and 358,282 living subjects (mean current age: 70.7) from UKB
83 (Supplementary Table 1) [14]. Two loci reached genome-wide significance (GWS) ($p <$
84 5.0×10^{-8}) on chromosomes 19 and 6 (Figure 1A). On chromosome 19, rs429358 was the
85 lead variant at the *APOE* locus ($\beta = 0.013$, $p = 6.4 \times 10^{-47}$, MAF = 15.6%). We tested whether
86 the presence of *APOE*- $\epsilon 4$ was enriched in certain primary causes of death. Among the top four
87 causes of death, each representing over 5% of total deaths (Figure 1B), only those due to
88 'Diseases of the circulatory system' (Chi-square $p = 1.6 \times 10^{-16}$) and 'Diseases of the nervous
89 system' ($p = 1.1 \times 10^{-71}$) showed a significant enrichment in the proportion of $\epsilon 4$ carriers
90 compared to the prevalence of $\epsilon 4$ carriers among all subjects (Figure 1C). In the chromosome
91 6 locus, overlapping *ZSCAN23*, two variants were GWS: rs6902687, located 2.2 kb upstream
92 of the transcription start site (TSS), and rs13190937 situated in the 5' untranslated region
93 (UTR) (rs6902687_C: $\beta = 0.004$, $p = 1.5 \times 10^{-8}$, MAF = 36.6%; rs13190937_A: $\beta = 0.004$, $p =$
94 1.5×10^{-8} , MAF = 36.6%, Figure 1D). To explore a potential regulatory function for variants
95 at the *ZSCAN23* locus, we investigated whether the lead SNPs were expression quantitative
96 trait loci (eQTLs) in the Genotype-Tissue Expression Project (GTEx) v8 database. rs13190937
97 was significantly associated with increased *ZSCAN23* expression in pancreatic tissue and the

98 longevity GWAS signal colocalized with the *ZSCAN23* expression quantitative trait loci
99 (eQTL) (posterior probability of colocalization (PP4) = 0.94; Figure 1E). Phenome-wide
100 association study analysis (PheWAS) using PheWeb [15] shows that the main associations of
101 rs13190937 are with celiac disease and intestinal malabsorption ($p = 1.8 \times 10^{-57}$)
102 (Supplementary Figure 1).

103 In sex-stratified GWAS (180,970 males and 212,863 females), the *APOE* locus was again
104 linked to longevity in both males and females (Supplementary Table 1 and Supplementary
105 Figure 2A and B). In males, we observed an additional GWS association for rs35705950_T
106 located between *MUC5AC* and *MUC5B* on chromosome 11 ($\beta = 0.01$, $p = 2.1 \times 10^{-8}$, MAF =
107 11.2%) (Supplementary Figure 2C), while no additional association was found in females. This
108 variant was notably linked to increased *MUC5B* expression in lung tissue with the longevity
109 GWAS signal aligning with a *MUC5B* eQTL (PP4 = 0.99; Supplementary Figure 2D). We also
110 confirmed through PheWAS that rs35705950 is associated with a diagnosis of pulmonary
111 fibrosis ($p = 4.4 \times 10^{-13}$) and “Other interstitial pulmonary diseases with fibrosis” listed as
112 primary cause of death ($p = 1.7 \times 10^{-5}$), and Illness of the father “Lung cancer” ($p = 2.1 \times 10^{-4}$),
113 but not with the mother’s ($p = 0.07$) (Supplementary Figure 2E).

114

115 **Gene-based rare variant association analyses in whole-exome data**

116 Among 26,230,624 variants with MAF < 1%, 1,830,070 variants (17,174 genes) were
117 annotated as loss-of-function (LoF) or missense variants. Of these, 628,362 were predicted LoF
118 variants (17,071 genes with a median of 28 variants per gene), 985,950 were missense variants
119 predicted as damaging by AlphaMissense (15,891 genes with a median of 47 variants per gene),
120 and 349,791 were missense variants predicted as damaging by rare exome variant ensemble
121 learner (REVEL) (12,219 genes with a median of 12 variants per gene). Of variants classified
122 by each, 23.7% of AlphaMissense and 66.9% of REVEL variants were also pathogenic by the
123 other classifier.

124 We identified six genes whose burden of LoF variants is significantly associated with reduced
125 lifespan: *TET2* ($p = 3.8 \times 10^{-30}$), *ATM* ($p = 6.0 \times 10^{-10}$), *BRCA2* ($p = 1.3 \times 10^{-34}$), *CKMT1B*
126 ($p = 4.5 \times 10^{-7}$), *BRCA1* ($p = 4.9 \times 10^{-12}$) and *ASXL1* ($p = 2.3 \times 10^{-44}$) (Figure 2A and Table
127 1). All of these but *CKMT1B* also showed gene-wide significance in a direction-agnostic
128 (SKAT-O) approach (Supplementary Figure 3A). Additionally, in eight genes, the burden of
129 missense variants predicted as pathogenic by AlphaMissense was associated with reduced
130 lifespan: *DNMT3A* ($p = 1.4 \times 10^{-9}$), *SF3B1* ($p = 6.7 \times 10^{-12}$), *CHL1* ($p = 5.0 \times 10^{-7}$), *TET2*
131 ($p = 4.2 \times 10^{-7}$), *PTEN* ($p = 1.0 \times 10^{-8}$), *SOX21* ($p = 2.9 \times 10^{-8}$), *TP53* ($p = 3.1 \times 10^{-15}$) and
132 *SRSF2* ($p = 9.8 \times 10^{-89}$) (Figure 2B). Lastly, three genes showed gene-wide significance for
133 burden of missense variants predicted by REVEL: *DNMT3A* ($p = 6.6 \times 10^{-9}$), *PTEN*
134 ($p = 6.6 \times 10^{-8}$), and *TP53* ($p = 6.2 \times 10^{-9}$) (Supplementary Figure 4 and Supplementary Table
135 2). SKAT-O identified additional associations with pathogenic missense variants predicted by
136 AlphaMissense in *C1orf52* ($p = 2.1 \times 10^{-7}$), *IDH2* ($p = 5.3 \times 10^{-39}$) and *RLIM* ($p = 3.7 \times 10^{-9}$)
137 (Supplementary Figure 3B), and by REVEL in *TERT* ($p = 8.1 \times 10^{-10}$) (Supplementary Figure
138 3C and Supplementary Table 2).

139 For sex-specific gene-based analysis, an additional six genes not identified in the whole-cohort
140 analysis showed gene-wide significance in males by either burden or SKAT-O: *CDKN1A* and
141 *PTPRK* (LoF); *COA7* and *TG* (AlphaMissense); *NMNAT2* and *PITRM1* (REVEL)
142 (Supplementary Figure 5A, 6A and Supplementary Table 3). In females, we identified three
143 additional genes associated with reduced lifespan: *PORCN* (AlphaMissense); *UGT1A8* and
144 *OLIG1* (REVEL) (Supplementary Figure 5B, 6B and Supplementary Table 4).

145

146 **Gene-burden survival analysis**

147 For the 13 gene-wide significant genes in the burden analyses, we assessed the association of
148 variant carrier status with lifespan using Cox proportional hazards regression. Carriers of LoF
149 variants in six genes were associated with decreased survival compared to non-carriers:
150 *CKMT1B* (HR=3.9, $p=2.1 \times 10^{-6}$), *ASXL1* (HR=2.2, $p=3.8 \times 10^{-26}$) (Figure 3A), *TET2*
151 (HR=1.7, $p=2.7 \times 10^{-18}$), *ATM* (HR=1.7, $p=2.5 \times 10^{-10}$), *BRC A2* (HR=2.4, $p=1.1 \times 10^{-40}$),
152 and *BRC A1* (HR =2.2, $p=1.0 \times 10^{-12}$) (Supplementary Figure 7A). Similarly, carriers of
153 AlphaMissense-predicted pathogenic variants exhibited significantly earlier mortality
154 compared to non-carriers on the following genes: *DNMT3A* (HR=1.4, $p=2.0 \times 10^{-7}$), *SF3B1*
155 (HR=2.1, $p=2.0 \times 10^{-8}$), *CHL1* (HR=1.3, $p=3.4 \times 10^{-7}$), *PTEN* (HR=4.2, $p=3.4 \times 10^{-10}$),
156 *SOX21* (HR=1.9, $p=5.2 \times 10^{-8}$), *TP53* (HR=3.7, $p=1.7 \times 10^{-13}$), *SRSF2* (HR=5.0, $p=2.4 \times$
157 10^{-52}) (Figure 3B) and *TET2* (HR=1.5, $p=3.8 \times 10^{-6}$) (Supplementary Figure 7B). Carriers
158 of pathogenic variants predicted by REVEL showed similar trends: *DNMT3A* (HR=1.5,
159 $p=1.9 \times 10^{-6}$), *PTEN* (HR=5.3, $p=1.3 \times 10^{-10}$), and *TP53* (HR=2.4, $p=6.6 \times 10^{-8}$)
160 (Supplementary Figure 7C).

161 To explore the contribution of individual rare variants to mortality in each gene-wide
162 significant gene in the burden and SKAT-O tests, we conducted Cox proportional hazards
163 regression for each variant with a minor allele count (MAC) of three or more (Table 2). In total,
164 587 variants including LoF, AlphaMissense and REVEL variants were examined. After
165 applying a Bonferroni correction for multiple testing, setting the significance threshold at
166 8.5×10^{-5} (0.05/587), we identified significant associations with reduced lifespan for four LoF
167 variants: rs370735654 in *TET2* (MAC=17, HR=7.0, $p=5.6 \times 10^{-9}$), rs587779834 in *ATM*
168 (MAC=113, HR=2.5, $p=2.7 \times 10^{-5}$), rs80359520 in *BRC A2* (MAC=10, HR=6.1, $p=1.8 \times$
169 10^{-6}), rs750318549 in *ASXL1* (MAC=201, HR=2.5, $p=2.3 \times 10^{-15}$). Additionally, significant
170 associations with AlphaMissense variants were noted in six genes, impacting lifespan:
171 rs769009649 in *C1orf52* (MAC=62, HR=3.2, $p=7.4 \times 10^{-7}$), rs147001633 in *DNMT3A*
172 (MAC=269, HR=1.8, $p=3.7 \times 10^{-5}$), rs377023736 in *SF3B1* (MAC=12, HR=6.0, $p=8.5$
173 $\times 10^{-8}$), rs116421102 in *CHL1* (MAC=1,842, HR=1.3, $p=2.5 \times 10^{-5}$), rs121913502 in *IDH2*
174 (MAC=45, HR=5.7, $p=1.0 \times 10^{-20}$), rs11540652 in *TP53* (MAC=5, HR=10.0, $p=6.6 \times 10^{-8}$)
175 and rs751713049 in *SRSF2* (MAC=51, HR=5.8, $p=1.9 \times 10^{-26}$). For missense variants
176 predicted by REVEL, rs1043358053 in *TERT* (MAC=5, HR=11.9, $p=7.4 \times 10^{-5}$) and
177 rs11540652 in *TP53* were significantly linked to reduced lifespan (Supplementary Table 5).

178

179 Phenome-wide association studies

180 For the nine novel longevity genes identified in the burden test (*CKMT1B*, *ASXL1*, *DNMT3A*,
181 *SF3B1*, *CHL1*, *PTEN*, *SOX21*, *TP53* and *SRSF2*), we examined the burden of LoF or
182 pathogenic missense variants through PheWASs across 1,670 UKB phenotypes including
183 disease occurrences derived from electronic health record, self-reported family history, and
184 physical measures (Supplementary Figure 8). The burden of LoF variants in *ASXL1* and
185 AlphaMissense variants in *DNMT3A*, *SF3B1*, *PTEN*, *TP53* and *SRSF2* were strongly linked to
186 an increased risk of leukemia: acute myeloid leukemia (*ASXL1*: Odds Ratio (OR)=1.05;
187 $p=8.6 \times 10^{-170}$; *DNMT3A*: OR=1.03, $p=2.1 \times 10^{-150}$; *SRSF2*: OR=1.3, $p=1.2 \times 10^{-195}$;
188 *TP53*: OR=1.05, $p=4.7 \times 10^{-35}$), monocytic leukemia (*DNMT3A*: OR=1.01, $p=2.5 \times 10^{-9}$),
189 chronic lymphoid leukemia (*SF3B1*: OR=1.07, $p=4.1 \times 10^{-68}$) and acute lymphoid leukemia
190 (*PTEN*: OR=1.01, $p=2.1 \times 10^{-14}$). Additionally, the burden of LoF in *CKMT1B* was
191 associated with hypopharynx cancer (OR=1.03, $p=3.9 \times 10^{-26}$), vertiginous syndromes
192 (OR=1.03, $p=3.0 \times 10^{-17}$) and salivary glands cancer (OR=1.03; $p=3.2 \times 10^{-12}$). *SOX21*
193 burden was associated with increased acne (OR=1.01, $p=6.9 \times 10^{-7}$) and spinocerebellar
194 disease (OR=1.01, $p=2.3 \times 10^{-6}$).

