
1 **Exome-wide association studies discover germline mutation patterns**
2 **and identify high-risk populations in human cancers**

3 **Running head:** Exome-wide association studies (ExWAS) in pan-cancer

4

5 Sipeng Shen, PhD^{1,2,3*}; Yunke Jiang, MS¹; Guanrong Wang, MS^{1,4}; Hongru Li, MS¹;
6 Dongfang You, PhD^{1,3}; Weiwei Duan, PhD⁶; Ruyang Zhang, PhD^{1,5}; Yongyue Wei,
7 PhD^{1,3}; Hongbing Shen^{2,7}, PhD; Zhibin Hu, PhD^{2,7}; David C. Christiani, MD^{8,9}; Yang
8 Zhao, PhD^{1,5*}; Feng Chen, PhD^{1,2,3*}

9

10 ¹Department of Biostatistics, Center for Global Health, School of Public Health,
11 Nanjing Medical University, Nanjing 211166, China

12 ²Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Jiangsu
13 Collaborative Innovation Center for Cancer Personalized
14 Medicine, Nanjing Medical University, 211166, Nanjing, China

15 ³China International Cooperation Center of Environment and Human Health, Nanjing
16 Medical University

17 ⁴Department of Epidemiology, Jiangsu Health Development Research Center, Jiangsu
18 Province, Nanjing 210036, China

19 ⁵Key Laboratory of Biomedical Big Data of Nanjing Medical University, Nanjing
20 211166, China

21 ⁶Department of Bioinformatics, School of Biomedical Engineering and Informatics,
22 Nanjing Medical University, Nanjing, Jiangsu 211166, China

23 ⁷Department of Epidemiology, Center for Global Health, School of Public Health,

24 Nanjing Medical University, Nanjing 211166, China, Nanjing, Jiangsu 211166, China

25 ⁸Department of Environmental Health, Harvard T.H. Chan School of Public Health,
26 Harvard University, Boston, MA 02115, USA

27 ⁹Pulmonary and Critical Care Division, Massachusetts General Hospital, Department
28 of Medicine, Harvard Medical School, Boston, MA 02114, USA

29

30 ***Send correspondence to:**

31 Dr. Feng Chen, SPH Building Room 412, 101 Longmian Avenue, Nanjing, Jiangsu
32 211166, China. Email: fengchen@njmu.edu.cn

33 Dr. Yang Zhao, SPH Building Room 404, 101 Longmian Avenue, Nanjing, Jiangsu
34 211166, China. Email: zhaoyang@njmu.edu.cn

35 Dr. Sipeng Shen, SPH Building Room 406, 101 Longmian Avenue, Nanjing, Jiangsu
36 211166, China. Email: sshen@njmu.edu.cn

37

38 **Consent for publication**

39 All authors have reviewed and approved this manuscript.

40

41 **Conflicts of interest statements/Financial Disclosure statement:** The authors report
42 no conflicts of interest.

43 **Funding**

44 This study was supported by the National Natural Science Foundation of China
45 (82103946 to S.S., 82173620 to Y.Z., 81530088 to F.C.), National Key Research and
46 Development Program of China (2016YFE0204900 to F.C.), Natural Science

47 Foundation of the Jiangsu Higher Education Institutions of China (21KJB330004 to
48 S.S.).

49 **Author contributions**

50 SS, YZ, and FC contributed to the study design. SS and YJ contributed to data
51 collection. SS performed statistical analyses and interpretation. SS drafted the
52 manuscript. GW, HL, DY, WD, RZ, YW, HS, ZH, and DC revised the final
53 manuscript. All authors approved the final version of the manuscript.

54

55

56 **Abstract**

57 Genome-wide association studies have discovered numerous common variants
58 associated with human cancers. However, the contribution of exome-wide rare
59 variants to cancers remains largely unexplored, especially for the protein-coding
60 variants. The UK Biobank provides detailed cancer follow-up information linked to
61 whole-exome sequencing (WES) for approximately 450,000 participants, offering an
62 unprecedented opportunity to evaluate the effect of exome variation on pan-cancer.
63 Here, we performed exome-wide association studies (ExWAS) based on single variant
64 levels and gene levels to detect their associations across 20 primary cancer types in
65 the discovery set (WES-300k, N = 284,456) and replication set (WES-150k, N =
66 143,478), separately. The ExWAS detected 143 independent variants at variant-level
67 and 49 genes at gene-level, while nine variants and eight genes were shared across
68 cancers. In the cross-trait meta-analysis, we identified 239 additional independent
69 pleiotropic variants, mapping to the genes which were functional through trans-omics
70 analyses in transcriptomics and proteomics. Further, we developed exome-wide risk
71 scores (ERS) to identify high-risk populations based on rare variants with minor allele
72 frequency (MAF) < 0.05. The ERS had satisfactory performance in cancer risk
73 stratification, especially for the extremely high-risk persons (top 5% ERS) that were
74 frequently risk allele carriers. The ERS (median C-index (IQR): 0.655 (0.636-0.667))
75 outperforms the traditional polygenic risk score (PRS) (median C-index (IQR): 0.585
76 (0.572-0.614)) for discrimination in the replication set. Our findings offer further
77 insight into the genetic architecture of human exomes for cancer susceptibility.

78 **Keywords:** exome-wide association studies, pan-cancer, rare variants, polygenic risk
79 score, cross-trait

80

81 **Introduction**

82 Cancer ranks as a leading cause of death and a critical barrier to increasing life
83 expectancy worldwide ¹. Population-based early screening approaches showed a
84 remarkable reduction in cancer mortality ², such as low-dose computed tomography
85 (CT) screening for lung cancer ^{3,4}. Considering the cost-effectiveness balance, it is
86 generally agreed that screening should be limited to the high-risk population.
87 However, precisely identifying high-risk persons is still challenging, while cancer is a
88 complex disease that derives from environmental exposure and inherent heredity ⁵.

89 Cancer shows substantial heritability from genetic variants ⁶. Genome-wide
90 association studies (GWAS) have identified numerous associations of genome-wide
91 significance between genetic variants and common diseases ⁷. However, common
92 single nucleotide polymorphisms (SNPs) identified in GWAS explain only a small
93 fraction of heritability, which might be limited to the coverage of SNP arrays ⁸.
94 Exome-wide association studies (ExWAS) have shown that rare coding variants tend
95 to have larger phenotypic effects than common SNPs and contribute an essential
96 component of heritability ⁹. Due to the effect allele frequency being generally low in
97 ExWAS, the sample size should be large (e.g., $n > 100,000$) to guarantee the statistical
98 power, especially for the rare variants ^{10,11}.

99 Moreover, it is widely recognized that the polygenic risk score (PRS) is a powerful
100 tool to discriminate the high-risk population susceptible to specific cancer ^{12,13}.
101 However, most PRSs are generated using common variants derived from the SNP
102 array, which ignore the rare variants with larger effects, especially those located on
103 exomes with remarkable biological significance ^{14,15}. Thus, the rare variants might
104 provide complementarity value for cancer risk stratification based on traditional PRS.

105 The UK Biobank (UKB) is a powerful resource for evaluating the associations
106 between coding variants and human diseases because of its large sample size with

107 high-quality whole-exome sequencing (WES) data ($n \approx 450,000$)^{16,17}. In our study, we
108 investigated the landscape of genetic variants with multiple primary cancers through
109 UKB WES project. Further, we leveraged the rare variants to improve the risk
110 stratification models to identify the high-risk population.

111

112 **Results**

113 *Exome-wide association study for single variants*

114 Our study included the whole-exome sequencing 450k (WES-450k in data-field
115 23148) population of European ancestry in UK Biobank (Table S1). To ensure the
116 robustness of the results, we conducted a two-stage association study in two separate
117 datasets: discovery set (interim WES-300k in data-field 23146, $N = 284,456$) and
118 replication set (the remaining WES-150k, $N = 143,478$). Overall, our study included
119 20 cancer types containing 106,836 primary cancer cases and 321,098 shared
120 cancer-free controls. The number of cancer cases ranged from 773 (thyroid cancer) to
121 32,307 (skin cancer). 1,769,329 exome single nucleotide variants (SNVs) annotated
122 putative loss-of-function (LoF), missense, and synonymous passed the quality control
123 procedures.

124 For single-variant analyses, 306,031 variants with minor allele count (MAC) ≥ 10 were
125 tested in ExWAS. The genomic inflation factor (GIF) values suggested no obvious
126 population stratification (Figure S1). When combining all the P values of pan-cancer,
127 the GIF was 1.105, indicating the existing pleiotropic effects (Figure S2).

