

Genetic analysis of functional rare germline variants across 9 cancer types from the DiscovEHR study

Manu Shivakumar^{1,2}, Jason E. Miller², Venkata Ramesh Dasari³, David Carey⁴, Radhika Gogoi^{3*},
Dokyoon Kim^{1,5,6*} on behalf of the DiscovEHR collaboration

¹ Biomedical & Translational Informatics Institute, Geisinger, Danville, PA, USA

² Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA

³ Weis Center for Research, Geisinger Clinic, Danville, PA, USA

⁴ Department of Molecular and Functional Genomics, Geisinger, Danville, PA, USA

⁵ Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA

⁶ Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, USA

Abstract

Rare variants play an essential role in the etiology of cancer and characterizing rare germline variants that impact the risk of cancer is an ongoing challenge. We performed a genome-wide rare variant analysis using germline whole exome sequencing (WES) data derived from the Geisinger MyCode initiative to discover cancer predisposition variants. The case-control association analysis was conducted by binning pathogenic and likely pathogenic variants in 5,538 cancer patients and 7,286 matched controls in a discovery set and 1,991 cancer patients and 2,504 matched controls in a validation set across nine cancer types. We discovered 87 genes and 106 pathways significantly associated with cancer (*Bonferroni-corrected* $P < 0.05$) out of which seven genes and 26 pathways replicated from the validation set (*suggestive threshold* $P < 0.05$). Further, four genes and 21 pathways were discovered to be associated with multiple cancers (*Bonferroni-corrected* $P < 0.05$). Additionally, we identified 13 genes and two pathways associated with survival outcome across seven cancers (*Bonferroni-corrected* $P < 0.05$), where

31 two genes, *PCDHB8* and *DCHS2*, were also associated with survival outcome in TCGA data. In
32 summary, we conducted one of the largest pan-cancer association studies using germline data
33 derived from a single hospital system to find novel predisposition genes and pathways
34 associated with nine cancers. Our results can inform future guidelines for germline genetic
35 testing in cancer, which will be helpful in screening for cancer high-risk patients. This work adds
36 to the knowledge base and progress being made in precision medicine.

37

38

39 **Introduction**

40 Cancer is the second most lethal disease in United States with an estimated 1,735,350 new
41 cases and 609,640 deaths in 2018¹. Cancer is caused by inherited germline variants and
42 acquired somatic mutations. A recent twin study showed ~33% heritability of cancer across 23
43 cancer types with a high estimate of 57% for prostate (MIM: 176807), 31% for breast (MIM:
44 114480), 38% for kidney (MIM: 144700), and 58% for skin melanoma (MIM: 155600)². Germline
45 genetic markers for cancers have been widely studied leading to the discovery of many
46 heritable predisposition genes such as *BRCA1* (MIM: 113705), *BRCA2* (MIM: 600185) and *PALB2*
47 (MIM: 610355) in breast cancer, *RB1* (MIM: 614041) in retinoblastoma and *MLH1* (MIM:
48 120436), *MSH2* (MIM: 609309), *MSH6* (MIM: 600678), and *PMS2* (MIM: 600259) in Lynch
49 syndrome (MIM: 120435). To date, many genome-wide association studies (GWAS) have been
50 conducted and many more variants and genes have been discovered as associated with various
51 cancer types³⁻⁸. However, a large portion of inherited genetic factors that result in
52 carcinogenesis is still unknown and many studies are being undertaken to discover these