195

196 **Somatic mutation and clonal hematopoiesis of indeterminate potential**

197 We computed the variant allelic fraction (VAF) per carrier for each variant included in the
198 analysis. Generally, germline variants have a mean VAF close to 50%, while somatic variants'
199 mean VAF will be lower [16]. Thus, when an association is linked to clonal hematopoiesis of
200 indeterminate potential (CHIP), we expect the distribution of VAF to be left-shifted compared
201 to a normal distribution centered at VAF = 50%. Considering LoF variants, *TET2* (mean VAF
202 across variants [95% bootstrap confidence interval for the mean VAF] = 0.33 [0.31,0.34]) and
203 *ASXL1* (mean VAF =0.32 [0.31,0.33]) burden test associations are supported by variants with
204 a left-shifted VAF distribution (Supplementary Table 6, Supplementary Figure 9A). Similarly,
205 considering pathogenic Alpha Missense variants, in *DNMT3A* (mean VAF= 0.24 [0.23-0.24]),
206 *TET2* (mean VAF=0.36 [0.34,0.38]), *TP53* (mean VAF=0.28 [0.24,0.34]), *SRSF2* (mean
207 VAF=0.30 [0.28,0.31]), *SF3B1* (mean VAF=0.31 [0.26,0.37]) and *CHL1* (mean
208 VAF=0.37[0.28-0.45]) are also left-shifted and the observed associations may be linked to
209 CHIP (Supplementary Table 5, Supplementary Figure 9B).

210

211 **Methods**

212

213 **Study participants**

214 The UKB is a large population-based longitudinal cohort study with recruitment from 2006 to
215 2010 in the United Kingdom [17]. In total, 502,664 participants aged 40-69 years were
216 recruited and underwent extensive phenotyping including health and demographic
217 questionnaires, clinic measurements, and blood draw at one of 22 assessment centers, of whom
218 468,541 subjects have been genotyped by both SNP array and WES.

219 We restricted our analysis to 393,833 individuals who self-reported their ethnic background as
220 'white British' and were categorized as European ancestry based on genetic ethnic grouping
221 (Field: 21000). Among them, 35,551 subjects were reported deceased, and their ages at death
222 were recorded from the UK Death Registry (Field: 40007). For the other 358,282 subjects
223 without death records, we assumed they were still alive by the latest censoring date (November
224 30, 2023). We determined their last known ages by subtracting their year and month of birth
225 (Field: 33) from the censoring date.

226

227 **SNP array genotyping and QC**

228 A total of 488,000 UKB participants were genotyped using one of two closely related
229 Affymetrix microarrays (UKB Axiom Array or UK BiLEVE Axiom Array) for ~820,000
230 variants. Quality control (QC), phasing, and imputation were performed as described
231 previously [18]. Briefly, the genotyped dataset was phased and imputed into UK10K, 1000
232 Genomes Project phase 3, and Haplotype Reference Consortium reference panels, resulting in
233 approximately 97 million variants. Additionally, we removed SNPs with imputation quality
234 score < 0.3, genotype missing rate > 0.05, minor allele frequency (MAF) < 1%, and Hardy-
235 Weinberg equilibrium $p < 1.0 \times 10^{-6}$.

236

237 **Genome-wide association studies**

238 We performed linear regression models using PLINK v2.0 [19] to test the association of
239 common variants (MAF \geq 1%) with longevity for the entire cohort, as well as stratified by sex.
240 For all three analyses, we used Martingale residuals calculated using the Cox proportional
241 hazards model as the outcome variable. The procedure for calculating Martingale residuals was
242 as follows. First, a Cox proportional hazards model [20] was fitted without genotype,

243

$$H(t|z) = H_0(t)e^{z\beta'}$$

244 where $H_0(t)$ is the baseline hazard function at time point t given the last-known age and
245 dead/alive status, $Z = [Z_1, \dots, Z_k]$ is a covariate matrix, and $\beta = [\beta_1, \dots, \beta_k]$ is a coefficient
246 matrix for Z . Here, we included sex and the first five principal components (PC) as covariates,
247 but for sex-specific analyses, sex was excluded. Then, Martingale residuals were calculated as:

$$248 \quad \widehat{M}_i = \delta_i - \widehat{H}_0(t)e^{Z\widehat{\beta}'}$$

249 where δ_i is the dead/alive status (0=alive, 1=dead) of the i th subject and $\widehat{\beta}$ is the estimated
250 coefficient matrix. We adapted the *coxph* function from the *survival* (v.3.6.4) R package [21]
251 to compute the Martingale residuals. Genome-wide significance threshold was set at the
252 standard GWAS level ($p=5.0 \times 10^{-8}$). We used *LocusZoom* [22] to generate regional plots
253 and Python v.3.7 to create Manhattan plots.

254

255 **Gene expression and colocalization analysis**

256 To evaluate the effect of the significant loci identified in our GWAS, we examined expression
257 quantitative trait loci (eQTLs) across 49 tissues having at least 73 samples from the Genotype-
258 Tissue Expression Project (GTEx) version 8 [23]. Bayesian colocalization analysis was
259 employed using the *COLOC* package (v.5.2.3) [24] in R and the posterior probability of
260 colocalization (PP4) was calculated between GWAS findings and eQTL associations within a
261 1 megabase (Mb) window. Additionally, colocalization was visualized using the
262 *locuscompareR* package [25].

263

264 **Whole-exome sequencing and QC**

265 Whole exome sequencing (WES) data was available for 469,835 UKB participants. The dataset
266 was generated by the Regeneron Genetics Center [26]. Details about the production and QC
267 for the WES data are described previously [26]. We restricted the WES analysis to rare variants
268 (MAF < 1%).

269

270 **Rare variant annotation**

271 Rare variants in WES data were annotated using Variant Effect Predictor (v. 112) provided by
272 Ensembl [27]. We defined LoF variants as those with predicted consequences: splice acceptor,
273 splice donor, stop gained, frameshift, start loss, stop loss, transcript ablation, feature elongation,
274 or feature truncation. Missense variants were annotated using AlphaMissense [28] and REVEL
275 [29] plugins and included if they had an AlphaMissense score ≥ 0.7 or REVEL score ≥ 0.75 .
276 All annotation was conducted based on GRCh38 genome coordinates.

277

278 **Gene-based rare variant association studies**

279 For testing groups of rare variants, genotype matrices were first transformed into a binary
280 variable describing whether samples carry a variant of a given class as follows:

$$281 \quad G_i = \begin{cases} 1, & \text{if } \sum_{j=1}^k g_{ij} > 0 \\ 0, & \text{if } \sum_{j=1}^k g_{ij} = 0 \end{cases},$$

282 Where g_{ij} is the minor allele count observed for subject i at variant j in the gene and k is the
283 number of variants in the gene. We carried out two gene-based tests: Burden test and sequence
284 kernel association test-optimized (SKAT-O) [30]. The burden test is a mean-based test that
285 assumes the same direction of effects for all variants within a gene. On the other hand, SKAT-
286 O employs a weighted average of the burden test and SKAT [31], the latter a variance-based
287 test that does not lose power when variants have opposing directions of effect.

288 Association tests were performed for each gene and rare variant class, including separately LoF
289 variants, missense variants with an AlphaMissense score ≥ 0.7 , and missense variants with a
290 REVEL score ≥ 0.75 , using Martingale residuals as the phenotype as in the common variant
291 analyses. We excluded genes with fewer than 10 variant carriers to ensure the reliability of our
292 analyses. A gene-wide significance threshold was established at $p=7.4 \times 10^{-7}$ based on the
293 Bonferroni method accounting for the number of genes, variant classes, and statistical methods.
294 Gene-based analyses were carried out using the *SKAT* package (v.2.2.5) in R.
295 To characterize the impacts of gene burden in significant genes, we compared lifespan survival
296 depending on gene burden using Kaplan–Meier survival curves, log-rank tests, and Cox
297 proportional hazard regression analyses. Additionally, we performed Cox proportional hazards
298 regression to assess the effect of each rare variant in a gene. The *survival* (v.3.6.4) package in
299 R was utilized for the survival analysis.

300

301 **Phenome-wide association studies**

302 For gene-wide significant genes, we conducted phenome-wide association studies (PheWAS)
303 of variant carrier status across 1,670 phenotypes in the UKB derived from binary, categorical,
304 and continuous traits. Phenotypes included the International Classification of Disease 10
305 (ICD-10) codes, family history (e.g. father's illness, father's age at death), blood count (e.g.
306 white blood cell count), blood biochemistry (e.g. Glucose levels), infectious diseases (e.g. pp
307 52 antigen for Human Cytomegalovirus), physical measures (e.g. BMI), cognitive test (e.g.
308 pairs matching) and brain measurements (e.g. subcortical volume of hippocampus). For ICD-
309 10 codes, we excluded phenotypes from the following ICD-10 chapters: 'Injuries, poisonings,
310 and certain other consequences of external causes' (Chapter XIX), 'External causes of
311 morbidity and mortality' (Chapter XX), 'Factors influencing health status and contacts with
312 health services' (Chapter XXI), and 'Codes for special purposes' (Chapter XXII). The ICD-10
313 codes were then converted into Phecodes (v.1.2) [32] which combine correlated ICD codes into
314 a distinct code and improve alignment with diseases commonly used in clinical practice.
315 For binary traits, we removed phenotypes with fewer than 100 cases, and for continuous traits,
316 those with fewer than 100 participants were excluded. Depending on the phenotype, we
317 employed various regression models including binary logistic regression, ordinal logistic
318 regression, multinomial logistic regression, and linear regression. All analyses included age,
319 sex, and first five PCs as covariates. Phenome-wide significance threshold was set at
320 $p=2.9 \times 10^{-5}$ based on the number of phenotypes.

321

322 **Variant allelic fraction**

323 To investigate whether some gene-level associations are enriched for somatic variants, we
324 computed the variant allele frequency (VAF) for each heterozygous sample, reporting the mean
325 VAF and VAF distribution per gene per variant class. VAF is defined as the number of reads
326 with an alternate allele divided by the read depth at a given variant position. We also calculated
327 the confidence interval for the mean VAF per gene using 10,000 bootstrap samples to ensure
328 robust statistical analysis.

329

330 **Discussion**

331 In this study, we report several known and novel findings related to genetic risks associated
332 with longevity analyzing 393,833 European participants from the UKB. In the common variant
333 GWAS, three independent loci associated with increased mortality risk were identified. In the
334 gene-based analysis of rare non-synonymous variants, 17 genes had their burden/SKAT-O test
335 associated with longevity.