128 Among the 20 cancer types, ExWAS detected 255 signals from 242 variants that
129 passed the genome-wide significance level ($P < 5 \times 10^{-8}$) in 12 cancer types. After LD
130 pruning, we observed 153 independent signals across 64 chromosome regions

131 (LD- $r^2 < 0.5$) (Figure 1a, Table S2). The top variant with maximum association signals
132 was rs555607708 (22:28695868:AG:A, frameshift variant of *CHEK2*, HGVS:
133 p.Thr367fs, effect allele frequency: 0.22%), which was associated with three cancer
134 types including breast [OR (95% CI): 4.15 (3.16-5.43), $P = 7.24 \times 10^{-25}$], prostate [OR
135 (95% CI): 2.40 (1.82-3.16), $P = 3.48 \times 10^{-10}$] and leukemia [OR (95% CI): 8.48
136 (4.12-17.48), $P = 1.21 \times 10^{-9}$].

137 Additionally, eight independent variants had significant associations with at least two
138 cancer types, leading by rs6998061 (OR=0.82~0.89, missense variant in *POU5F1B*),
139 rs16891982 (OR=1.42~1.46, missense variant in *SLC45A2*), rs387907272
140 (beta=24.1~76.2, LoF variant in *MYD88*), rs3787220 (OR=0.85~0.88, synonymous
141 variant in *NCOA6*), rs77681059 (OR=1.11~1.32, missense variant in *TUBB3*),
142 rs1805007 (OR=1.34~1.64, missense variant in *MC1R*), rs56288641 (OR=1.28~1.59,
143 missense variant in *VPS9DI*), rs1126809 (OR=1.13~1.23, missense variant in *TYR*).
144 More than half of the variants were shared between skin cancer and melanoma,
145 including three variants located on 16q24.3. The remaining three variants were shared
146 among breast, prostate, leukemia, colorectal cancers, and non-Hodgkin's lymphoma
147 (NHL) (Table S3).

148 ***Exome-wide association study based on gene-level***

149 In addition to the variant-level analyses, we performed gene-based association studies
150 to capture the effects of ultra-rare variants. We used different genetic models to detect
151 the signals, according to minor allele frequency threshold (0.05, 0.01, 0.001) and
152 variant functional annotation (LoF, LoF+missense, LoF+missense+synonymous).
153 After Bonferroni correction, 49 genes were considered significant ($P < 2.5 \times 10^{-6}$)
154 (Figure 1b, Table S4). The top genes with the highest hit frequency were *CHEK2*
155 [breast (OR = 1.03, $P = 1.08 \times 10^{-22}$), prostate (OR = 1.02, $P = 1.28 \times 10^{-12}$), leukemia
156 (OR = 1.04, $P = 1.55 \times 10^{-13}$), *BRCA2* [breast (OR = 1.08, $P = 1.43 \times 10^{-38}$), ovary (OR

157 = 1.14, $P = 1.94 \times 10^{-30}$), prostate (OR = 1.04, $P = 5.98 \times 10^{-8}$), and *ATM* [breast (OR =
158 1.10, $P = 4.33 \times 10^{-10}$), prostate (OR = 1.04, $P = 1.89 \times 10^{-9}$), pancreas (OR = 1.04, $P =$
159 2.67×10^{-7})], which were associated with three cancer types. *VPS9D1*, *SLC45A2*,
160 *BRCA1*, *MSH6*, and *MC1R* were associated with two cancer types (Table S5).

161 We summarized the association results of genes that reached significance level in at
162 least two cancer types from variant-level and gene-level analyses (Figure 1c). Four
163 genes, including *CHEK2*, *ATM*, and *BRCA1/2*, showed a close relationship with
164 human cancers.

165 *Shared genetic cross-trait meta-analyses identify pleiotropic signals*

166 To identify additional potential pleiotropic variants associated with multiple cancers,
167 we performed a cross-trait meta-analysis using Association analysis based on
168 SUBSETs (ASSET). 1,572 variants that reached $P < 10^{-4}$ in any cancer type were
169 included. We identified 150 independent variants (mapping to 86 genes, 43
170 chromosome cytobands) with significant one-directional effects (Table S6) and 89
171 independent variants (mapping to 38 genes, 16 cytobands) with significant
172 bidirectional effects (Table S7) (Figure 2a).

173 We checked the pleiotropic variants in NHGRI-EBI GWAS Catalog¹⁸, a publicly
174 available database that collected all published GWAS signals phenome-wide. The due
175 date of associations collection was April 7, 2022. More than half of the identified
176 variants were not reported in GWAS Catalog [100 novel one-directional variants (66.7%
177 of all), 59 novel bidirectional variants (66.3% of all)] (Table S6-S7).

178 We also summarized the shared genetic variants across cancers identified in ASSET
179 subgroups according to their effect direction (Figure 2b). The top cancer pairs with
180 the highest shared one-directional variants were skin & melanoma (79 variants), skin
181 & thyroid (69 variants), sarcoma & skin (51 variants), NHL & skin (50 variants),
182 esophagus & skin (47 variants). The top cancer pairs with the highest heterogeneous

183 bidirectional variants were leukemia & thyroid (51 variants), liver & thyroid (50
184 variants), lung & sarcoma (48 variants), skin & liver (48 variants), lung & thyroid (48
185 variants) ([Table S8](#)).

186 ***Trans-omics functional analysis for the identified genes***

187 We performed a trans-omics analysis to integrate the identified genes from single
188 variant, gene based and cross-trait meta-analyses into transcriptomics and proteomics.
189 Gene expression was obtained from the recompute transcriptomic data of The Cancer
190 Genome Atlas (TCGA) and The Genotype-Tissue Expression (GTEx). 98 unique
191 protein-coding genes across 15 tissue types that passed quality control were included.
192 Through comparisons of gene expression in tumor and healthy normal tissues, we
193 observed remarkable differences in gene expression. The well-known *HLA* family
194 genes, *BRCA1/2*, *CHEK2*, and *TUBB3* were up-regulated in tumor tissues, while *ATM*,
195 *VPS9DI*, and *MC1R* were down-regulated. However, some genes showed
196 heterogenous trends across cancers, such as the *KRT* family, *POU5F1B*, and *TET2*
197 ([Figure 3a, 3b](#)).

198 Further, we performed a KEGG pathway enrichment analysis for the identified genes.
199 Numerous immune-related pathways were identified, such as Th1 and Th2 cell
200 differentiation ($P = 1.46 \times 10^{-3}$), Th17 cell differentiation ($P = 2.63 \times 10^{-3}$), and
201 inflammatory bowel disease ($P = 3.94 \times 10^{-4}$), as well as the classical cancer-related
202 pathways, including cell adhesion molecules ($P = 1.29 \times 10^{-5}$), platinum drug
203 resistance ($P = 6.14 \times 10^{-4}$), p53 signaling pathway ($P = 6.14 \times 10^{-4}$), and NF-kappa B
204 signaling pathway ($P = 0.018$) ([Figure 3e, Table S9](#)).

205 Proteomic data were collected across ten tissue types from The National Cancer
206 Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC). We observed
207 moderate differences by comparing protein abundance in tumor and adjacent normal

208 tissues. Interestingly, some proteins showed similar patterns with the corresponding
209 gene expression, such as *HLA* family genes and *CHEK2*. However, a reverse trend
210 was also found for some genes. For example, *BRCA2* was down-regulated in
211 proteomics of tumor tissues (Figure 3c, 3d).

212 Using the STRING database to integrate all known and predicted proteins-protein
213 interactions, we identified two large clusters: one was related to the HLA (e.g., HLA
214 family genes, *MICA*, *BTNL2*) and immune function (e.g., *CTLA4*, *CTSS*, *SAMHD1*);
215 another was related to DNA damage response and repair (e.g., *BRCA2*, *ATM*, *CHEK2*,
216 *MSH6*) (Figure 3f). In addition, three small clusters were also identified, including the
217 keratin family, BPI fold containing family B, and melanogenesis.

218 ***Identify high-risk population based on rare variants***

219 We developed exome-wide risk scores (ERS) to identify high-risk population based
220 on rare variants with minor allele frequency (MAF) < 0.05 for incident cancers. The
221 discovery set (WES-300k population) was used for risk score training, while the
222 replication set (WES-150k) was used for external validation. After screening using the
223 least absolute shrinkage and selection operator (LASSO) with 10-fold cross-validation,
224 we included a moderate number of rare variants in ERS (median [interquartile range
225 (IQR)]: 103 (82-199)), ranging from 39 variants (thyroid) to 585 variants (skin)
226 (Table S10).