53 genetic variants. For instance, the genetic contribution explained by all variants discovered to
54 date is about 39% in prostate cancer^{2; 9} and 30% in breast cancer^{2; 10}.
55
56 Since common variants discovered to be associated with multiple cancers have only modest
57 effect size, the missing heritability could be further explained by rare variants. Moreover, rare
58 variants have been known to contribute to various complex diseases including cancer¹¹⁻¹³. The
59 aggregation of rare variants in a gene can lead to loss of function of the gene or change in
60 expression¹⁴. Similarly, since pathways perform a sequence of biochemical actions leading to a
61 cellular function or product, changes in the expression of genes involved within a pathway can
62 lead to cancer^{15; 16}. Previous studies have also indicated that cancer is caused by an
63 accumulation of a number of singular or rare variants in particular genes or pathways¹². To that
64 effect, binning the pathogenic and likely pathogenic rare variants into genes and pathways
65 would help us increase statistical power to detect associations and also infer biological
66 mechanisms^{13; 17; 18}.
67
68 The MyCode community initiative is a precision medicine project, launched at Geisinger in
69 2007, which enabled the storage of blood, serum, and DNA samples in a system-wide
70 biorepository that is available for use in broad research¹⁹. To date, over 244,000 patients have
71 signed up for the MyCode initiative and over 90,000 patient blood samples have been
72 sequenced as part of the DiscovEHR project in collaboration with the Regeneron Genetics
73 Center²⁰. The sequenced data can be easily linked to the electronic health record (EHR) of the
74 patient, allowing access to rich longitudinal data. Apart from the EHR, Geisinger also maintains

75 a cancer registry that contains all the patients diagnosed or treated for cancer at any Geisinger
76 medical facility. Further, as part of the MyCode program, the genetic data is also being used to
77 detect increased risk of developing one or more of 21 medically actionable conditions, including
78 breast cancer, ovarian cancer (MIM: 167000), Marfan syndrome (MIM: 154700), Lynch
79 syndrome, etc., and the results are returned to the patients through a “Return of Results”
80 program²¹. Moreover, many similar programs around the world are helping to integrate
81 genomics into clinical practice²². Manolio et al. emphasize that the integration of genomic
82 findings to clinical practice has been relatively slow and insist on the need to have an openly
83 accessible knowledge base of variants, phenotypes and clinically actionable variants²³. The
84 sharing of genetic findings is likely to help the scientific research community to improve our
85 understanding of the phenotype of interest and propel precision medicine by bringing more
86 genomics into clinical practice.

87

88 In summary, we conducted one of the largest pan-cancer association studies using germline
89 data from a single hospital system to find novel genes and pathways associated with nine
90 cancers. Our study also validates several genes and pathways that have already been implicated
91 in other genome-wide association studies. We identified 87 genes that were significant across
92 cancers, of which seven were replicated in an independent dataset and four genes were shared
93 among multiple cancers. We also identified 106 pathways that reached genome-wide
94 significance, of which 26 pathways were replicated. Further, 21 pathways were significant
95 across multiple cancers. In addition to the genes and pathways associated with cancer risk, we
96 also identified 13 genes and two pathways associated with survival outcome across cancers.

97

98 **Results**

99

100 **Study design and population characteristics**

101

102 This study was based on a subset of 7,449 cancer cases and 9,792 controls selected by matching
103 age, BMI and gender from ~90,000 sequenced samples from the DiscovEHR study. The samples
104 were sequenced in two phases using different platforms as described in the Methods section. In
105 phase 1, 60,000 samples were sequenced, and 5,538 cancer patients across nine cancers and
106 7,286 matched controls were pulled. In phase 2, 30,000 samples were sequenced which
107 included 1,991 cancer patients and 2,504 matched controls. Consequently, the phase 2 dataset
108 was used to replicate results from phase 1. Cancer patient IDs that were retrieved from the
109 cancer registry were classified into particular cancers using International Classification of
110 Diseases for Oncology (ICD-O) codes. After classifying the cancer patients to their respective
111 cancers, only nine cancers, including bladder (MIM: 109800), breast, colorectal (MIM: 114500),
112 kidney, lung (MIM: 211980), melanoma, prostate, thyroid (MIM: 188550), and uterine cancer,
113 had more than 300 samples in the discovery set. A low number of samples in association
114 studies results in a higher type 1 error rate and lower statistical power to detect associations²⁴.
115 Thus, the rest of the cancers were excluded from this study. The distribution and basic
116 demographics of patients across these cancers are shown in Table 1. A common set of controls
117 were used for all the cancers except breast, uterine and prostate cancer as they are sex-specific

118 cancers. The sex-matched controls were separately pulled to match the same number of
119 controls across all cancers.