336 Consistent with previous reports, rs429358, determining the *APOE-ε4* allele dosage, was
337 associated with decreased lifespan across both sexes. *APOE-ε4* is well known for its
338 associations with Alzheimer's Disease [33] and cardiovascular disease [34]. In our dataset, the
339 proportion of $\epsilon 4$ carriers was significantly higher for deaths caused by 'Disease of the
340 circulatory system' and 'Diseases of the nervous system' compared to the general prevalence
341 of $\epsilon 4$ carriers, which could explain the effect of $\epsilon 4$ on longevity. Examining the subcategories
342 of these ICD-10 chapters, 'Disease of the circulatory system' includes cardiovascular disease
343 (I51.6), while 'Diseases of the nervous system' covers Alzheimer's disease (G30). We also
344 identified a GWS association at the *ZSCAN23* locus, which had not been previously reported.
345 Our colocalization analysis revealed that the longevity-associated signal colocalizes with a
346 *ZSCAN23* eQTL in pancreatic tissue with increased expression observed in minor allele
347 carriers. Although the role of *ZSCAN23* remains unclear, recent studies have linked its
348 expression to pancreatic tumors, supporting our colocalization findings [35]. For sex-specific
349 GWAS, a GWS association specific to males was found between *MUC5AC* and *MUC5B*,
350 which highly colocalizes with a *MUC5B* eQTL in lung tissue and many studies have linked
351 this variant to pulmonary disease like idiopathic pulmonary fibrosis [36, 37] and COVID-19
352 [38, 39]. Previously reported SNP associations with longevity were concordant in our dataset
353 but none of these passed the GWAS suggestive threshold ($p=1.0 \times 10^{-5}$) except for those at
354 the *APOE* locus. This phenomenon likely resulted from previous studies relying on proxy data
355 such as parental age at death, which may capture a different set of genetic factors than direct
356 proband mortality data.

357 In our gene-based rare variant analysis, 17 genes achieved gene-wide significance (p
358 $< 7.4 \times 10^{-7}$) in either the burden or SKAT-O test. Four of these, *TET2*, *ATM*, *BRCA2*, and
359 *BRCA1*, were reported in a previous rare-variant analysis of longevity in UKB [9]. We
360 identified 13 novel genes associated with longevity—*CKMT1B*, *ASXL1*, *DNMT3A*, *SF3B1*,
361 *PTEN*, *SOX21*, *TP53*, *SRSF2*, *C1orf52*, *CHL1*, *IDH2*, *RLIM*, and *TERT*— when assessing
362 variants causing genetic LoF or missense variants classified as pathogenic by REVEL or
363 AlphaMissense. Of note, LoF and missense variant analyses identified mostly separate genes
364 with only one overlap (*TET2*). This supports the use of both categories in rare variant analyses
365 and may indicate that missense variants as classified by AlphaMissense capture a wider range
366 of variation missed when only assessing LoF variants, which are generally interpreted as
367 resulting in haploinsufficiency. Importantly, missense variants may lead to increased or
368 decreased protein function. In our analyses, *IDH2* was not gene-wide significant with the
369 burden test ($p=1.9 \times 10^{-4}$, Table 1) but was highly significant with SKAT-O ($p=5.3 \times 10^{-39}$,
370 Table 1). Since SKAT-O does not lose power when variants have differing directions of effect,
371 this suggests that different mutations in *IDH2* can lead to either increased or decreased
372 longevity. These results underline the gain in information achieved when studying rare
373 missense variants as well as LoF using appropriate statistical techniques.

374 Strikingly, most of the genes we identified carrying longevity-associated rare variants have
375 been previously linked to cancer. *TET2*, *ASXL1*, *DNMT3A*, and *SF3B1* are all known to harbor
376 causal leukemia variants [40-43], and somatic variants in *SRSF2* have been described in
377 myelodysplastic syndrome [44]. *ATM*, *BRCA2*, and *BRCA1* mutations have been well
378 characterized in breast, ovarian, and other cancers [45-47]. *RLIM* appears to be a regulator of
379 estrogen-dependent transcription, an important pathway in breast cancer [48], and has been
380 recently described as a potential tumor suppressor [49]. *PTEN* and *TP53* are well studied due
381 to their critical role in genomic stability and are the two most mutated genes in human cancer
382 [50]. *IDH2* is also frequently mutated in many kinds of cancer [51]. The antisense long
383 noncoding RNA *SOX21-AS1*, but not *SOX21*, has been linked to oral, cervical, and breast
384 cancer [52-54]. A recent study found potential for *CKMT1B* expression as a prognostic

385 biomarker in glioma [55]; similarly, alterations in *CHL1* expression have been associated with
386 development and metastasis of many types of cancer [56]. Finally, variation in both the coding
387 and promoter sequences of *TERT* has been associated with a variety of cancer types [57, 58].
388 Our PheWAS results also suggest that most of these genes are associated with cancer,
389 specifically blood-based tumors such as myeloid leukemia. Combined with the common
390 *ZSCAN23* locus we identified, associated with pancreatic tumors, this points to cancer being
391 the major genetic factor currently affecting lifespan in UKB. This is consistent with a previous
392 study of healthspan that found cancer to be the first emerging disease in over half of disease
393 cases in UKB [59]. These results likely reflect the characteristics of the cohort, comprised of
394 predominantly middle-aged individuals, with age-at-death ranging from 40.9 to 85.2 years and
395 last-known ages between 52.6 and 88.7 years.

396 For sex-specific rare variant analyses, we identified six novel genes (*CDKN1A*, *PTPRK*,
397 *COA7*, *TG*, *NMNAT2* and *PITRM1*) in males and three genes (*PORCN*, *UGT1A8* and *OLIG1*)
398 in females. Some of these genes have been found to associate with sex-specific diseases. In one
399 study, advanced prostate cancer patients had a higher frequency of a variant on the 3'UTR of
400 *CDKN1A* [60] and the gene has received attention as a potential therapeutic target for prostate
401 cancer [61]. *PORCN* is located on the X chromosome and mutations on it can cause Goltz-
402 Gorlin Syndrome [62], but it has also been found to regulate a signaling pathway that controls
403 cancer cell growth [63]. *UGT1A8* expression is altered in endometrial cancer [64] and amino
404 acid substitutions in it may modulate estradiol metabolism leading to increased risk of breast
405 and endometrial cancer [65].

406 Since UKB collected DNA from peripheral blood mononuclear cell samples, we explored
407 whether the variants were potentially of somatic origin, picked up by WES genotyping due to
408 CHIP. The VAF distribution of variants included in our analysis emphasizes that several
409 associations are likely linked to CHIP, and notably include the well-established CHIP-related
410 genes *TET2*, *ASLX1*, *DNMT3A*, *SF3B1*, *TP53* and *SRSF2*. While WES heterozygote genotypes
411 for these variants will not include all variants with some degree of CHIP within these genes (as
412 evidenced by many more individuals having non-zero alternate allele count at these locations,
413 data not shown), it does capture CHIP-related somatic variants sufficiently to establish robust
414 associations with longevity. In UKB the mean duration between the primary visit (blood draw
415 date) and death is currently 9.2 years (± 3.8) and suggests that WES screening for CHIP
416 variants may be used as a precision health tool to contribute to earlier cancer detection by
417 assessing individuals with higher susceptibility risks. In addition to known cancer variants,
418 such as breast cancer-related *BRCA1/BRCA2*, our study highlights novel associations that
419 should be considered in cancer susceptibility screenings.

420 By combining large-scale GWAS with rare variant analysis, this study enhances our
421 understanding of the genetic basis of human longevity. Our results emphasize the importance
422 of understanding the genetic factors driving the most prevalent causes of mortality on a
423 population level, highlighting the potential for early genetic testing to identify germline and
424 somatic variants that place some individuals at risk of early death. Understanding the biological
425 pathways through which these genes influence cancer and aging, as well as the environmental
426 factors interacting with these pathways, will be essential for developing therapeutic targets
427 aimed at extending healthy lifespan. Our study's implications thus extend beyond genetics, as
428 they touch on the broader aspects of health care, public health policy, and preventive strategies
429 against age-related diseases.

430 In conclusion, this study enhances our understanding of the genetic basis of human longevity
431 by combining large-scale GWAS with detailed rare variant analysis. The novel loci identified
432 warrant further exploration to understand their biological roles and interactions with

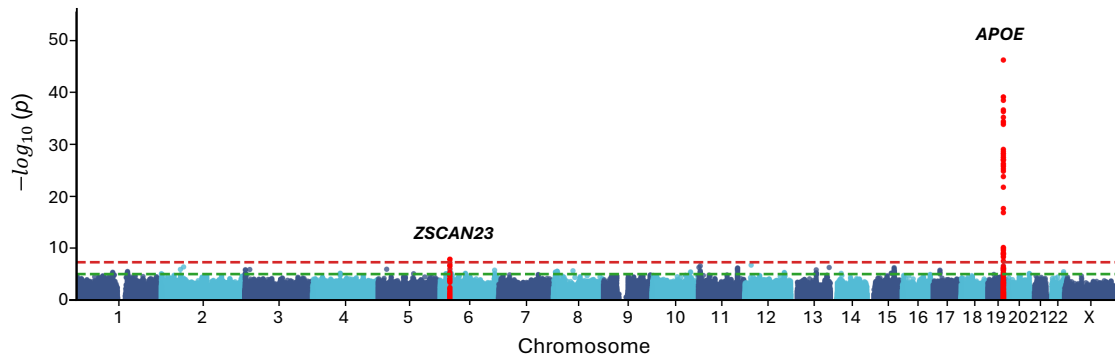
433 environmental factors, which will be crucial for unraveling the complex nature of aging and
434 developing strategies to mitigate its adverse effects.

435 **Figures and Tables**

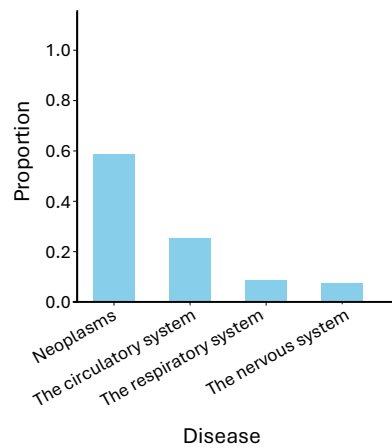
436

437 **Figure 1. Common variant GWAS of longevity.** (A) Manhattan plot. (B) The proportion of
 438 cause of death for the top 4 categories, each accounting for more than 5% of total deaths) (C)
 439 Association of causes of death with *APOE*- ϵ 4 genotype. (D) Locuszoom and (E) colocalization
 440 plots at the *ZSCAN23* locus, colocalized with *ZSCAN23* eQTL in pancreatic tissue in GTEx.
 441 PP4: posterior probability of colocalization.
 442

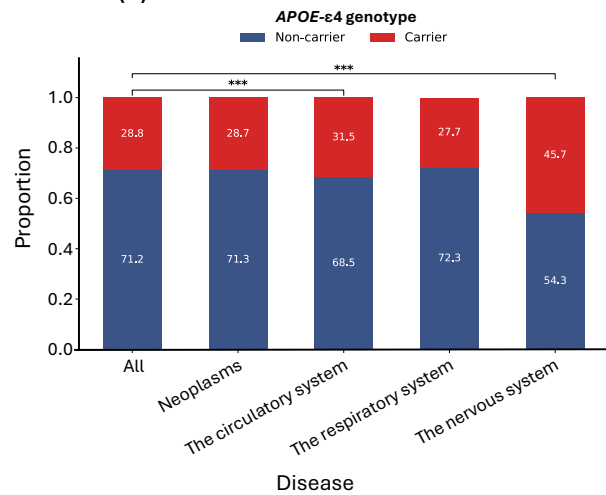
(A)



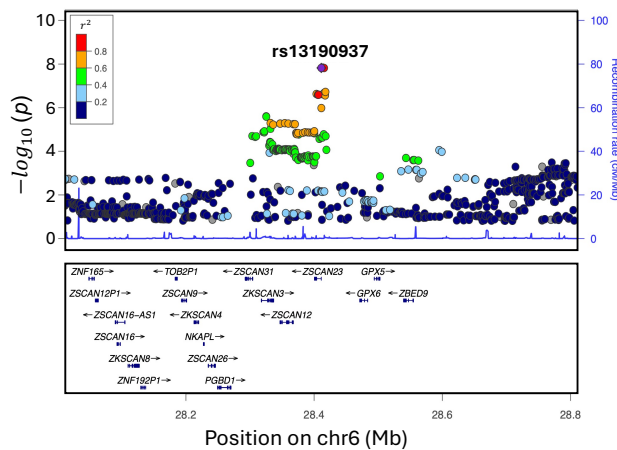
(B)