227 We performed a risk stratification analysis to identify the high-risk populations
228 susceptible to specific cancer types. We defined three risk levels: extremely high-risk
229 (top 5% ERS), high-risk (5%-25% ERS), and low-risk (bottom 75% ERS), which
230 might be applicable to different medical screening strategies. The ERS could stratify
231 the cancer absolute incidence risk significantly in the replication set (Figure 4) and the
232 whole UKB-450k population (Figure S4) (all log-rank $P < 2.2 \times 10^{-16}$). Using the
233 low-risk population as the reference, the extremely high-risk persons had hazard ratios

234 (HRs) ranging from 4.52 to 9.92 [median (IQR): 7.20 (6.44-7.85)], which
235 outperformed PRS [HR median (IQR): 2.44 (1.95-2.80)]. The high-risk persons had
236 HRs ranging from 1.37 to 4.43 [median (IQR): 1.97 (1.75-2.33)], while the PRS had
237 HRs ranging from 1.11 to 2.15 with median (IQR): 1.56 (1.41-1.79) (Figure 5a, 5b).
238 Thus, the ERS had satisfactory performance in cancer risk stratification, especially for
239 the extremely high-risk persons that were frequently risk allele carriers. The
240 distributions of ERS in each cancer type were shown in Figure S5. Most people did
241 not carry the causal alleles, while only a few were high-risk carriers. The density plot
242 of cases and controls indicated the cancer cases had obvious larger ERS than controls
243 (Figure S6).

244 Further, we evaluated the discrimination abilities of ERS and PRS using the C-index
245 for ten-year cancer incidence. The C-index values in the replication set were stable,
246 ranging from 0.601 to 0.686 [median (IQR): 0.655 (0.636-0.667)], outperforming the
247 traditional PRS [median (IQR): 0.585 (0.572-0.614)] (Figure 5c). Through Spearman
248 correlation analysis, we found low correlation between ERS and PRS (average $r_s =$
249 0.016), indicating the complementary predictive value for the rare exome variants
250 (Table S11).

251

252 **Methods**

253 *Study population and phenotype definition*

254 The UK Biobank (UKB) is a population-based prospective cohort of individuals aged
255 40–69 years, enrolled between 2006 and 2010. The work described herein was
256 approved by the UK Biobank under application no. 83445. All the phenotype data
257 were accessed in March 2022.

258 Health-related outcomes were ascertained via individual record linkage to national

259 cancer and mortality registries and hospital in-patient encounters. Cancer diagnoses
260 were coded by International Classification of Diseases version 10 (ICD-10) codes.
261 Individuals with at least one recorded incident diagnosis of a borderline, in situ, or
262 primary malignant cancer were defined as cases collected from data fields 41270
263 (Diagnoses - ICD10), 41202 (Diagnoses - main ICD10), 40006 (Type of cancer:
264 ICD10), and 40001 (primary cause of death: ICD10). Finally, we analyzed 20 primary
265 cancer types with overall cases > 500, including bladder, brain, breast, colorectal,
266 esophagus, head and neck, kidney, lung, leukemia, liver, non-Hodgkin's lymphoma
267 (NHL), ovary, pancreas, prostate, sarcoma, melanoma, skin (non-melanoma), stomach,
268 thyroid, uterus cancers.

269 To ensure the robustness of the results and validate the risk stratification model, we
270 used the WES-300k population released on Sep 28th 2021 as the discovery set and the
271 remaining WES-150k population additionally released on Oct 29th 2021 as the
272 replication set. We included only participants of European ancestry. The demographic
273 and cancer characteristics were described in [Table S1](#). All the data analyses were
274 performed on DNAnexus Research Analysis Platform (RAP).

275 *Quality control for the genetic variants*

276 Whole-exome sequencing data for UKB participants were generated using the IDT
277 xGen v1 capture kit on the NovaSeq6000 platform. The UK Biobank WES 450k
278 release includes CRAM and gVCF files processed using the OQFE protocol ¹⁹.
279 Single-sample variants were called from OQFE CRAMs with DeepVariant 0.0.10
280 employing a retrained model and are provided as single-sample gVCFs. All gVCFs
281 were aggregated with GLnexus 1.2.6 using the default joint-genotyping parameters
282 for DeepVariant. The OQFE protocol maps to a full GRCh38 reference version
283 including all alternative contigs in an alt-aware manner. Genotype depth filters (SNV
284 DP \geq 7, indel DP \geq 10) were applied prior to variant site filters requiring at least one

285 variant genotype passing an allele balance filter (heterozygous SNV $AB > 0.15$,
286 heterozygous indel < 0.20). The detailed parameters were described in Category 170 of
287 the UKB showcase. In addition, we filtered out the variants with low allele counts if a
288 variant had minor allele count (MAC) < 3 in the discovery set or $MAC < 2$ in the
289 replication set.

290 *Exome-wide association for single variants and gene levels*

291 Single-variant and gene-based association analyses were performed using SAIGE
292 v1.0.5^{20,21}. SAIGE is a toolkit developed for genome-wide association tests in
293 biobank level datasets, which uses saddlepoint approximation to handle extremely
294 case-control imbalance of binary traits and linear mixed models to account for sample
295 relatedness. The variant-level association tests included high-quality and reliable
296 variants with $MAC \geq 10$. The exact inclusion criteria for single variants were: (i) MAC
297 ≥ 3 in both cases and controls in the discovery set; (ii) $MAC \geq 2$ in both cases and
298 controls in the replication set. The variants were functionally annotated using Variant
299 Effect Predictor (VEP) software²². In addition to the coding variants that alter protein
300 sequences, synonymous variants could also disturb the level of mRNA expression and
301 have non-neutral functions^{23,24}. Thus, variants annotated as putative loss-of-function
302 (LoF, including nonsense, splice site, and frameshift variants), missense, and
303 synonymous were included in the analysis. Independent variants were pruned out
304 using the PLINK v1.9 clump function (*--clump-r2 0.50 --clump-kb 500*).

305 For gene-based analysis, we included rare and ultra-rare variants with minor allele
306 frequency (MAF) < 0.05 , 0.01 , or 0.001 . Three genetic models were considered: LoF,
307 LoF+missense, LoF+missense+synonymous. Of all the combinations, we reported the
308 association results with the lowest P value to collectively capture a wide range of
309 genetic architectures²⁵. The effect sizes and 95% confidence interval (CI) of genes
310 were estimated by burden tests.

311 In all the association analyses, we adjusted the covariates including age, gender
312 (excluding sex-specific tumors), Body Mass Index (BMI), smoking status (binary),
313 drinking status (binary), and the top 10 principal components (PCs). Meta-analysis
314 was used to summarize the results between discovery and replication sets for single
315 variants by METAL software ²⁶. At the same time, the gene-based P values were
316 aggregated by aggregated Cauchy association test (ACAT) method ²⁷.

317 We reported the significant associations when the variants/genes meet the following
318 criteria: (i) $P < 0.05$ in both discovery and replication sets; (ii) reach genome-wide
319 significant level ($P < 5 \times 10^{-8}$) in the meta-analysis for variant-level or pass Bonferroni
320 correction threshold $P < 2.5 \times 10^{-6}$ (nearly 20,000 protein-coding genes tested) for
321 gene-level.

322 ***Shared genetics analyses to identify potential genes across cancers***

323 To discover more potential variants associated with multiple cancers, we applied a
324 cross-trait meta-analysis using Association analysis based on SubSETs (ASSET) ²⁸.
325 ASSET is a statistical tool specifically designed to be powerful for pooling
326 association signals across multiple cancers when true effects may exist only in a
327 subset of the cancers. We considered both one-sided (one-directional) and two-sided
328 (bidirectional) ASSET. Variants with association $P < 1 \times 10^{-4}$ in any cancer type were
329 included. Variants were considered significant while reaching $P_{\text{overall}} < 5 \times 10^{-8}$ and $P <$
330 0.05 for each side in bidirectional tests.