120

121 Table 1 provides age, BMI, female ratio, and vital status (alive or deceased) information across
122 cancers. Among all cancers, breast cancer had the largest number of cases (N = 1,214) followed
123 by prostate cancer (N = 1,146). Further, uterine cancer had a significantly higher average BMI as
124 compared to other cancers. The average BMI for uterine cancer patients was 38.33 kg/m² in the
125 discovery dataset and 36.76 kg/m² in the replication dataset. Additionally, lung cancer had the
126 highest number of cases who are deceased, which is expected as lung cancer is by far the
127 leading cause of death due to cancer¹. Further, observing the female ratio across the cancers
128 also shows a gender disparity in some cancers. Specifically, the incidence rate in bladder cancer
129 was found to be 4.46 fold higher in male than female, and in thyroid cancer it was 3.95 fold
130 higher in female than male. The difference in incidence rates have been well documented in
131 other studies with a 3-4 times increased risk of bladder cancer risk in men²⁵ and 2.9 times
132 increased risk of thyroid cancer in women²⁶.

133

134 Variant filtering based on functional annotation and scores improves power and has been
135 successfully used in many association studies^{14; 27}. In this study, the variants from whole exome
136 sequence data were annotated using Variant Effect Predictor (VEP)²⁸ and ClinVar²⁹.

137 Subsequently, only the variants categorized as pathogenic and likely-pathogenic based on the
138 annotations were retained for further analysis. A strategy for classifying variants as pathogenic
139 and likely pathogenic is elaborated in the Methods section. Additionally, all common variants

140 were removed, and only rare variants (MAF < 0.05) were retained. The number of variants
141 available after filtering out for each cancer cohort is listed in Table S1.

142

143 **Pathway-based rare variant analysis**

144

145 Association analysis using rare variants usually suffers from a lack of power as very large
146 datasets are required. Therefore, rare variants are often binned into a biologically informed-
147 unit such as a gene or pathway to improve the power¹⁷. In this study, pathogenic and likely
148 pathogenic rare variants with minor allele frequency (MAF) < 0.05 were binned into genes
149 followed by KEGG pathways using BioBin^{17; 30; 31}. Next, an association test was performed to
150 determine if the gene/pathway is significantly associated with the phenotype. SKAT-O is an
151 optimal unified approach that combines a burden and non-burden sequence kernel association
152 test (SKAT) test, and maintains power regardless of the direction of effect and causality of the
153 variants³². After determining the association p-values, they were adjusted for multiple testing in
154 each cancer type separately using a Bonferroni correction. Any genes/pathways with a
155 Bonferroni-corrected $P < 0.05$ were considered as significant results. The same procedure was
156 followed to conduct rare variant analysis in discovery and replication datasets. Figure 2 shows
157 all the pathways that were Bonferroni significant across all cancers and the same are listed in
158 Table S4-S6.

159

160 In total, 106 pathways were found to be significantly associated across all cancers (Figure 2).

161 However, no significant pathways were found in prostate cancer. Further, 26 pathways: 12 in

162 bladder cancer, five in colorectal cancer, five in kidney cancer, two in lung cancer and two in
163 thyroid cancer, marked in red in Figure 2, were replicated from the replication dataset (SKAT-O
164 $P < 0.05$). More information regarding the 26 pathways including locus count, minor allele
165 count (MAC) in cases, MAC in controls, SKAT-O p-value and Bonferroni-corrected p-value in
166 discovery and replication datasets can be found in Table 2. Additionally, 21 pathways were
167 found to be significantly associated in more than one cancer, with the FoxO signaling pathway
168 and GnRH signaling pathway significantly associated with four cancers, followed by apoptosis
169 and bladder cancer significantly associated with three cancers and the rest of the 17 pathways
170 significantly associated with two cancers (Table 3).

171

172 **Gene-based rare variant analysis**

173

174 All pathogenic and likely pathogenic variants below $MAF < 0.05$ were binned into gene
175 boundaries defined by Entrez annotations derived from Library of Knowledge Integration (LOKI)
176 using BioBin^{17; 30}. The total number of genes that the variants were binned across all cancers is
177 listed in Table S2. The bar plot in Figure 3 shows the total number of loci binned for a given
178 gene and variant types as annotated by VEP. In total, there were 87 genes that were
179 significantly associated with a specific cancer (Bonferroni-corrected $P < 0.05$) (Figure 3).