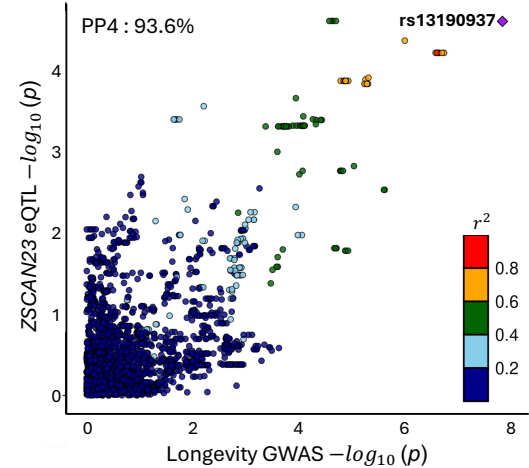
(C)



(D)

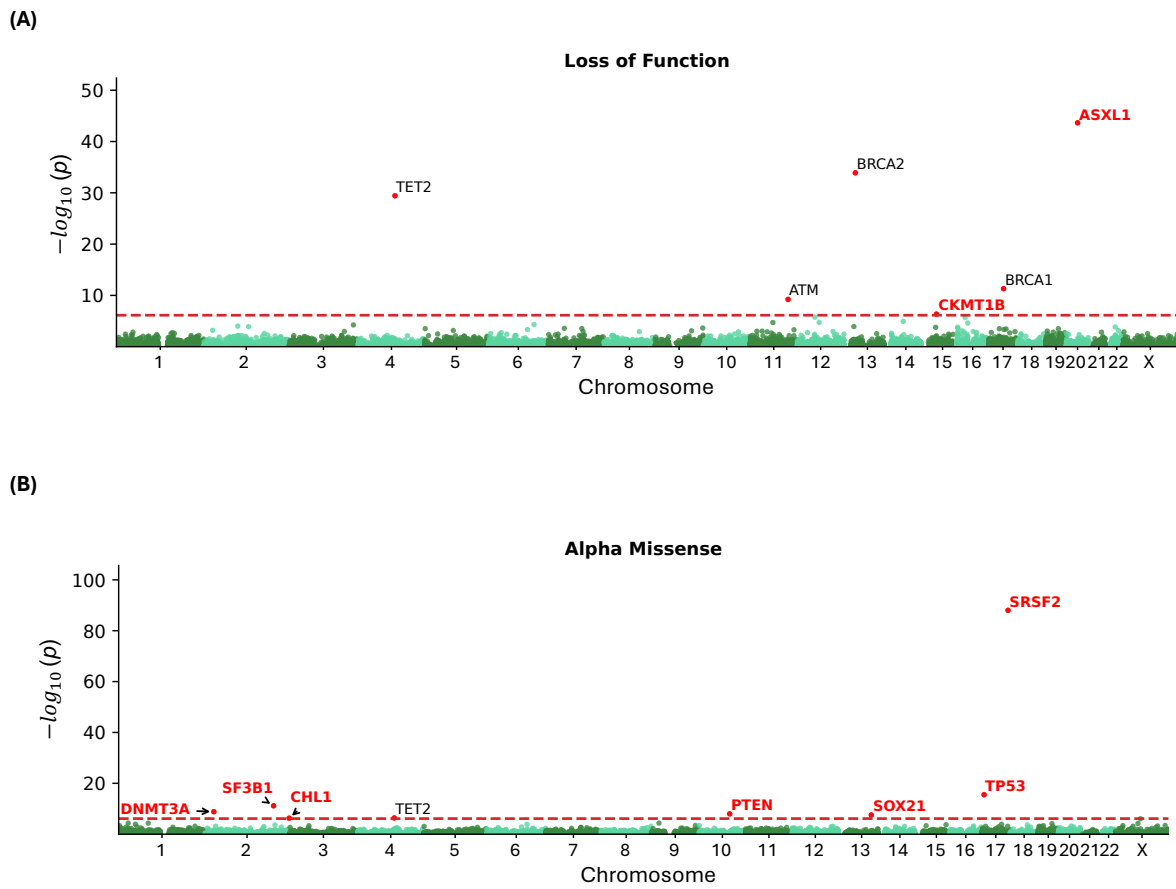


(E)



443

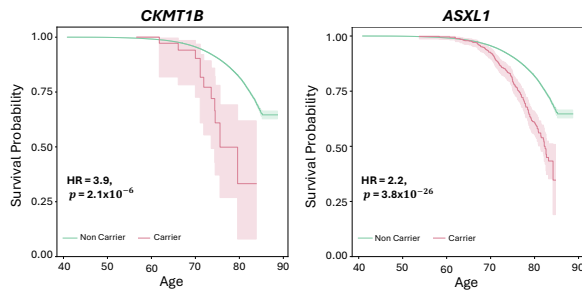
444 **Figure 2. Rare variant burden association with longevity, considering loss-of-functions**
445 **(A) and Alpha Missense pathogenic variants (B).** Novel genes are highlighted in red.
446



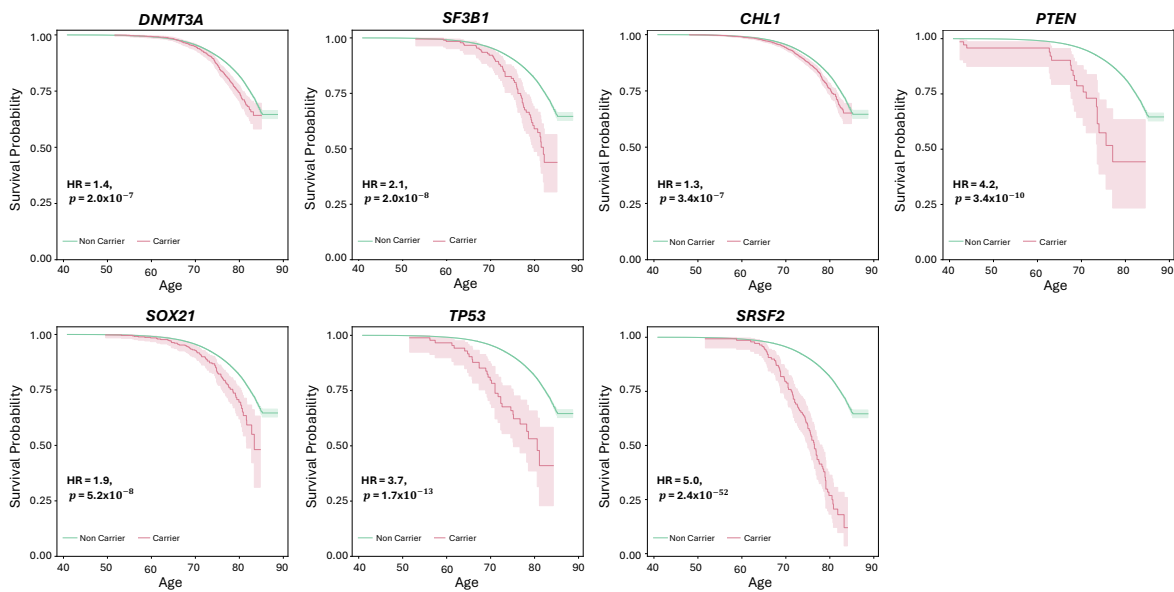
447

448 **Figure 3. Survival curves comparing carriers and non-carriers of variants on genes with**
449 **a significant burden of loss-of-function (A) and AlphaMissense pathogenic (B) variants.**
450

(A) Loss of Function

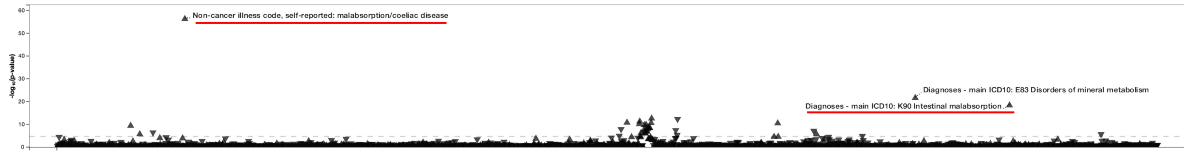


(B) Alpha Missense



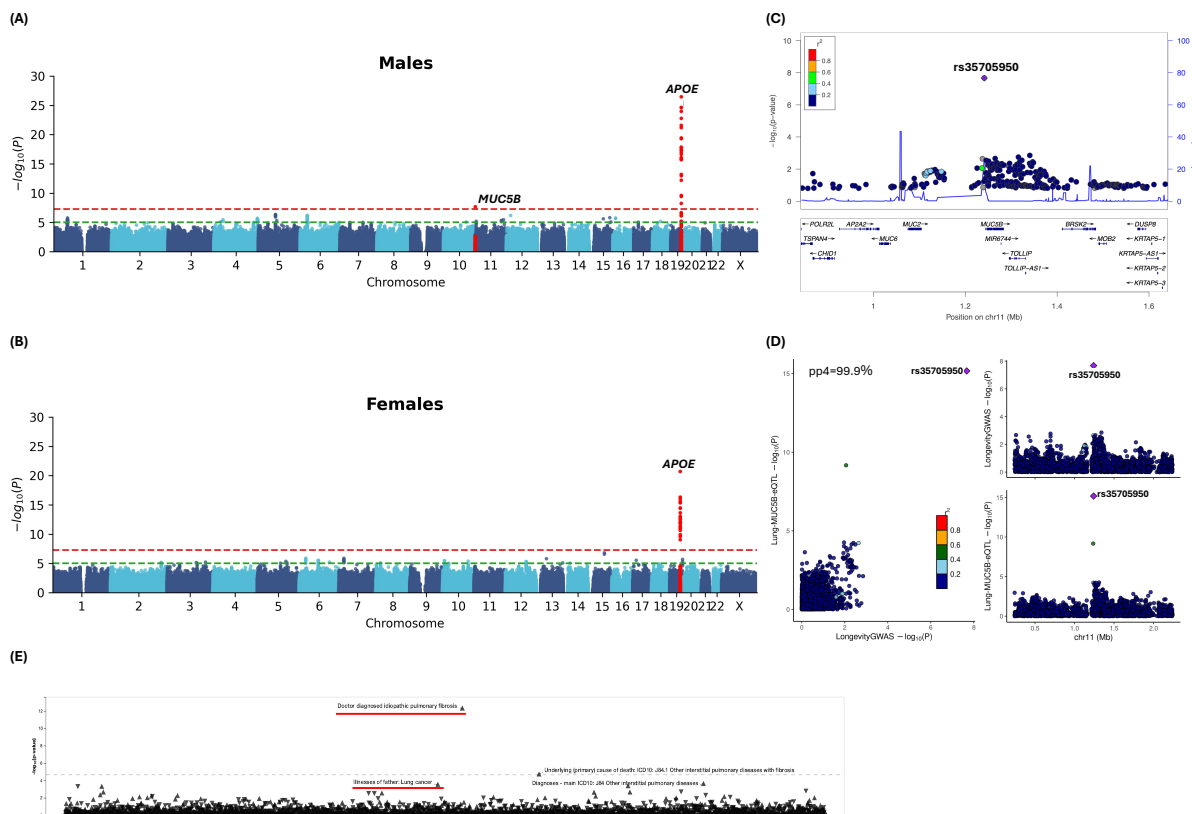
451
452

453 **Supplementary Figure 1. Phenome-wide association of rs13190937 on ZSCAN23.** This
454 analysis is based on PheWeb (<https://pheweb.org/UKB-Neale/>).
455