331 ***Comparison analyses for gene expression in tumor and healthy normal tissues***

332 The UCSC Toil Recompute Compendium provides processed transcript-level
333 RNA-Seq data from TCGA tumor tissues and GTEx healthy normal tissues quantified
334 using a unified computational pipeline to remove computational batch effects. We

335 used this data to perform comparative analysis across tumor and health normal tissues
336 from both projects ²⁹. All the gene expression values were normalized to transcripts
337 per million (TPM) and then logarithmically transformed. After filtering, 15 tissue
338 types with sample size > 100, including 7,085 TCGA tumor tissues and 4,311 GTEx
339 normal tissues, were analyzed (Table S12). We used Student's t-test ($P_{\text{bonferroni}} < 0.05$)
340 and fold change ($FC > 1.5$ or $FC < 0.5$) to identify the differential gene expression
341 between tumor and healthy normal tissues.

342 ***Pathway enrichment analysis***

343 We collected the pathway information with gene sets from the KEGG database,
344 containing a total of 186 pathways up to March 2022. All enrichment analyses were
345 performed using the R package *clusterProfiler* ³⁰.

346 ***Comparison analyses for protein abundance in tumor and adjacent normal tissues***

347 The National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium
348 (CPTAC) is a national effort to accelerate the understanding of the molecular basis of
349 cancer through the application of large-scale proteome and genome analysis ³¹. Eleven
350 cancer types with available protein abundance data of 1,270 tumor tissues and 845
351 adjacent normal tissues were collected from ten tissue types (Table S13). The
352 differential proteins were identified following the same criteria with gene expression.

353 ***Protein-protein interaction analysis***

354 To further understand the protein-protein interactions, we used the STRING database,
355 which considered both physical interactions as well as functional associations ³². The
356 protein interaction network was clustered into different colors using Markov
357 Clustering (MCL).

358 ***Polygenic risk score generation***

359 The polygenic risk score (PRS) aggregates the effects of numerous genetic variants
360 into a single number which predicts genetic predisposition for a phenotype. To
361 investigate the association of PRS with cancer risk, we generated the PRS based on
362 previous literature reports. The SNP information and score weights were collected
363 from the PGS Catalog database³³. PGS Catalog is an open database of published
364 PRSs, covering >2,000 scores for various phenotypes. If multiple PRSs were reported
365 for one cancer, we selected the one generated from the largest sample size. After
366 filtering, 17 PRSs were collected and generated in UKB imputed genotype population
367 (data field 22828: Imputation from genotype), except for liver, stomach, and sarcoma
368 cancers that did not have reported PRSs (Table S14).

369 ***Development and validation of the risk stratification models based on rare variants***

370 We developed an exome-wide risk score (ERS) for population risk stratification based
371 on rare variants (MAF<0.05) for each cancer separately. The analyses were restricted
372 to incident cancers. To perform independent training and validation phases, the
373 selected variants and weights were determined using the association information in
374 the discovery set.

375 We select the rare genetic variants reaching $P < 1 \times 10^{-4}$ in the discovery set as the first
376 step. The variants identified in this analysis were further screened through penalized
377 regression using the least absolute shrinkage and selection operator (LASSO) after
378 10-fold cross-validation to reduce the overfitting and collinearity problem. When high
379 correlation exists, variants representing independent loci with the strongest statistical
380 significance were retained. The ERS was generated as: $ERS = \sum_1^n \beta_i G_i$, where β_i
381 denoted the coefficient of the i^{th} variant G_i calculated by SAIGE in the discovery

382 set.

383 In the validation phase, the variant panel with previously determined weights was
384 used to generate ERS in the replication set.

385 We used person-year to describe the absolute cancer incidence risk, which was
386 defined as the time gap from the date of cohort enrollment to cancer diagnosis or the
387 last follow-up, whichever came first. Hazard ratios (HRs) and 95% confidence
388 interval (CI) were used to evaluate the association between ERS and cancer risk based
389 on Cox proportional hazards models, adjusting for age, sex (excluding sex-specific
390 cancer), and top ten principal components. We compared effect sizes of ERS for
391 cancer risk based on the top 5% (extremely high-risk), 5~25% (high-risk), and bottom
392 75% (low-risk) percentile of ERS. The discrimination performance of the risk scores
393 were evaluated by Harrell's C-index.

394

395

396 **Discussion**

397 In this study, we comprehensively evaluated the susceptibility between genetic
398 variants on human exome and 20 primary cancer types in approximately 420,000
399 UKB participants of European ancestry. To our knowledge, this is the first
400 exome-wide pan-cancer study including almost the whole UKB population, which
401 could improve the statistical power compared with some previous studies using the
402 early-phase UKB 200k population^{34,35}. Trans-omics analyses were performed to
403 evaluate the functional evidence of identified signals, including genomics,
404 transcriptomics, and proteomics. Moreover, through establishing the independent
405 discovery and replication sets, the signals identified and the risk stratification models
406 could be validated externally to ensure their robustness, especially for the rare
407 variants with smaller MAF.

408 Our first major finding discovers exome-wide signals associated with multiple cancers.
409 In the ExWAS analyses, the identified protein-coding variants in *CHEK2* (known as
410 c.1100del), *BRCA1/2*, *ATM* have been reported associated with cancers in previous
411 studies ³⁶. Based on such a large-scale population, novel variants and genes were also
412 identified. For example, the missense variant rs6998061 in *POU5F1B*, which was
413 predicted as a possibly damaging SNV by VEP, was associated with prostate and
414 colorectal cancers. *POU5F1B* was a protein-coding gene highly homologous to *OCT4*
415 ³⁷, which was recently shown to be transcribed in cancer cells. It has been found
416 related to tumorigenicity and tumor growth *in vivo* and could promote angiogenesis
417 and cell proliferation and inhibits apoptosis in cancer cells ³⁸. The stop lost variant
418 rs387907272, predicted as probably damaging in *MYD88*, was associated with
419 leukemia and NHL. *MYD88* encodes a cytosolic adapter protein that plays a central
420 role in the innate and adaptive immune response ³⁹. It functions as an essential signal
421 transducer in the interleukin-1 and Toll-like receptor signaling pathways ⁴⁰. Moreover,
422 mutations in *MYD88* could activate NF- κ B and its associated signaling pathways,
423 thereby promoting B-cell proliferation and survival ⁴¹. Thus, this germline mutation
424 might be a promising target for clinical implications. Moreover, we identified multiple
425 SNVs located in 16q24.3 that were shared between skin cancer and melanoma,
426 including *VPS9DI*, *MC1R*, and *TUBB3*. *VPS9DI*, a protein-coding gene that affects
427 protein binding activity, was significantly associated with skin cancer and melanoma.
428 Although its role in cancer has not been reported, its antisense *VPS9DI-AS1* could
429 promote tumorigenesis and progression by mediating micro RNAs via the
430 Wnt/ β -catenin signaling pathway in multiple cancers ^{42,43}. Human *MC1R* has an
431 inefficient poly(A) site allowing intergenic splicing with its downstream neighbor
432 *TUBB3*, which were involved in melanogenesis. Melanogenesis is a key parameter of
433 differentiation in melanocytes and melanoma cells that could affect the treatment of
434 pigmentary disorders ⁴⁴. Therefore, the SNVs and genes we identified had remarkable
435 biological functions that were practical for precise clinical implications.

436 Our second major finding identified the pleiotropic variants shared among cancers.
437 We observed strong functional evidence for the identified genes from KEGG pathway
438 network and protein-protein interaction network through trans-omics analyses for
439 gene expression and protein abundance. We found several essential function modules
440 from proteomics. The DNA damage response and repair agents are widely used in
441 clinical oncology given the expanding role of immune checkpoint blockade as a
442 therapeutic strategy⁴⁵. The HLA locus, located on chromosome 6, is among the most
443 polymorphic regions of the human genome⁴⁶. HLA dysfunction is deeply involved in
444 the immune evasion events in the development and progression of certain cancers⁴⁷.
445 A previous study has reported that somatic mutation in HLA was associated with
446 multiple cancers⁴⁸; we hereby demonstrated that germline mutation in HLA was also
447 relevant. In addition, the keratin family and BPI fold containing family B (BPIFB)
448 were also related to cancers that had certain biological functions^{49,50}.