180 Furthermore, seven genes - *MLNR* (MIM: 602885), *CPAMD8* (MIM: 608841), *CHRNE* (MIM:
181 100725), *HOXB13* (MIM: 604607), *SCML4* (HGNC: 256380), *BST1* (MIM: 600387) and *TMEM186*
182 (HGNC: 25880), marked in red in Figure 3, were replicated in the phase 2 dataset ($P < 0.05$)
183 (Table 4).

184

185 Additionally, four genes, including *MAPK12* (MIM: 602399), *ECE2* (MIM: 610145), *DNMT3A*
186 (MIM: 602769) and *CHIA* (MIM: 606080), were significantly associated in multiple cancers. The
187 gene *MAPK12* was significantly associated with bladder cancer and colorectal cancer, *ECE2* and
188 *CHIA* for melanoma and colorectal cancer, and *DNMT3A* for bladder and lung cancer. Further
189 association test statistics on these genes are described in Table 5. The PhenoGram³³ plot in
190 Figure 4 shows all the genes found to be significantly associated across all cancers. Additionally,
191 the lollipop³⁴ plots in Figure 5 and Figure S1 shows the type of variants – frameshift, missense,
192 stop gained, stop lost, splice acceptor, splice donor, start lost, and their relative position in the
193 gene. Variants that were found in the Catalog of Somatic Mutation in Cancer (COSMIC)
194 database were marked with COSMIC ids. The nonsense stop gained variants marked in yellow
195 usually results in a truncated protein, which are non-functional and the frameshift variants
196 marked in red usually cause a loss of function due to a shift in the reading frame.

197

198 **Pathogenic and likely pathogenic variants in known oncogenes and tumor suppressor genes**

199

200 Previously, Huang et al.¹⁴ identified potential genes in cancers with a higher enrichment of
201 pathogenic or likely pathogenic variants identified in the Exome Aggregation Consortium (ExAC)
202 non-TCGA cohort from a curated list of genes that contribute to cancer susceptibility. They
203 identified 28 cancer gene associations (FDR < 0.05) and 16 suggestive associations (FDR < 0.15)
204 by conducting total frequency test (TFT)¹⁴ on germline data across 33 cancers. We wanted to
205 see if any of these genes also had higher mutational germline burden in our study as it would

206 help validate cancer susceptibility genes and even discover new gene associations that were
207 not statistically significant in Huang et al. As such, the list of gene-cancer pairs were filtered to
208 the nine cancer types under consideration in this study and TFT was run on the genes to
209 validate the enrichment of pathogenic and likely pathogenic variants in cancer patients against
210 control group. Seven genes – *ATM* (lung, MIM: 607585), *CHECK2* (breast, HGNC: 11200), *MSH2*
211 (colorectal), *BRCA2* (thyroid), *POLE* (kidney, MIM: 174762), *PALB2* (breast), *MLH1* (colorectal)
212 were found to be significant at a TFT p-value < 0.05. Moreover, two of the genes – *ATM* (lung)
213 and *BRCA2* (thyroid) found to be previously significant¹⁴ were replicated in this study. Further,
214 three genes – *ATM* (lung), *CHECK2* (breast) and *MSH2* (colorectal) were significant at FDR <
215 0.15. The carrier frequency of these genes and distribution of pathogenic and likely pathogenic
216 variants across oncogenes and tumor suppressor genes is shown in Figure 6 (Table S8). In
217 summary, we were able to validate two genes (TFT p-value < 0.05) and we discovered two more
218 genes with a suggestive association (FDR < 0.15).

219

220 **Survival analysis**

221

222 In this study, we also sought to discover variants that have an impact on the survival of
223 patients. To this aim, a weighed burden matrix was used to run cox regression adjusting for age
224 and BMI as covariates. The cox p-values were further adjusted using a Bonferroni correction
225 separately on each cancer type to account for multiple testing. Thirteen genes and two
226 pathways across seven cancers were found to be significantly associated with survival at
227 Bonferroni < 0.05. Further, to confirm that the p-values were not a random effect, permutation