456

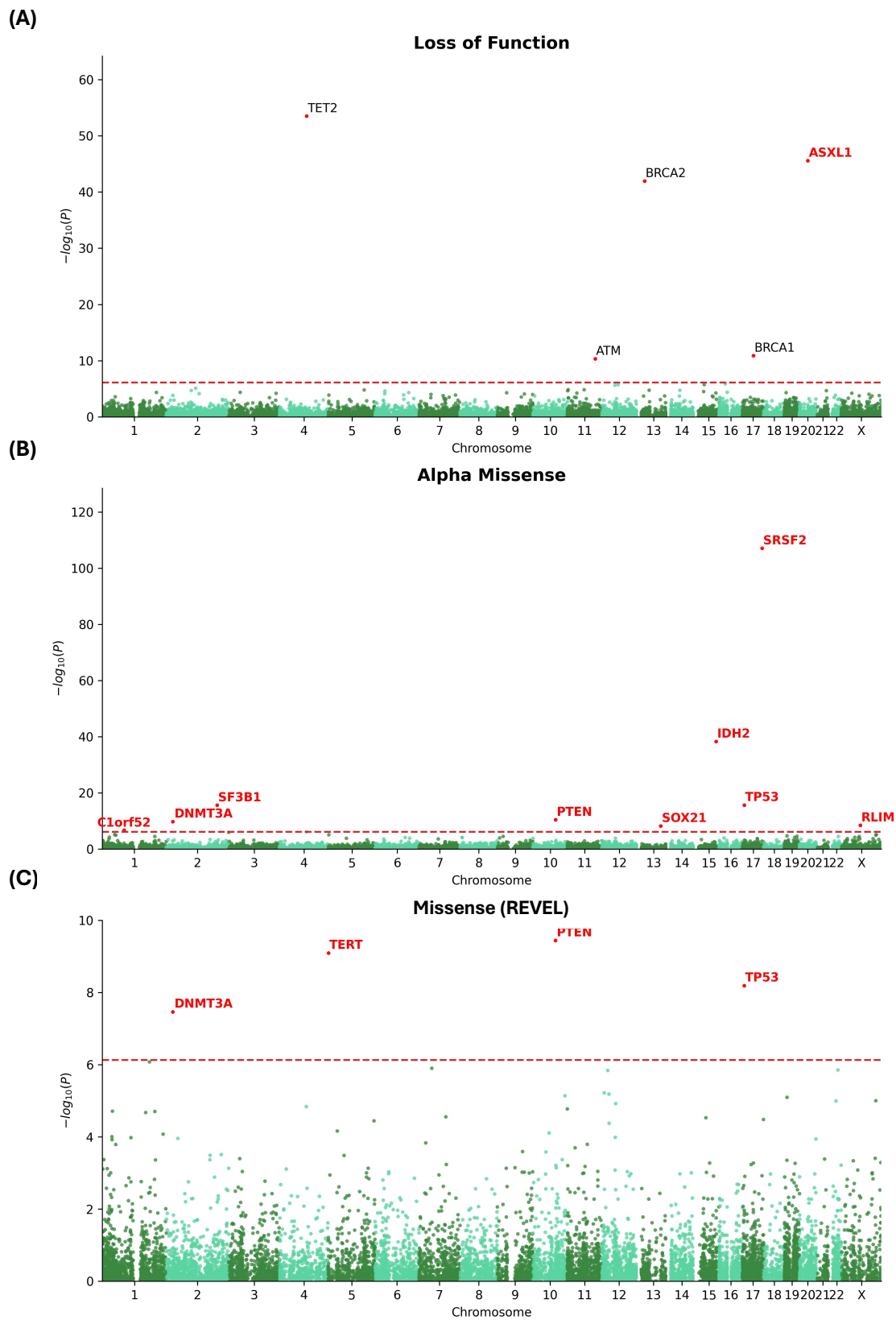
457 **Supplementary Figure 2. Sex-stratified common variant GWAS of longevity.** Manhattan
458 plot in males (A), and females (B). Locuszoom (C) and colocalization (D) plots at the *MUC5B*
459 locus in males, colocalized with *MUC5B* eQTL in lung tissue in GTEx. PP4: posterior
460 probability of colocalization. (E) Phenom-wide association of rs35705950. This analysis is
461 based on PheWeb (<https://pheweb.org/UKB-Neale/>).
462



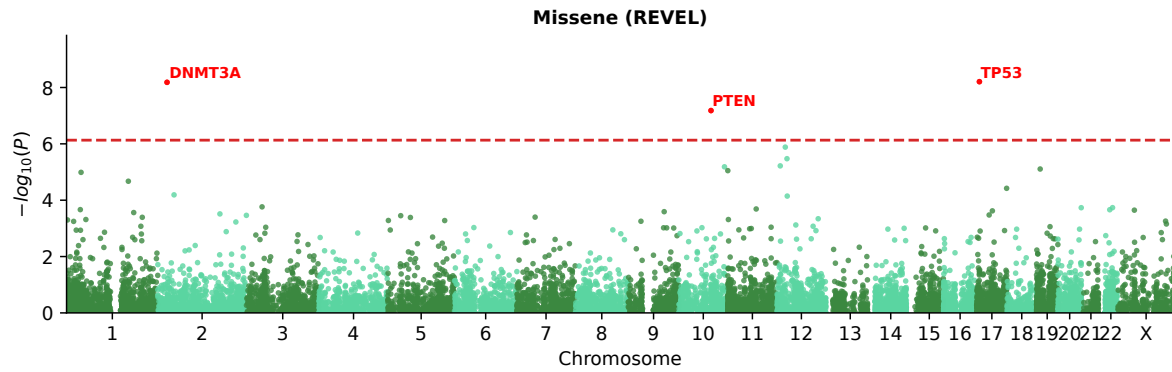
463
464

465

466 **Supplementary Figure 3. Rare variant SKAT-O association with longevity considering 3**
467 **categories: Loss-of-function (A), Alpha Missense (B), and REVEL (C). Novel genes are**
468 **highlighted in red.**
469

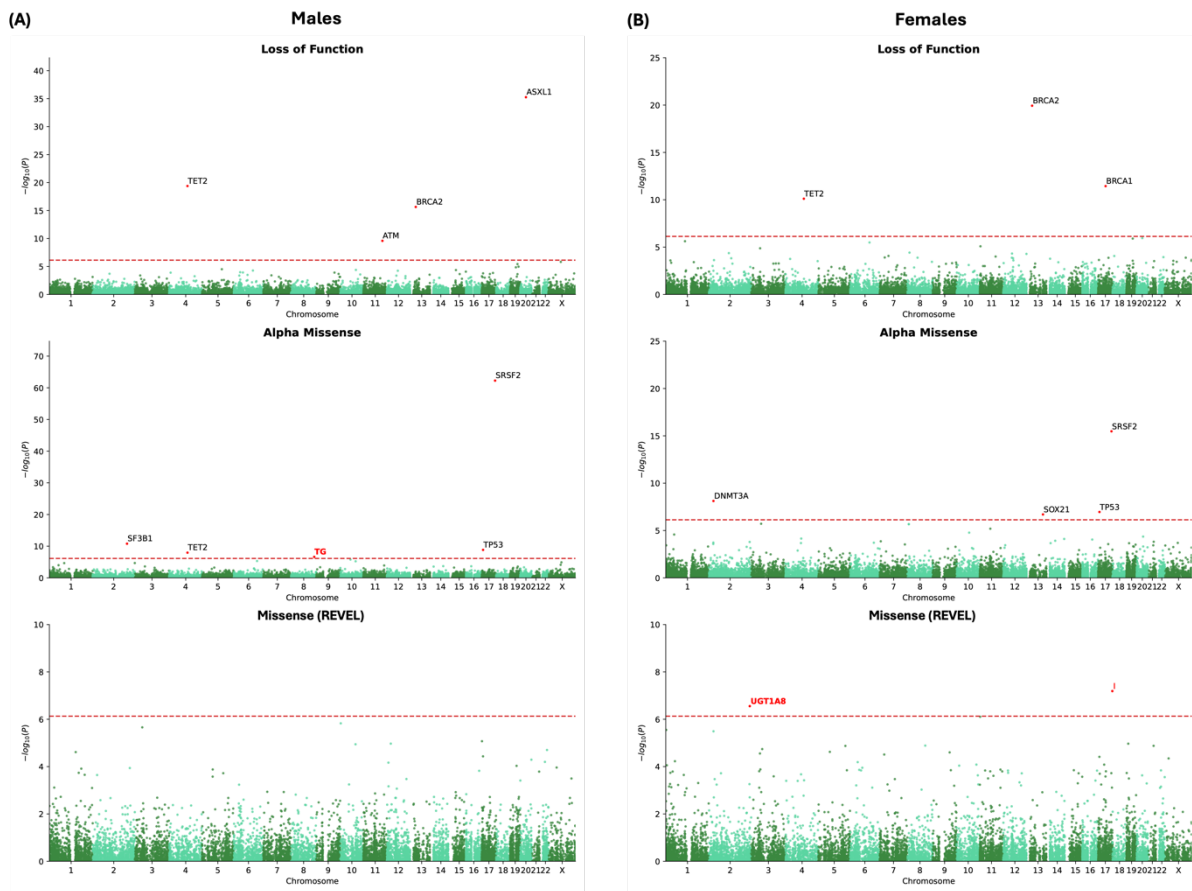


471 **Supplementary Figure 4. Rare variant burden association with longevity considering**
472 **REVEL pathogenic missense variants.** Novel genes are highlighted in red.
473



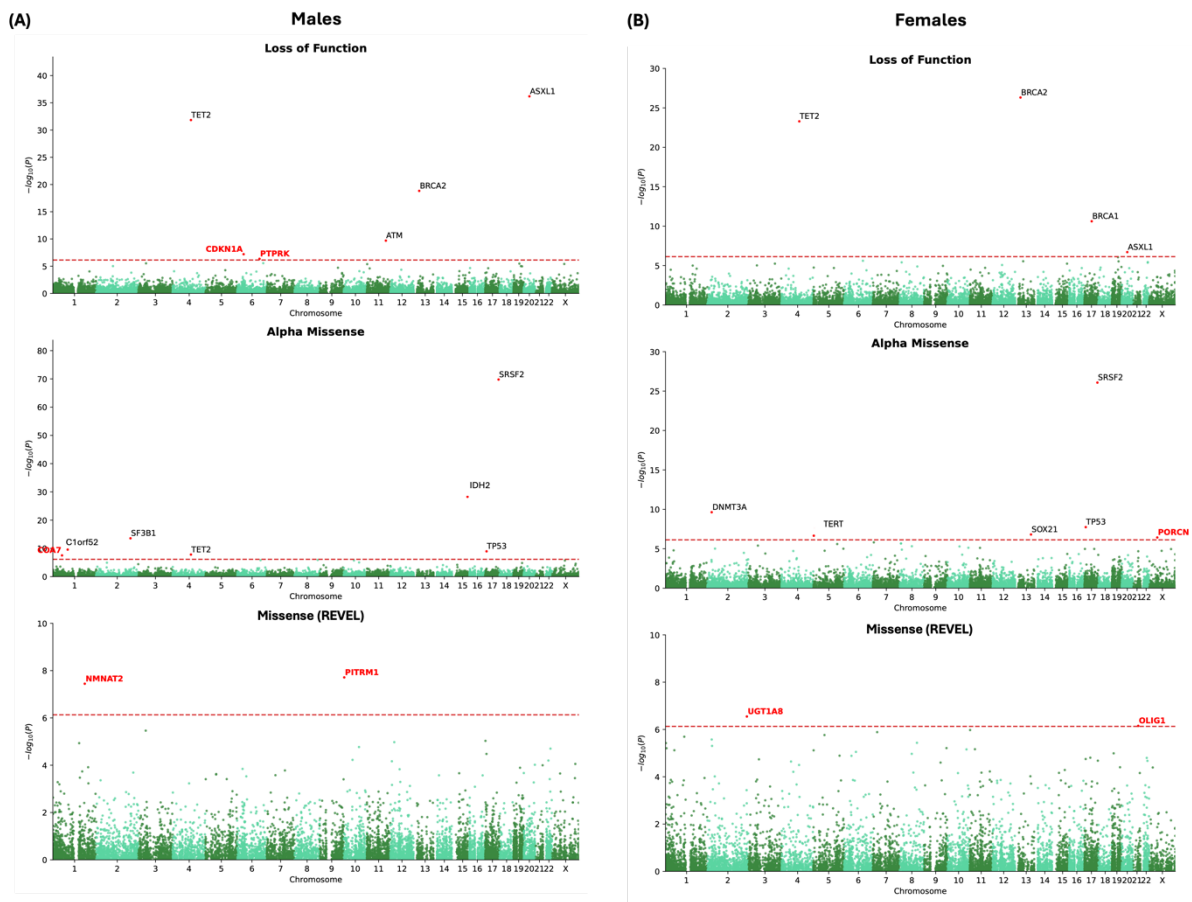
474
475
476

477 **Supplementary Figure 5. Sex-stratified rare variant burden association with**
478 **longevity considering 3 categories for each sex: Loss-of-function (A), Alpha Missense (B),**
479 **and REVEL (C). Novel genes are highlighted in red.**
480