449 Our third major finding improves the ability for high-risk population identification. It
450 is widely recognized that early screening for cancers is most likely beneficial when
451 the target tumor type has relatively uniform biology and a slower rate of progression².
452 Targeting on high-risk populations with appropriate strategies for early detection
453 could get remarkable benefits of mortality reduction⁵¹⁻⁵³. However, the selection of
454 individual to be screened is a complex procedure, with difficulty accurately
455 identifying high-risk persons who are most likely to benefit from screening. Because
456 cancer is heritable, PRS is emerging as the quantitative measurement for individual
457 genetic risk. However, the heritability for common variants identified in GWAS is
458 limited, while the contribution of rare variants could not be ignored. Therefore, we
459 leveraged the rare variants to construct the ERS to offset this limitation. By evaluating
460 C-index and risk stratification, we demonstrated the added values of rare variants.

461 ERS could be combined with specific tumor screening strategies, suggesting that
462 people with extremely high risk should be screened frequently (e.g., once a year), and

463 those with high risk should be screened regularly (e.g., once every three years), which
464 is expected to further reduce the cancer mortality. Therefore, ERS is expected to serve
465 as an informative benchmark to incorporate the PRS and baseline information that
466 have been used in cancer risk assessment.

467 Our work has several strengths. First, we comprehensively evaluated the exome-wide
468 genetic variants in 20 cancer types on variant and gene levels among 420k
469 participants and analyzed the cross-cancer pleiotropy through cross-trait meta-analysis.
470 Second, we explored the relationship between identified genes and cancers at
471 multi-omics levels, including genomics, transcriptomics, and proteomics. The
472 trans-omics analyses revealed that the identified signals were functional. Third, we
473 focused on the high-risk population identification based on exome-wide variants,
474 while few studies developed risk scores using rare variants. We demonstrated the
475 stable performance of ERS across pan-cancer in the replication set, especially for its
476 ability to identify the extremely high-risk persons. Therefore, the ERS might serve as
477 a complementary genetic risk assessment tool combined with the existing screening
478 guidelines.

479 It is essential to acknowledge the limitations of our study. First, this study was
480 conducted in the UKB population only. Although we established a discovery set and
481 replication set, it is not strictly independent validation. Therefore, future large-scale
482 population studies should be conducted to replicate these findings. Second, we
483 focused on individuals of European ancestry only. Moreover, it is essential to evaluate
484 the associations of variants and performance of ERS in non-European populations.
485 Third, we mainly investigated the genetic effects of population risk stratification.
486 However, the contribution of environmental factors should not be ignored.
487 Well-established risk models incorporated with environmental factors, PRS, and ERS
488 should be developed for specific cancer.

489 In conclusion, our study provides novel insights into human exomes and rare variants

490 through comprehensive analyses of genetic susceptibility to human cancers and
491 subsequent target analyses on specific genes and risk stratification.

492

493 **Data Availability**

494 UK Biobank data is available from <https://www.ukbiobank.ac.uk/>. TCGA data is
495 available from <https://portal.gdc.cancer.gov/>. GTEx data is available from
496 <https://www.gtexportal.org/home/>. CPTAC data is available from
497 <https://pdc.esacinc.com/pdc/pdc>.

498 **Code Availability**

499 The R software codes that support our findings are available from the corresponding
500 author by a reasonable request.

501

502

503 **Reference**

- 504 1 Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of
505 Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a*
506 *cancer journal for clinicians* **71**, 209-249, doi:10.3322/caac.21660 (2021).
- 507 2 Shieh, Y. *et al.* Population-based screening for cancer: hope and hype. *Nat Rev*
508 *Clin Oncol* **13**, 550-565, doi:10.1038/nrclinonc.2016.50 (2016).
- 509 3 Li, N. *et al.* One-off low-dose CT for lung cancer screening in China: a
510 multicentre, population-based, prospective cohort study. *The Lancet.*
511 *Respiratory medicine* **10**, 378-391, doi:10.1016/S2213-2600(21)00560-9
512 (2022).
- 513 4 Kovalchik, S. A. *et al.* Targeting of low-dose CT screening according to the
514 risk of lung-cancer death. *N Engl J Med* **369**, 245-254,
515 doi:10.1056/NEJMoa1301851 (2013).
- 516 5 Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of
517 cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N*
518 *Engl J Med* **343**, 78-85, doi:10.1056/NEJM200007133430201 (2000).
- 519 6 Rashkin, S. R. *et al.* Pan-cancer study detects genetic risk variants and shared
520 genetic basis in two large cohorts. *Nat Commun* **11**, 4423,
521 doi:10.1038/s41467-020-18246-6 (2020).
- 522 7 Visscher, P. M., Yengo, L., Cox, N. J. & Wray, N. R. Discovery and
523 implications of polygenicity of common diseases. *Science* **373**, 1468-1473,
524 doi:10.1126/science.abi8206 (2021).
- 525 8 Tam, V. *et al.* Benefits and limitations of genome-wide association studies.
526 *Nature reviews. Genetics* **20**, 467-484, doi:10.1038/s41576-019-0127-1
527 (2019).
- 528 9 Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P. R. Whole-exome
529 imputation within UK Biobank powers rare coding variant association and
530 fine-mapping analyses. *Nat Genet* **53**, 1260-1269,
531 doi:10.1038/s41588-021-00892-1 (2021).
- 532 10 Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960
533 individuals in the UK Biobank. *Nature* **586**, 749-756,
534 doi:10.1038/s41586-020-2853-0 (2020).
- 535 11 Cirulli, E. T. *et al.* Genome-wide rare variant analysis for thousands of
536 phenotypes in over 70,000 exomes from two cohorts. *Nat Commun* **11**, 542,
537 doi:10.1038/s41467-020-14288-y (2020).

-
- 538 12 Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to
539 clinical instruments. *Genome medicine* **12**, 44,
540 doi:10.1186/s13073-020-00742-5 (2020).
- 541 13 Polygenic Risk Score Task Force of the International Common Disease, A.
542 Responsible use of polygenic risk scores in the clinic: potential benefits, risks
543 and gaps. *Nat Med* **27**, 1876-1884, doi:10.1038/s41591-021-01549-6 (2021).
- 544 14 Wainschtein, P. *et al.* Assessing the contribution of rare variants to complex
545 trait heritability from whole-genome sequence data. *Nat Genet* **54**, 263-273,
546 doi:10.1038/s41588-021-00997-7 (2022).
- 547 15 Singh, T. *et al.* Rare coding variants in ten genes confer substantial risk for
548 schizophrenia. *Nature* **604**, 509-516, doi:10.1038/s41586-022-04556-w
549 (2022).
- 550 16 Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank
551 participants. *Nature* **599**, 628-634, doi:10.1038/s41586-021-04103-z (2021).
- 552 17 Sun, B. B. *et al.* Genetic associations of protein-coding variants in human
553 disease. *Nature* **603**, 95-102, doi:10.1038/s41586-022-04394-w (2022).
- 554 18 Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published
555 genome-wide association studies, targeted arrays and summary statistics 2019.
556 *Nucleic Acids Res* **47**, D1005-D1012, doi:10.1093/nar/gky1120 (2019).
- 557 19 Szustakowski, J. D. *et al.* Advancing human genetics research and drug
558 discovery through exome sequencing of the UK Biobank. *Nat Genet* **53**,
559 942-948, doi:10.1038/s41588-021-00885-0 (2021).
- 560 20 Zhou, W. *et al.* Scalable generalized linear mixed model for region-based
561 association tests in large biobanks and cohorts. *Nat Genet* **52**, 634-639,
562 doi:10.1038/s41588-020-0621-6 (2020).
- 563 21 Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample
564 relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335-1341,
565 doi:10.1038/s41588-018-0184-y (2018).
- 566 22 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome biology* **17**,
567 122, doi:10.1186/s13059-016-0974-4 (2016).
- 568 23 Sauna, Z. E. & Kimchi-Sarfaty, C. Understanding the contribution of
569 synonymous mutations to human disease. *Nature reviews. Genetics* **12**,
570 683-691, doi:10.1038/nrg3051 (2011).
- 571 24 Shen, X., Song, S., Li, C. & Zhang, J. Synonymous mutations in
572 representative yeast genes are mostly strongly non-neutral. *Nature*,
573 doi:10.1038/s41586-022-04823-w (2022).