228 testing was conducted by randomly shuffling the weighed burden values among patients and
229 running cox regression 100,000 times. We observed considerably less significant permutation p-
230 values across all genes and pathways. The permutation p-values and other statistics are listed in
231 Table 6 for significant genes and Table 7 for significant pathways. All the Kaplan-Meier survival
232 curves are shown in Figure S2. The genes that were significantly associated were further tested
233 for association with survival in The Cancer Genome Atlas (TCGA) provisional data on the cBio
234 Cancer Genomics Portal (<http://cbioportal.org>)³⁵. The two groups were formed using somatic
235 mutations and mRNA expression (RNA Seq V2 RSEM) using a z-score threshold ± 2 . Two genes
236 *PCDHB8* (*Logrank P* = 9.22E-03, MIM: 606334) and *DCHS2* (*Logrank P* = 0.036, MIM: 612486)
237 were significant at a *Logrank P* < 0.05.

238

239 Discussion

240

241 In this study, we present results from a rare variant analysis conducted across nine cancers
242 using a cohort of 7,449 cancer cases and 9,792 controls from a single hospital system using
243 whole exome sequencing data and clinical data from a patient EHR. A total of 133 pathways (26
244 replicated) and 91 genes (7 replicated) were identified as associated with cancers. Furthermore,
245 21 pathways and four genes were associated with multiple cancer types. Additionally, we
246 identified 13 genes and two pathways as associated with survival across multiple cancers.

247

248 Many KEGG pathways identified in this study have already been implicated in cancer, such as
249 “pathways in cancer”, “GnRH signaling pathway”³⁶, “bladder cancer”, “FoxO signaling pathway”

250 ³⁷, “metabolic pathways”, “gap junction” ³⁸, “apoptosis”, “base excision repair”, “melanoma”,
251 “choline metabolism in cancer” and “basal cell carcinoma”. One pathway of interest that is
252 associated with bladder cancer is the “insulin secretion pathway”. Previous studies have shown
253 diabetes mellitus increases the risk of bladder cancer³⁹ and is possibly due to administration of
254 the anti-diabetic drug Pioglitazone⁴⁰. Another pathway “HTLV-I infection” was found to be
255 associated with kidney cancer and was also replicated. HTLV-I is a known oncovirus that causes
256 cancer⁴¹. Further studies on the “HTLV-I infection” pathway could elucidate the role of germline
257 variants in cancers. Another pathway “Legionellosis” was found to be associated with kidney
258 and bladder cancer, Legionella pneumonia in cancer has a very high mortality rate ~31%⁴², and
259 variants in pathway could play a role in susceptibility or recovery of the patients. Another
260 pathway, the “Hippo signaling pathway” was found to be associated with uterine cancer. The
261 “Hippo tumor suppressor pathway” is known to phosphorylate YAP and TAZ which are critical
262 for cell growth, reprogramming and development⁴³. The Hippo pathway also interacts with the
263 PI3K/AKT pathway which is commonly involved in cancer⁴³. Additionally, Hippo pathway is
264 known to affect the survival of cancer patients⁴⁴ and in this study, one of the genes *DCHS2*,
265 which is part of Hippo pathway, was also associated with survival in uterine cancer and it was
266 also replicated in the TCGA data.
267
268 A number of previous studies have shown *HOXB13* to be associated with prostate cancer⁴⁵⁻⁴⁸,
269 and in this study as well, *HOXB13* was found to be associated with prostate cancer in the
270 discovery dataset and was replicated. Another gene, *CPAMD8*, which is involved in broad-
271 spectrum protease inhibition, innate immunity and damage control was found to be associated

272 with kidney cancer in the discovery and replication datasets⁴⁹. *CPAMD8* is known to be
273 substantially expressed in kidney^{49; 50} and given its functional role, rare gene-disruptive variants
274 in *CPAMD8* could lead to carcinogenesis. We also identified two genes associated with uterine
275 cancer that replicated - *CHRNE* which is a subunit of nicotinic acetylcholine receptors (nAChRs)
276 and *TMEM186* which is a member of the transmembrane protein family. Nicotine, a compound
277 present in cigarettes, mediates cell proliferation and angiogenesis through nicotinic
278 acetylcholine receptors (nAChRs) and its subunits⁵¹. Still, its mechanism of action is not well
279 understood for uterine cancer where some studies have shown smoking to reduce the risk of
280 uterine cancer contrary to other cancers^{51; 52}. Again, the exact role of *TMEM186* in uterine
281 cancer is also unexplored. TMEMs are differentially regulated in many types of cancers and
282 some TMEMs are known to act as tumor suppressors while others as oncogenes⁵³. Further, in
283 bladder cancer, the Putative Polycomb group (PcG) protein gene (*SCML4*), is involved in the
284 regulation of crucial developmental and physiological processes and is known to promote
285 proliferation and inhibit apoptosis. *SCML4* was replicated in bladder cancer⁵⁴.