481

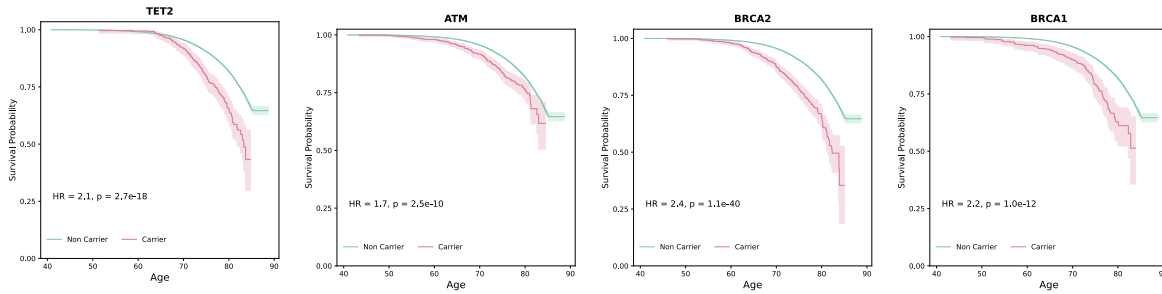
482 **Supplementary Figure 6. Sex-stratified rare variants SKAT-O association with**
483 **longevity considering 3 categories for each sex: Loss-of-function (A), Alpha Missense (B),**
484 **and REVEL (C). Novel genes are highlighted in red.**
485



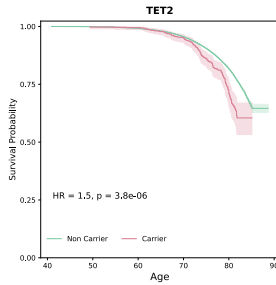
486

487 **Supplementary Figure 7. Survival curves comparing carriers and non-carriers of**
488 **variants considered on genes with a significant burden of loss-of-function (*TET2*, *ATM*,**
489 ***BRCA2* and *BRCA1*) (A), AlphaMissense pathogenic (B) variants (*TET2*), missense**
490 **variants predicted by REVEL (*DNMT3A*, *PTEN* and *TP53*) (C)**
491

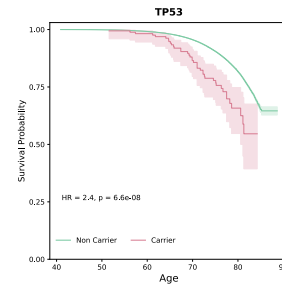
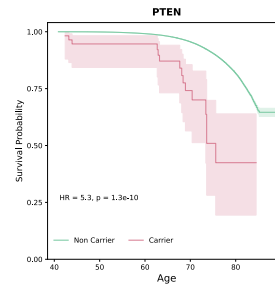
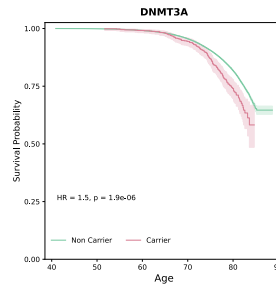
(A) Loss of Function



(B) Alpha Missense

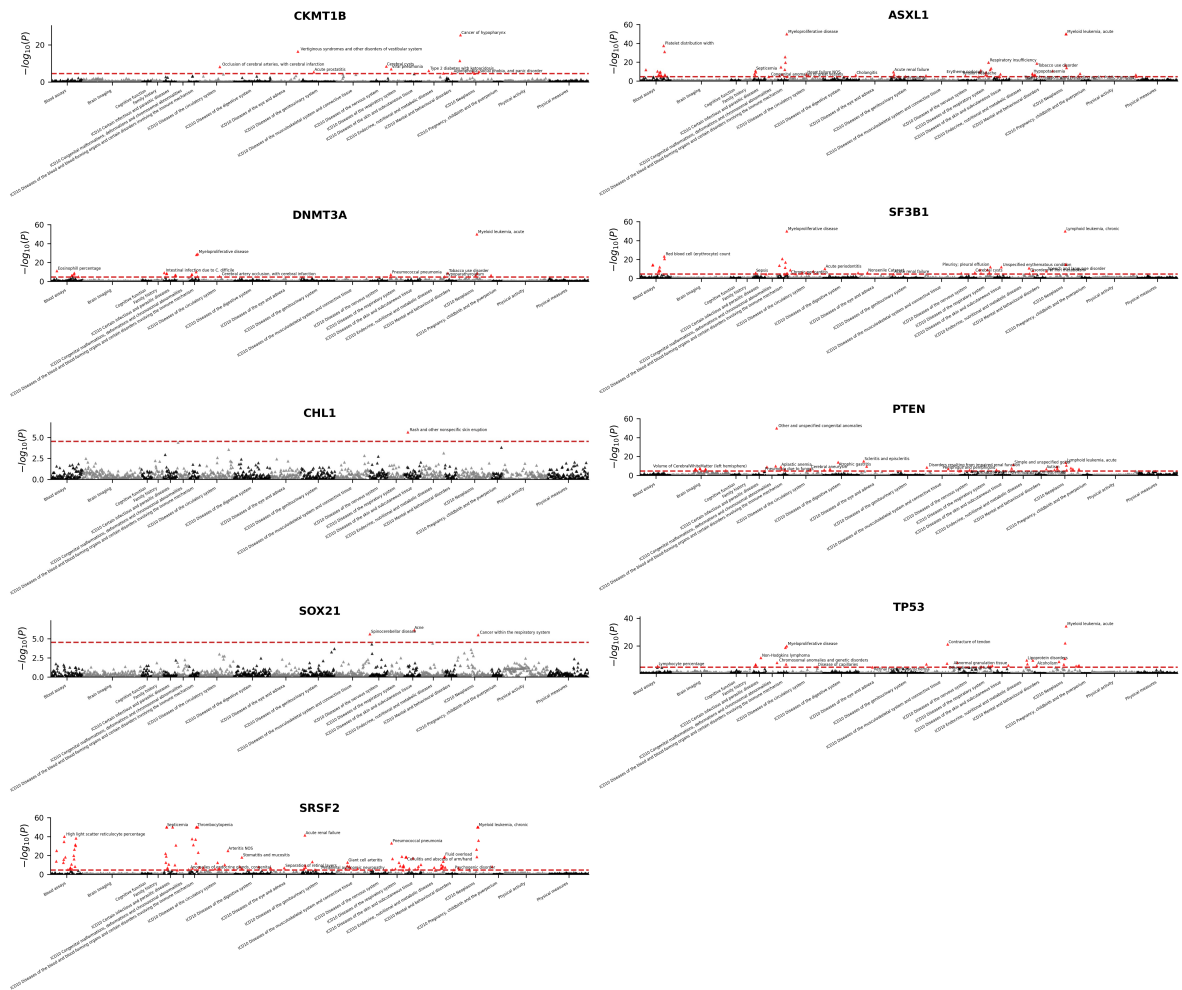


(C) Missense (REVEL)



492
493
494

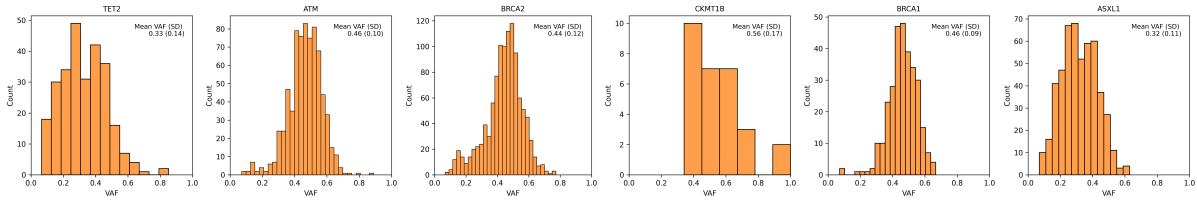
495 **Supplementary Figure 8. Phenome-wide association of the burden of rare variants at the**
496 **nine novel genes identified in our burden test. Variants considered correspond to loss-of-**
497 **function and Alpha missense defined variants. P-values less than 1.0×10^{-50} are capped at 50.**
498



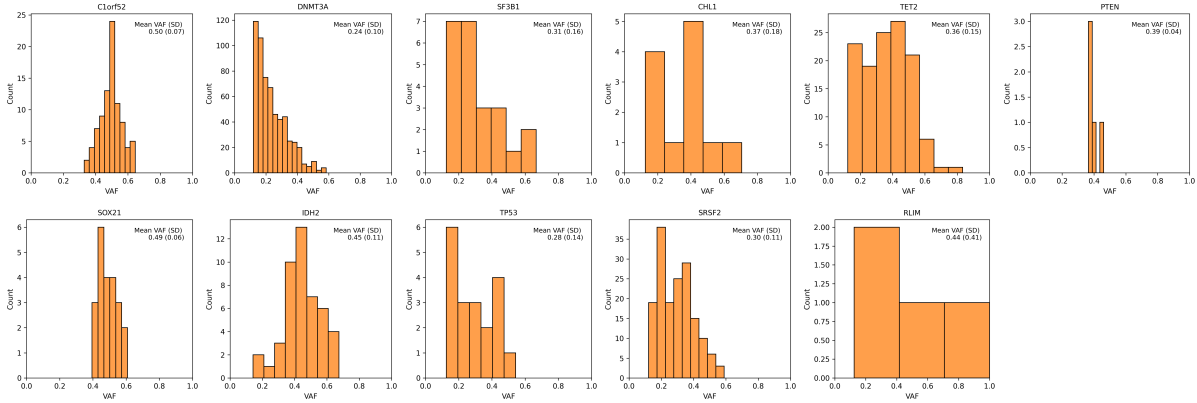
499
500
501

502 **Supplementary Figure 9. Variant allelic fraction distribution per gene for variants**
503 **considered in each category: Loss-of-function (A) and Alpha Missense (B).**
504

(A) Loss of Function



(B) Alpha Missense



505

Table 1. Significant genes for rare variants association with burden and SKAT-O tests ($p < 7.4 \times 10^{-7}$). Gene names in bold font represent novel associations.

Variant Class	Chr	Gene	# of variants	# of carriers	Burden p -value	SKAT-O p -value
LoF	4	<i>TET2</i>	243	563	3.8×10^{-30}	3.4×10^{-54}
	11	<i>ATM</i>	247	1,170	6.0×10^{-10}	4.8×10^{-11}
	13	<i>BRCA2</i>	245	1,271	1.3×10^{-34}	1.2×10^{-42}
	15	<i>CKMT1B</i>	15	40	4.5×10^{-7}	1.8×10^{-6}
	17	<i>BRCA1</i>	120	456	4.9×10^{-12}	1.3×10^{-11}
	20	<i>ASXL1</i>	72	533	2.3×10^{-44}	2.9×10^{-46}
Alpha Missense	1	<i>C1orf52</i>	23	175	4.5×10^{-5}	2.1×10^{-7}
	2	<i>DNMT3A</i>	167	1,229	1.4×10^{-9}	1.8×10^{-10}
	2	<i>SF3B1</i>	64	195	6.7×10^{-12}	2.5×10^{-16}
	3	<i>CHL1</i>	33	3,666	5.0×10^{-7}	1.1×10^{-6}
	4	<i>TET2</i>	159	826	4.2×10^{-7}	7.4×10^{-7}
	10	<i>PTEN</i>	50	71	1.0×10^{-8}	3.9×10^{-11}
	13	<i>SOX21</i>	52	463	2.9×10^{-8}	6.7×10^{-9}
	15	<i>IDH2</i>	89	349	1.9×10^{-4}	5.3×10^{-39}
	17	<i>TP53</i>	35	90	3.1×10^{-15}	2.5×10^{-16}
17	<i>SRSF2</i>	14	141	9.8×10^{-89}	8.3×10^{-108}	
X	<i>RLIM</i>	25	51	8.8×10^{-7}	3.7×10^{-9}	

LoF: Loss of Function; Chr: chromosome

Table 2. Lead variant association per gene among significant genes in the burden and SKAT-O tests. Only variants with at least 3 minor alleles are reported.