-
- 574 25 Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK
575 Biobank exomes. *Nature* **597**, 527-532, doi:10.1038/s41586-021-03855-y
576 (2021).
- 577 26 Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis
578 of genomewide association scans. *Bioinformatics* **26**, 2190-2191,
579 doi:10.1093/bioinformatics/btq340 (2010).
- 580 27 Liu, Y. *et al.* ACAT: A Fast and Powerful p Value Combination Method for
581 Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet* **104**, 410-421,
582 doi:10.1016/j.ajhg.2019.01.002 (2019).
- 583 28 Bhattacharjee, S. *et al.* A subset-based approach improves power and
584 interpretation for the combined analysis of genetic association studies of
585 heterogeneous traits. *Am J Hum Genet* **90**, 821-835,
586 doi:10.1016/j.ajhg.2012.03.015 (2012).
- 587 29 Vivian, J. *et al.* Toil enables reproducible, open source, big biomedical data
588 analyses. *Nat Biotechnol* **35**, 314-316, doi:10.1038/nbt.3772 (2017).
- 589 30 Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for
590 comparing biological themes among gene clusters. *OmicS: a journal of*
591 *integrative biology* **16**, 284-287 (2012).
- 592 31 Satpathy, S. *et al.* A proteogenomic portrait of lung squamous cell carcinoma.
593 *Cell* **184**, 4348-4371 e4340, doi:10.1016/j.cell.2021.07.016 (2021).
- 594 32 Szklarczyk, D. *et al.* The STRING database in 2021: customizable
595 protein-protein networks, and functional characterization of user-uploaded
596 gene/measurement sets. *Nucleic Acids Res* **49**, D605-D612,
597 doi:10.1093/nar/gkaa1074 (2021).
- 598 33 Lambert, S. A. *et al.* The Polygenic Score Catalog as an open database for
599 reproducibility and systematic evaluation. *Nat Genet* **53**, 420-425,
600 doi:10.1038/s41588-021-00783-5 (2021).
- 601 34 Zeng, C. *et al.* Association of Pathogenic Variants in Hereditary Cancer Genes
602 With Multiple Diseases. *JAMA Oncol*, doi:10.1001/jamaoncol.2022.0373
603 (2022).
- 604 35 Cheng, S. *et al.* Exome-wide screening identifies novel rare risk variants for
605 major depression disorder. *Mol Psychiatry*, doi:10.1038/s41380-022-01536-4
606 (2022).
- 607 36 Reiner, A. S. *et al.* Radiation Treatment, ATM, BRCA1/2, and
608 CHEK2*1100delC Pathogenic Variants and Risk of Contralateral Breast
609 Cancer. *J Natl Cancer Inst* **112**, 1275-1279, doi:10.1093/jnci/djaa031 (2020).

-
- 610 37 Breyer, J. P. *et al.* An expressed retrogene of the master embryonic stem cell
611 gene POU5F1 is associated with prostate cancer susceptibility. *Am J Hum*
612 *Genet* **94**, 395-404, doi:10.1016/j.ajhg.2014.01.019 (2014).
- 613 38 Hayashi, H. *et al.* The OCT4 pseudogene POU5F1B is amplified and
614 promotes an aggressive phenotype in gastric cancer. *Oncogene* **34**, 199-208,
615 doi:10.1038/onc.2013.547 (2015).
- 616 39 Cohen, P. & Strickson, S. The role of hybrid ubiquitin chains in the MyD88
617 and other innate immune signalling pathways. *Cell Death Differ* **24**,
618 1153-1159, doi:10.1038/cdd.2017.17 (2017).
- 619 40 Salcedo, R., Cataisson, C., Hasan, U., Yuspa, S. H. & Trinchieri, G. MyD88
620 and its divergent toll in carcinogenesis. *Trends Immunol* **34**, 379-389,
621 doi:10.1016/j.it.2013.03.008 (2013).
- 622 41 de Groen, R. A. L., Schrader, A. M. R., Kersten, M. J., Pals, S. T. & Vermaat, J.
623 S. P. MYD88 in the driver's seat of B-cell lymphomagenesis: from molecular
624 mechanisms to clinical implications. *Haematologica* **104**, 2337-2348,
625 doi:10.3324/haematol.2019.227272 (2019).
- 626 42 Wang, X. *et al.* ZEB1 activated-VPS9D1-AS1 promotes the tumorigenesis and
627 progression of prostate cancer by sponging miR-4739 to upregulate MEF2D.
628 *Biomed Pharmacother* **122**, 109557, doi:10.1016/j.biopha.2019.109557
629 (2020).
- 630 43 Ettinger, D. S. *et al.* NCCN Guidelines Insights: Non-Small Cell Lung Cancer,
631 Version 2.2021. *J Natl Compr Canc Netw* **19**, 254-266,
632 doi:10.6004/jnccn.2021.0013 (2021).
- 633 44 Kleszczynski, K. *et al.* Melatonin exerts oncostatic capacity and decreases
634 melanogenesis in human MNT-1 melanoma cells. *J Pineal Res* **67**, e12610,
635 doi:10.1111/jpi.12610 (2019).
- 636 45 Mouw, K. W., Goldberg, M. S., Konstantinopoulos, P. A. & D'Andrea, A. D.
637 DNA Damage and Repair Biomarkers of Immunotherapy Response. *Cancer*
638 *Discov* **7**, 675-693, doi:10.1158/2159-8290.CD-17-0226 (2017).
- 639 46 Cabrera, T., Lopez-Nevot, M. A., Gaforio, J. J., Ruiz-Cabello, F. & Garrido, F.
640 Analysis of HLA expression in human tumor tissues. *Cancer Immunol*
641 *Immunother* **52**, 1-9, doi:10.1007/s00262-002-0332-0 (2003).
- 642 47 Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes
643 across 21 tumour types. *Nature* **505**, 495-501, doi:10.1038/nature12912
644 (2014).
- 645 48 Shukla, S. A. *et al.* Comprehensive analysis of cancer-associated somatic

646 mutations in class I HLA genes. *Nat Biotechnol* **33**, 1152-1158,
647 doi:10.1038/nbt.3344 (2015).

648 49 Karantza, V. Keratins in health and cancer: more than mere epithelial cell
649 markers. *Oncogene* **30**, 127-138, doi:10.1038/onc.2010.456 (2011).

650 50 Li, J. *et al.* Molecular biology of BPIFB1 and its advances in disease. *Ann*
651 *Transl Med* **8**, 651, doi:10.21037/atm-20-3462 (2020).

652 51 Independent, U. K. P. o. B. C. S. The benefits and harms of breast cancer
653 screening: an independent review. *Lancet* **380**, 1778-1786,
654 doi:10.1016/S0140-6736(12)61611-0 (2012).

655 52 Jonas, D. E. *et al.* Screening for Lung Cancer With Low-Dose Computed
656 Tomography: Updated Evidence Report and Systematic Review for the US
657 Preventive Services Task Force. *JAMA* **325**, 971-987,
658 doi:10.1001/jama.2021.0377 (2021).

659 53 Buskermolen, M. *et al.* Colorectal cancer screening with faecal
660 immunochemical testing, sigmoidoscopy or colonoscopy: a microsimulation
661 modelling study. *BMJ* **367**, 15383, doi:10.1136/bmj.15383 (2019).

662

663

664 **Figure legends**

665 **Figure 1.** (a) Manhattan plot for the single variant association results in ExWAS. The
666 blue dash line indicates the genome-wide significance level ($P < 5 \times 10^{-8}$). (b)
667 Manhattan plot for the gene-based association results in ExWAS. The blue dash line
668 indicates the Bonferroni correction level ($P < 2.5 \times 10^{-6}$). (c) The heatmap of
669 variant-level and gene-level association results for the genes shared at least two
670 cancer types. We report the signal if it reaches nominal $P < 0.05$ in the corresponding
671 cancer type (red: odds ratio (OR) > 1; blue: OR < 1). The grey color indicates
672 association $P > 0.05$.

673 **Figure 2.** (a) Circos plot for the genes identified in the cross-trait meta-analysis. (b)
674 Heatmap for the number of shared genetic variants across each cancer pair. The red
675 color indicates the shared variants with one-directional effects. The blue color
676 indicates the shared variants with bidirectional effects between the cancer pairs.

677 **Figure 3.** (a) Heatmap of fold change (FC) values to compare gene expression
678 between TCGA tumor tissues and GTEx healthy normal tissues. (b) Volcano plot for
679 the FC values and $-\log(P)$ values for comparison of gene expression. (c) Heatmap of
680 FC values to compare protein abundance between CPTAC tumor tissues and adjacent
681 normal tissues. NA: not available. (d) Volcano plot for the FC values and $-\log(P)$
682 values for comparison of protein abundance. (e) KEGG pathway network from the
683 enrichment analysis of the pleiotropic genes. (f) Protein-protein interaction network of
684 the signal genes and pleiotropic genes.