Variant Class	Chr	Gene	Variant	MA	MAC	AM	HR	<i>p</i> -value	Reported
LoF	4	<i>TET2</i>	rs370735654	T	17	-	7.0	5.6×10^{-9}	-
	11	<i>ATM</i>	rs587779834	A	113	-	2.5	2.7×10^{-5}	-
	13	<i>BRCA2</i>	rs80359520	C	10	-	6.1	1.8×10^{-6}	Breast Cancer [66]
	15	<i>CKMT1B</i>	rs1355844751	T	8	-	4.9	5.9×10^{-3}	-
	17	<i>BRCA1</i>	rs80357508	C	44	-	2.8	3.5×10^{-3}	-
	20	<i>ASXL1</i>	rs750318549	AG	201	-	2.5	2.3×10^{-15}	-
Alpha Missense	1	<i>C1orf52</i>	rs769009649	A	62	0.876	3.2	7.4×10^{-7}	-
	2	<i>DNMT3A</i>	rs147001633	T	269	0.995	1.8	3.7×10^{-5}	Leukemia [67, 68]
	2	<i>SF3B1</i>	rs377023736	A	12	0.999	6.0	8.5×10^{-8}	-
	3	<i>CHL1</i>	rs116421102	C	1,842	0.745	1.3	2.5×10^{-5}	-
	4	<i>TET2</i>	rs76428136	G	5	0.913	8.1	2.9×10^{-4}	-
	10	<i>PTEN</i>	rs587782350	T	3	0.941	20.4	2.6×10^{-3}	-
	13	<i>SOX21</i>	rs1172148601	A	67	0.856	2.3	1.7×10^{-3}	-
	15	<i>IDH2</i>	rs121913502	T	45	0.987	5.7	1.0×10^{-20}	Leukemia [68, 69]
	17	<i>TP53</i>	rs11540652	T	5	0.996	10.0	6.6×10^{-5}	Gastric Cancer [70], Ovarian Cancer [71]
	17	<i>SRSF2</i>	rs751713049	T	51	0.982	5.8	1.9×10^{-26}	-
X	<i>RLIM</i>	X:74592998:C:A	A	3	0.970	3.9	1.3×10^{-3}	-	

Chr: chromosome; MAC: minor allele count; AM: AlphaMissense score; HR: hazard ratio

Supplementary Table 1. Demographics of European ancestry in the analyses.

	All			Male			Female		
	Total	Living	Deceased	Total	Living	Deceased	Total	Living	Deceased
N	393,833	358,282	35,551	180,970	159,911	21,059	212,863	198,371	14,492
Last known age	70.8 ± 7.9	70.7 ± 8.0	71.2 ± 7.5	70.9 ± 8.0	70.8 ± 8.1	71.3 ± 7.4	70.7 ± 7.9	70.7 ± 7.9	71.1 ± 7.6
<i>APOE</i> ε4 carrier	113,437 (28.8%)	102,360 (28.6%)	11,077 (31.2%)	52,190 (28.8%)	45,635 (28.5%)	6,555 (31.1%)	61,247 (28.8%)	56,725 (28.6%)	4,522 (31.2%)

Supplementary Table 2. Significant genes for burden and SKAT-O association of rare variants, considering missense variants with REVEL > 75. Gene names in bold font represent novel associations in REVEL.

Variant Class	Chr	Gene	# of variants	# of carriers	Burden <i>p</i> -value	SKAT-O <i>p</i> -value
REVEL (>75)	2	<i>DNMT3A</i>	116	831	6.6×10^{-9}	3.5×10^{-8}
	5	<i>TERT</i>	31	60	2.6×10^{-3}	8.1×10^{-10}
	10	<i>PTEN</i>	42	45	6.6×10^{-8}	3.6×10^{-10}
	17	<i>TP53</i>	47	173	6.2×10^{-9}	6.5×10^{-9}

Supplementary Table 3. Significant genes for burden and SKAT-O association of rare variants in males. Genes in bold font represent novel associations in males.

Variant Class	Chr	Gene	# of variants	# of carriers	Burden p -value	SKAT-O p -value
LoF	4	<i>TET2</i>	162	280	4.1×10^{-20}	1.5×10^{-32}
	6	<i>CDKN1A</i>	8	33	1.1×10^{-4}	6.3×10^{-8}
	6	<i>PTPRK</i>	26	40	5.9×10^{-3}	4.5×10^{-7}
	11	<i>ATM</i>	318	520	2.6×10^{-10}	2.0×10^{-10}
	13	<i>BRCA2</i>	172	596	2.3×10^{-16}	1.5×10^{-19}
	20	<i>ASXL1</i>	59	347	5.4×10^{-36}	6.4×10^{-37}
Alpha Missense (>70)	1	<i>Clorf52</i>	19	76	2.2×10^{-5}	2.7×10^{-10}
	1	<i>COA7</i>	8	11	1.0×10^{-4}	3.1×10^{-8}
	2	<i>SF3B1</i>	43	122	1.6×10^{-11}	2.8×10^{-14}
	4	<i>TET2</i>	107	405	1.1×10^{-8}	1.6×10^{-8}
	8	<i>TG</i>	113	657	2.4×10^{-7}	1.2×10^{-6}
	15	<i>IDH2</i>	58	171	2.2×10^{-3}	5.5×10^{-29}
	17	<i>TP53</i>	24	48	1.5×10^{-9}	1.1×10^{-9}
17	<i>SRSF2</i>	10	104	5.4×10^{-63}	1.7×10^{-70}	
REVEL (>75)	1	<i>NMNAT2</i>	21	34	1.2×10^{-4}	1.9×10^{-8}
	10	<i>PITRM1</i>	6	10	1.5×10^{-6}	3.6×10^{-8}

Supplementary Table 4. Significant genes for burden and SKAT-O association of rare variants in females. Genes in bold font represent novel associations in females.

Variant Class	Chr	Gene	# of variants	# of carriers	Burden p -value	SKAT-O p -value
LoF	4	<i>TET2</i>	151	283	7.8×10^{-11}	5.1×10^{-24}
	13	<i>BRCA2</i>	182	675	1.2×10^{-20}	4.8×10^{-27}
	17	<i>BRCA1</i>	80	217	3.6×10^{-12}	2.4×10^{-11}
	20	<i>ASXL1</i>	46	186	1.1×10^{-6}	2.0×10^{-7}
Alpha Missense (>70)	2	<i>DNMT3A</i>	135	671	7.6×10^{-9}	2.3×10^{-10}
	5	<i>TERT</i>	22	27	1.7×10^{-3}	2.2×10^{-7}
	13	<i>SOX21</i>	34	251	2.0×10^{-7}	1.6×10^{-7}
	17	<i>SRSF2</i>	10	37	3.3×10^{-16}	8.4×10^{-27}
	17	<i>TP53</i>	23	52	1.1×10^{-7}	1.8×10^{-8}
	X	<i>PORCN</i>	11	32	5.6×10^{-4}	3.7×10^{-7}
REVEL (>75)	2	<i>UGT1A8</i>	2	18	2.8×10^{-7}	2.8×10^{-7}
	21	<i>OLIG1</i>	7	18	1.3×10^{-5}	7.0×10^{-7}

Supplementary Table 5. Lead variant association per gene among significant genes in the burden and SKAT-O tests. Only significant variant associations with at least 3 minor allele counts per gene are reported in this table.

Variant Class	Chr	Gene	Variant	MA	MAC	AM	HR	<i>p</i> -value	Reported
REVEL (>75)	2	<i>DNMT3A</i>	rs367909007	G	14	0.983	4.3	1.2×10^{-3}	-
	5	<i>TERT</i>	rs1043358053	C	5	0.926	11.9	7.4×10^{-7}	-
	10	<i>PTEN</i>	rs587782350	C	3	0.941	20.4	2.6×10^{-3}	-
	17	<i>TP53</i>	rs11540652	T	5	0.996	10.0	6.6×10^{-5}	Gastric Cancer [70], Ovarian Cancer [71]

Supplementary Table 6. Mean variant allelic fraction per gene across participants included in the corresponding gene-level Burden/SKAT-O analysis.

Variant Class	Chr	Gene	# of subjects	# of variants	Mean VAF (SD)
LoF	4	<i>TET2</i>	266	133	0.33 (0.14)
	11	<i>ATM</i>	734	128	0.46 (0.10)
	13	<i>BRCA2</i>	1,061	162	0.44 (0.12)
	15	<i>CKMT1B</i>	29	8	0.56 (0.17)
	17	<i>BRCA1</i>	302	74	0.46 (0.09)
	20	<i>ASXL1</i>	502	30	0.32 (0.11)
Alpha Missense	1	<i>C1orf52</i>	87	11	0.50 (0.07)
	2	<i>DNMT3A</i>	593	33	0.24 (0.10)
	2	<i>SF3B1</i>	23	5	0.31 (0.16)
	3	<i>CHL1</i>	12	3	0.37 (0.18)
	4	<i>TET2</i>	123	41	0.36 (0.15)
	10	<i>PTEN</i>	5	3	0.39 (0.04)
	13	<i>SOX21</i>	22	6	0.49 (0.06)
	15	<i>IDH2</i>	46	14	0.45 (0.11)
	17	<i>TP53</i>	19	7	0.28 (0.14)
17	<i>SRSF2</i>	164	6	0.30 (0.11)	
X	<i>RLIM</i>	4	2	0.44 (0.41)	

Acknowledgements

This research has been conducted using the UK Biobank Resource under application number 45420. We thank all the participants and researchers of UK Biobank for making these data open and accessible to the research community.

Authors' contributions

J.P conducted all the analyses, prepared all figures and wrote the manuscript. A.P.T contributed to the manuscript writing. L.T provided critical comment on the manuscript. M.D.G and Y.L.G planned, organized and supervised the entire study and revised the manuscript. All authors have approved the submitted version.

Funding

This research was supported by the Dean's Postdoctoral Fellowship at the School of Medicine, Stanford University. Additionally, this research was partially supported by the Biostatistics Shared Resource (B-SR) of the NCI-sponsored Stanford Cancer Institute: P30CA124435 and by the following NIH funding source of Stanford's Center for Clinical and Translational Education and Research award, under the Biostatistics, Epidemiology and Research Design (BERD) Program: 1UM1TR004921-01.

Data availability

GWAS summary statistics for this study are available in the GWAS Catalog. Data supporting the findings of this study are available from the UK Biobank (UKB). Access to these data is available from the authors with UKB permission

Code availability

The codes used for analyses in the present study are available at the following link:

<https://github.com/Junkkkk/Lifespan-studies>

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Passarino, G., F. De Rango, and A. Montesanto, *Human longevity: Genetics or Lifestyle? It takes two to tango*. Immunity & Ageing, 2016. **13**: p. 1-6.
2. Ruby, J.G., et al., *Estimates of the Heritability of Human Longevity Are Substantially Inflated due to Assortative Mating*. Genetics, 2018. **210**(3): p. 1109-1124.
3. v, B.H.J., et al., *Genetic influence on human lifespan and longevity*. Hum Genet, 2006. **119**(3): p. 312-21.
4. van den Berg, N., et al., *Longevity defined as top 10% survivors and beyond is transmitted as a quantitative genetic trait*. Nature Communications, 2019. **10**(1): p. 35.
5. Ryu, S., et al., *Genetic landscape of APOE in human longevity revealed by high-throughput sequencing*. Mech Ageing Dev, 2016. **155**: p. 7-9.
6. Sebastiani, P., et al., *APOE Alleles and Extreme Human Longevity*. J Gerontol A Biol Sci Med Sci, 2019. **74**(1): p. 44-51.
7. Joshi, P.K., et al., *Variants near CHRNA3/5 and APOE have age- and sex-related effects on human lifespan*. Nature Communications, 2016. **7**(1): p. 11174.
8. Joshi, P.K., et al., *Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity*. Nature Communications, 2017. **8**(1): p. 910.
9. Liu, J.Z., et al., *The burden of rare protein-truncating genetic variants on human lifespan*. Nature Aging, 2022. **2**(4): p. 289-294.
10. Liu, J.Z., Y. Erlich, and J.K. Pickrell, *Case-control association mapping by proxy using family history of disease*. Nature Genetics, 2017. **49**(3): p. 325-331.
11. Bae, H., et al., *A genome-wide association study of 2304 extreme longevity cases identifies novel longevity variants*. International Journal of Molecular Sciences, 2022. **24**(1): p. 116.
12. Sebastiani, P., et al., *Four Genome-Wide Association Studies Identify New Extreme Longevity Variants*. The Journals of Gerontology: Series A, 2017. **72**(11): p. 1453-1464.
13. Deelen, J., et al., *A meta-analysis of genome-wide association studies identifies multiple longevity genes*. Nature Communications, 2019. **10**(1): p. 3669.
14. Therneau, T.M., P.M. Grambsch, and T.R. Fleming, *Martingale-based residuals for survival models*. Biometrika, 1990. **77**(1): p. 147-160.
15. Gagliano Taliun, S.A., et al., *Exploring and visualizing large-scale genetic associations by using PheWeb*. Nature Genetics, 2020. **52**(6): p. 550-552.
16. Baer, C., et al., *"Somatic" and "pathogenic" - is the classification strategy applicable in times of large-scale sequencing?* Haematologica, 2019. **104**(8): p. 1515-1520.
17. Sudlow, C., et al., *UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age*. PLoS Med, 2015. **12**(3): p. e1001779.
18. Bycroft, C., et al., *The UK Biobank resource with deep phenotyping and genomic data*. Nature, 2018. **562**(7726): p. 203-209.
19. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses*. The American journal of human genetics, 2007. **81**(3): p. 559-575.

20. Cox, D.R., *Regression models and life-tables*. Journal of the Royal Statistical Society: Series B (Methodological), 1972. **34**(2): p. 187-202.
21. Therneau, T., *A package for survival analysis in S. R* package version, 2015. **2**(7): p. 2014.
22. Pruim, R.J., et al., *LocusZoom: regional visualization of genome-wide association scan results*. Bioinformatics, 2010. **26**(18): p. 2336-2337.
23. Mohammadi Pejman 5 6 Park YoSon 11 Parsana Princy 12 Segrè Ayellet V. 1 Strober Benjamin J. 9 Zappala Zachary 7 8, G.C.L.a.A.F.B.A.A.C.S.E.D.J.R.H.Y.J.B., et al., *Genetic effects on gene expression across human tissues*. Nature, 2017. **550**(7675): p. 204-213.
24. Giambartolomei, C., et al., *Bayesian test for colocalisation between pairs of genetic association studies using summary statistics*. PLoS genetics, 2014. **10**(5): p. e1004383.
25. Liu, B., et al., *Abundant associations with gene expression complicate GWAS follow-up*. Nature Genetics, 2019. **51**(5): p. 768-769.
26. Szustakowski, J.D., et al., *Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank*. Nature genetics, 2021. **53**(7): p. 942-948.
27. McLaren, W., et al., *The Ensembl Variant Effect Predictor*. Genome Biology, 2016. **17**(1): p. 122.
28. Cheng, J., et al., *Accurate proteome-wide missense variant effect prediction with AlphaMissense*. Science, 2023. **381**(6664): p. eadg7492.
29. Ioannidis, N.M., et al., *REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants*. Am J Hum Genet, 2016. **99**(4): p. 877-885.
30. Lee, S., et al., *Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies*. Am J Hum Genet, 2012. **91**(2): p. 224-37.
31. Wu, M.C., et al., *Rare-variant association testing for sequencing data with the sequence kernel association test*. Am J Hum Genet, 2011. **89**(1): p. 82-93.
32. Wu, P., et al., *Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation*. JMIR medical informatics, 2019. **7**(4): p. e14325.
33. Bellenguez, C., et al., *New insights into the genetic etiology of Alzheimer's disease and related dementias*. Nature Genetics, 2022. **54**(4): p. 412-436.
34. Mahley, R.W., *Apolipoprotein E: from cardiovascular disease to neurodegenerative disorders*. Journal of Molecular Medicine, 2016. **94**(7): p. 739-746.
35. Du, Q., et al., *Epigenetic silencing ZSCAN23 promotes pancreatic cancer growth by activating Wnt signaling*. Cancer Biology & Therapy, 2024. **25**(1): p. 2302924.
36. Lee, M.-G. and Y.H. Lee, *A meta-analysis examining the association between the MUC5B rs35705950 T/G polymorphism and susceptibility to idiopathic pulmonary fibrosis*. Inflammation Research, 2015. **64**(6): p. 463-470.
37. Nakano, Y., et al., *MUC5B promoter variant rs35705950 affects MUC5B expression in the distal airways in idiopathic pulmonary fibrosis*. American journal of respiratory and critical care medicine, 2016. **193**(4): p. 464-466.
38. van Moorsel, C.H., et al., *The MUC5B promoter polymorphism associates with severe*

- COVID-19 in the European population*. *Frontiers in medicine*, 2021. **8**: p. 668024.
39. Verma, A., et al., *A MUC5B gene polymorphism, rs35705950-T, confers protective effects against COVID-19 hospitalization but not severe disease or mortality*. *American journal of respiratory and critical care medicine*, 2022. **206**(10): p. 1220-1229.
 40. Quivoron, C., et al., *TET2 Inactivation Results in Pleiotropic Hematopoietic Abnormalities in Mouse and Is a Recurrent Event during Human Lymphomagenesis*. *Cancer Cell*, 2011. **20**(1): p. 25-38.
 41. Abdel-Wahab, O., et al., *ASXL1 Mutations Promote Myeloid Transformation through Loss of PRC2-Mediated Gene Repression*. *Cancer Cell*, 2012. **22**(2): p. 180-193.
 42. Yang, L., R. Rau, and M.A. Goodell, *DNMT3A in haematological malignancies*. *Nature Reviews Cancer*, 2015. **15**(3): p. 152-165.
 43. Wang, L., et al., *<i>SF3B1</i> and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia*. *New England Journal of Medicine*, 2011. **365**(26): p. 2497-2506.
 44. Kim, E., et al., *SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition*. *Cancer Cell*, 2015. **27**(5): p. 617-630.
 45. Cremona, C.A. and A. Behrens, *ATM signalling and cancer*. *Oncogene*, 2014. **33**(26): p. 3351-3360.
 46. Wooster, R., et al., *Identification of the breast cancer susceptibility gene BRCA2*. *Nature*, 1995. **378**(6559): p. 789-792.
 47. Ford, D., et al., *Risks of cancer in BRCA1-mutation carriers*. *The Lancet*, 1994. **343**(8899): p. 692-695.
 48. Johnsen, S.A., et al., *Regulation of Estrogen-Dependent Transcription by the LIM Cofactors CLIM and RLIM in Breast Cancer*. *Cancer Research*, 2008. **69**(1): p. 128-136.
 49. Gao, R., et al., *E3 Ubiquitin Ligase RLIM Negatively Regulates c-Myc Transcriptional Activity and Restrains Cell Proliferation*. *PLOS ONE*, 2016. **11**(9): p. e0164086.
 50. Yin, Y. and W.H. Shen, *PTEN: a new guardian of the genome*. *Oncogene*, 2008. **27**(41): p. 5443-5453.
 51. Yang, H., et al., *IDH1 and IDH2 Mutations in Tumorigenesis: Mechanistic Insights and Clinical Perspectives*. *Clinical Cancer Research*, 2012. **18**(20): p. 5562-5571.
 52. Yang, C.-M., et al., *Aberrant DNA hypermethylation-silenced SOX21-AS1 gene expression and its clinical importance in oral cancer*. *Clinical Epigenetics*, 2016. **8**(1): p. 129.
 53. Zhang, X., et al., *Long noncoding RNA SOX21-AS1 promotes cervical cancer progression by competitively sponging miR-7/VDAC1*. *Journal of Cellular Physiology*, 2019. **234**(10): p. 17494-17504.
 54. Sheng, X.-Y., et al., *Long-Chain Non-Coding SOX21-AS1 Promotes Proliferation and Migration of Breast Cancer Cells Through the PI3K/AKT Signaling Pathway*. *Cancer Management and Research*, 2020. **12**(null): p. 11005-11014.
 55. Shi, H., et al., *CKMT1B is a potential prognostic biomarker and associated with immune infiltration in Lower-grade glioma*. *PLOS ONE*, 2021. **16**(1): p. e0245524.
 56. Senchenko, V.N., et al., *Differential expression of CHL1 gene during development of major*

- human cancers*. PloS one, 2011. **6**(3): p. e15612.
57. Baird, D.M., *Variation at the TERT locus and predisposition for cancer*. Expert Reviews in Molecular Medicine, 2010. **12**: p. e16.
 58. Vinagre, J., et al., *Frequency of TERT promoter mutations in human cancers*. Nature Communications, 2013. **4**(1): p. 2185.
 59. Zenin, A., et al., *Identification of 12 genetic loci associated with human healthspan*. Communications Biology, 2019. **2**(1): p. 41.
 60. Kibel, A.S., et al., *CDKN1A and CDKN1B Polymorphisms and Risk of Advanced Prostate Carcinoma*. Cancer Research, 2003. **63**(9): p. 2033-2036.
 61. Li, S., et al., *miR-3619-5p inhibits prostate cancer cell growth by activating CDKN1A expression*. Oncology Reports, 2017. **37**(1): p. 241-248.
 62. Lombardi, M.P., et al., *Mutation update for the PORCN gene*. Human Mutation, 2011. **32**(7): p. 723-728.
 63. Covey, T.M., et al., *PORCN moonlights in a Wnt-independent pathway that regulates cancer cell proliferation*. Plos one, 2012. **7**(4): p. e34532.
 64. Zhao, F., et al., *The function of uterine UDP-glucuronosyltransferase 1A8 (UGT1A8) and UDP-glucuronosyltransferase 2B7 (UGT2B7) is involved in endometrial cancer based on estrogen metabolism regulation*. Hormones, 2020. **19**: p. 403-412.
 65. Thibaudeau, J., et al., *Characterization of common UGT1A8, UGT1A9, and UGT2B7 variants with different capacities to inactivate mutagenic 4-hydroxylated metabolites of estradiol and estrone*. Cancer research, 2006. **66**(1): p. 125-133.
 66. Yang, X.R., et al., *Prevalence and spectrum of germline rare variants in BRCA1/2 and PALB2 among breast cancer cases in Sarawak, Malaysia*. Breast Cancer Res Treat, 2017. **165**(3): p. 687-697.
 67. Clay-Gilmour, A., et al., *Pathogenic and likely pathogenic germline variation in patients with myeloid malignancies and their unrelated HLA-matched hematopoietic stem cell donors*. 2024.
 68. Petiti, J., et al., *Comprehensive Molecular Profiling of NPM1-Mutated Acute Myeloid Leukemia Using RNAseq Approach*. International Journal of Molecular Sciences, 2024. **25**(7): p. 3631.
 69. Bayram, D.M., F.M. Lafta, and B.F. Matti, *Impact of IDH Mutations on DNA Methylation of Acute Myeloid Leukemia Related Genes: A Review Article*. Journal of the Faculty of Medicine Baghdad, 2024. **66**(1): p. 116-125.
 70. Hnatyszyn, A., et al., *Mutations in Helicobacter pylori infected patients with chronic gastritis, intestinal type of gastric cancer and familial gastric cancer*. Hereditary Cancer in Clinical Practice, 2024. **22**(1): p. 9.
 71. Richau, C.S., et al., *BRCA1, BRCA2, and TP53 germline and somatic variants and clinicopathological characteristics of Brazilian patients with epithelial ovarian cancer*. Cancer Medicine, 2024.