685 **Figure 4.** Cumulative cancer incidence plot for the 20 cancer types in the replication
686 set (WES-150k). The red line indicates the extremely high-risk persons, the green line
687 indicates the high-risk population, and the blue line indicates the low-risk persons.
688 The P values were calculated using the log-rank tests.

689 **Figure 5.** (a) The dot plot of the hazard ratios (HRs) and 95% confidence intervals
690 (CIs) of exome-wide risk scores (ERS). The low-risk subgroup (blue dot) was set as
691 the reference group. The red dot indicates the extremely high-risk persons, the green
692 dot indicates the high-risk population. (b) The dot plot of the HRs and 95% CIs of
693 polygenic risk scores (PRS) for 17 cancer types, while liver, stomach, and sarcoma
694 cancers did not have reported PRSs. (c) The C-index values of ERS and PRS
695 generated by the Cox regression model.

696 **Supplementary Figures**

697 Figure S1. Q-Q plots of single variant tests for each cancer

698 Figure S2. Q-Q plot combining all the P values of pan-cancer

699 Figure S3. Manhattan plot for the single variant analyses of each cancer

700 Figure S4. Cumulative cancer incidence plot for the 20 cancer types in the whole
701 UKB-450k population.

702 Figure S5. Histogram of the exome-wide risk scores (ERS) in each cancer

703 Figure S6. Density plot of the exome-wide risk scores (ERS) in cases and controls

704 **Supplementary Tables**

705 Table S1. Demographic characteristics in the UK Biobank cohort

706 Table S2. Association results for independent single variants with $P < 5 \times 10^{-8}$ in the
707 whole UKB-450k population

708 Table S3. Association results for single variants with $P < 5 \times 10^{-8}$ in at least two cancer
709 types

710 Table S4. Association results for genes with $P < 2.5 \times 10^{-6}$ in the whole UKB-450k
711 population

712 Table S5. Association results for genes with $P < 2.5 \times 10^{-6}$ in at least two cancer types

713 Table S6. Cross-trait meta-analysis results for single variants with one-directional
714 effects

715 Table S7. Cross-trait meta-analysis results for single variants with bidirectional effects

716 Table S8. Number of independent shared genetic variants in the cross-trait
717 meta-analysis

718 Table S9. KEGG enrichment analysis for the genes identified in cross-trait
719 meta-analysis

720 Table S10. Model parameters for the exome-wide risk scores (ERS) in 20 cancer types

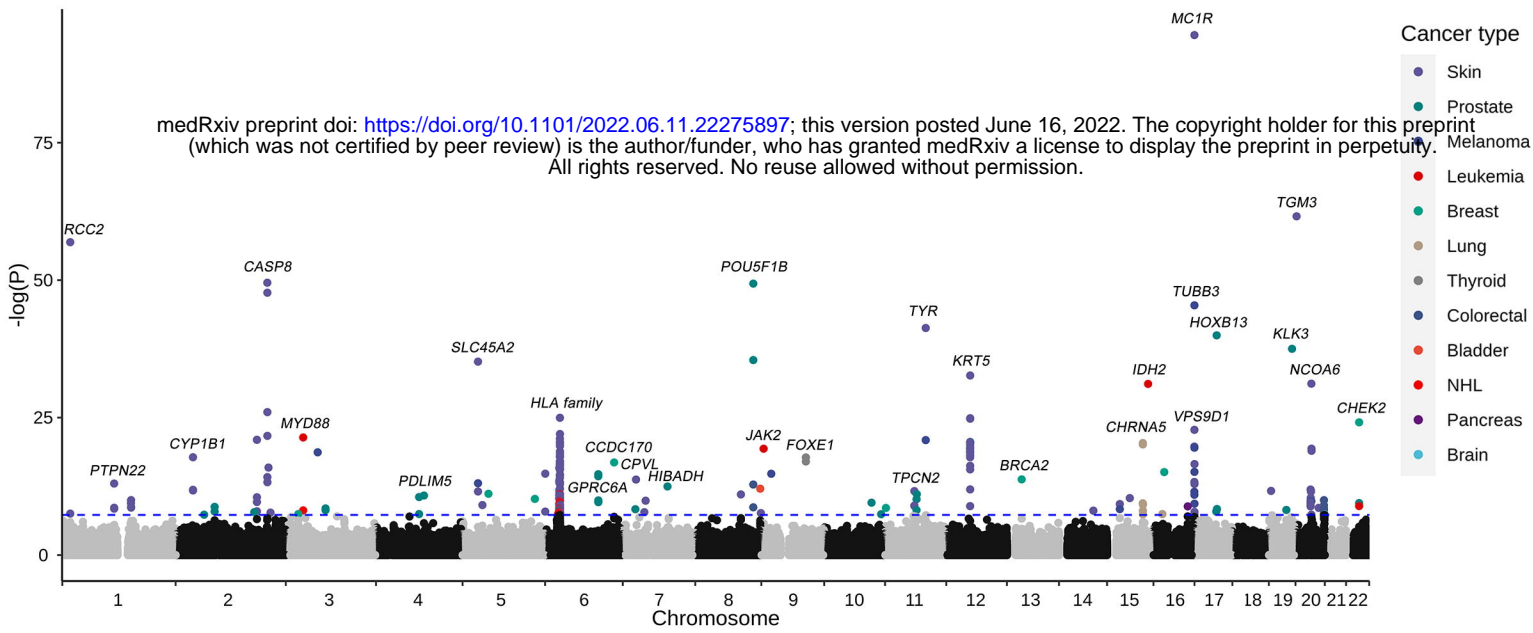
721 Table S11. Correlation analysis for ERS and PRS

722 Table S12. Number of tissues included in the comparison analyses for gene
723 expression in tumor and healthy normal tissues

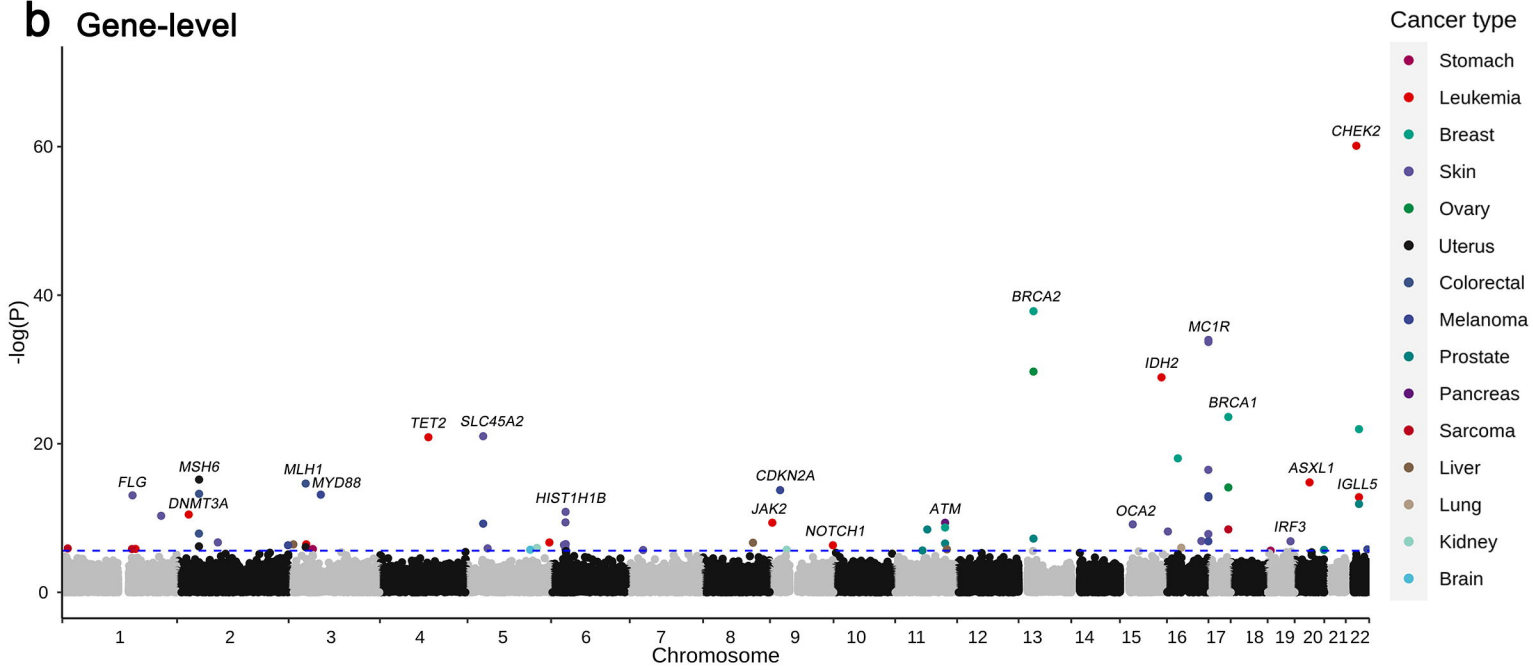
724 Table S13. Number of tissues included in the comparison analyses for protein
725 abundance in CPTAC

726 Table S14. Information of the selected polygenic risk score (PRS) in PGS catalog
727

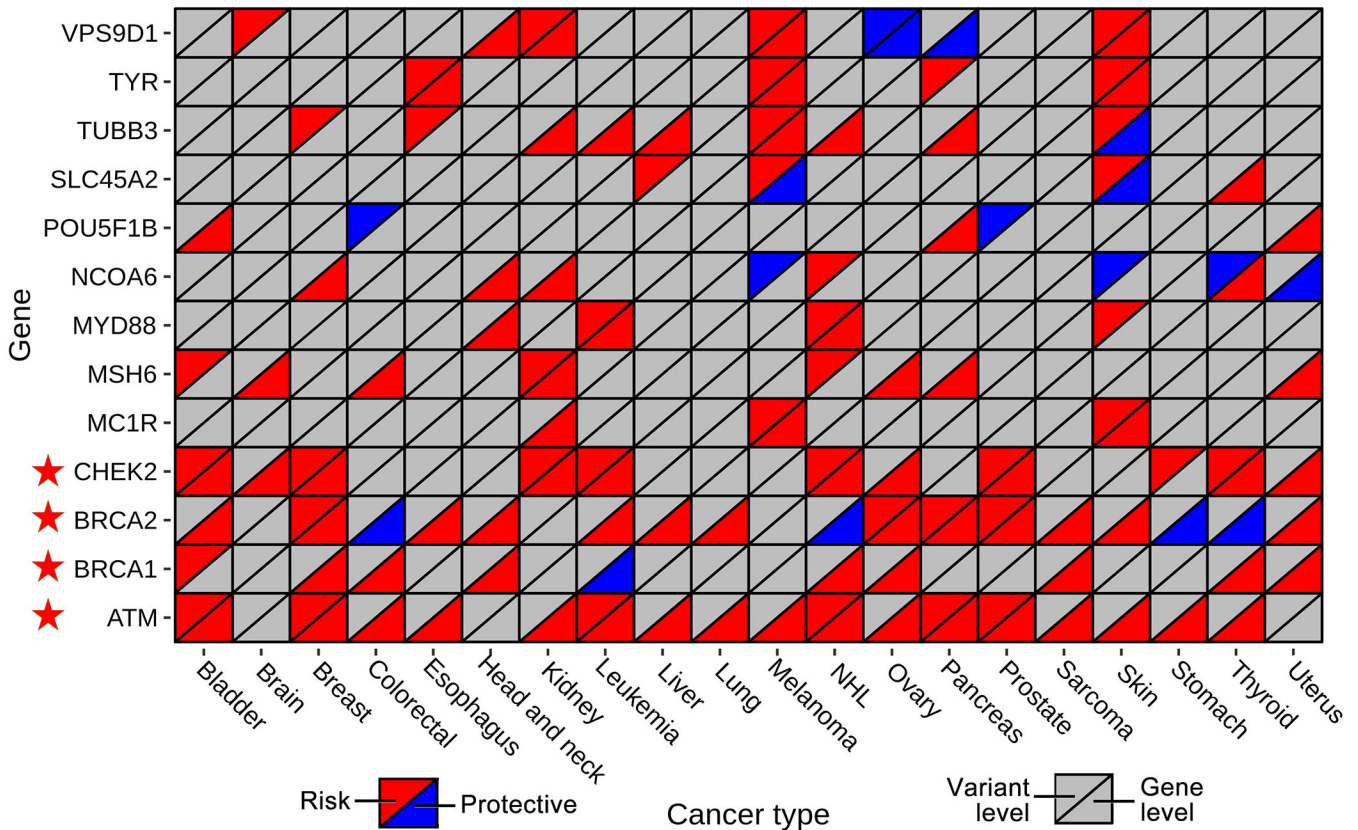
a Variant-level

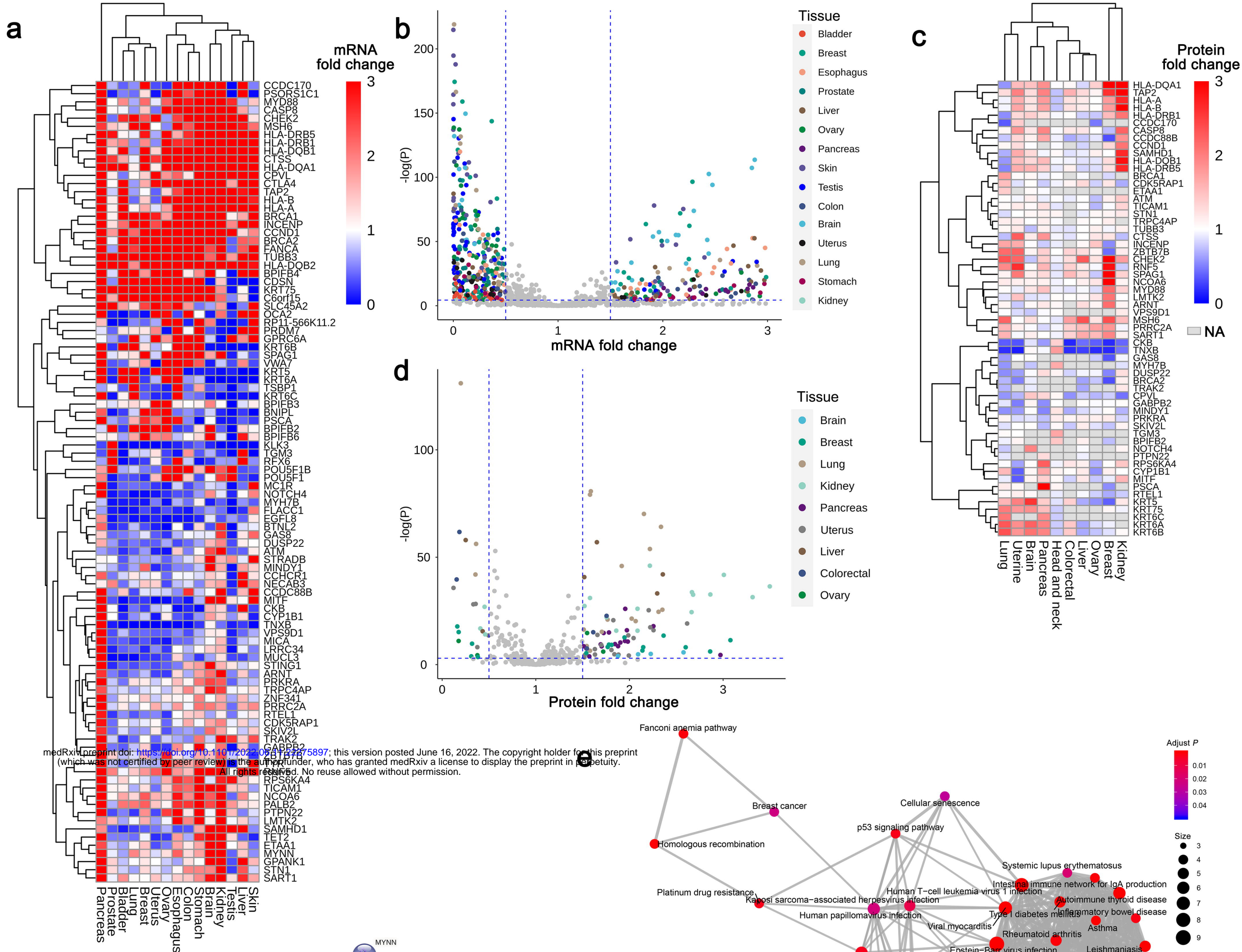


b Gene-level



c





medRxiv preprint doi: <https://doi.org/10.1101/2022.06.16.22275897>; this version posted June 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

- DNA damage response and repair
- HLA and immune
- Keratin family (type II cyto keratins)
- BPI fold containing family B
- Melanogenesis