286

287 Different cancer types share some pathways and genes, which generally include common
288 tumor suppressor genes and oncogenes¹⁴. In this study, we identified 21 pathways and four
289 genes that were associated with multiple cancers. Two of the genes *MAPK12* and *DNMT3A* are
290 well known genes involved in cancer with *MAPK12* acting as *p38 MAPK*, which is involved in cell
291 differentiation, apoptosis and autophagy, whereas *DNMT3A* is involved in DNA methylation and
292 its disruption leads to tumorigenesis^{55; 56}. Gene *ECE2* cleaves endothelin-1 (ET-1) which is a
293 potent vasoconstrictor peptide and ET-1 is known to be involved in angiogenesis, apoptosis and

294 growth in colorectal cancer and melanoma⁵⁷. Elevated levels of plasma levels and increased
295 immunopositivity of ET-1 has been observed in colorectal cancer⁵⁷.

296

297 We also discovered 13 genes that are associated with survival in cancers. The presence of rare
298 pathogenic and likely pathogenic variants reduced the overall survival rate for all the genes
299 identified across cancers. Some of the genes discovered were already known to be associated
300 with survival in cancers and a subset of them have also been suggested as a target for cancer
301 therapy like *SAXO2* (HGNC: 283726) which is involved in microtubule binding in uterine
302 cancer⁵⁸, *PCDHB8* whose downregulation is known to result in poor prognosis in bladder
303 cancer⁵⁹, *ATXN3* (MIM: 607047) whose downregulation increases expression of tumor
304 suppressor *PTEN* (MIM: 601728)⁶⁰, and *TPTE2* (MIM: 606791), a homolog of *PTEN*, whose
305 upregulation suppresses metastasis and/or tumorigenesis⁶¹. Other genes were associated with
306 survival in this study - *ANO5* (MIM: 608662) is known to regulate cell migration and invasion⁶²,
307 the *HOGA1* (MIM: 613597) gene is involved in metabolism, *CSH2* (MIM: 118820) is involved in
308 postnatal and intrauterine growth⁶³ and *HLA-G* (MIM: 142871) offers an immune escape
309 mechanism as it is involved in cytokine signaling in the immune system and class I MHC
310 mediated antigen processing and presentation⁶⁴. Thus, disruption of the normal activity of
311 these genes could promote cancer. Another gene *NAA38* (MIM: 617990) which was associated
312 with survival in thyroid has been shown to be associated with survival in glioblastoma (MIM:
313 137800)⁶⁵. Furthermore, two genes *PCDHB8* and *DCHS2* were significantly associated with
314 survival in the TCGA provisional dataset.

315

316 Even though many associations were identified in this study, further studies would be required
317 to elucidate the molecular mechanisms. A shortcoming of this study was that the replication
318 cohort was underpowered to replicate all the findings. Additionally, the participants in the
319 replication cohort were derived from participants who enrolled into the MyCode program more
320 recently than the discovery dataset. Therefore, most of the patients in the replication set were
321 alive and it was not practical to run survival analysis using the replication dataset. The
322 limitations imposed by the sample size and power would be addressed in the future as the
323 MyCode and DiscovEHR programs are still ongoing and more samples are being sequenced.
324 Another limitation of our study is that our population predominantly consists of European
325 ancestry, mainly due to the patient population at Geisinger which is predominantly of European
326 ancestry.

327

328 In conclusion, this study conducted genome-wide rare-variant analysis to find novel genes and
329 pathways associated across nine cancers. We also replicated many genes and pathways that are
330 very well known in cancers, which further emphasizes the fact that some portion of the missing
331 heritability is attributed to rare variants. We also identified some genes associated with the
332 survival of the patients which have already been suggested as targets in cancer therapy. We
333 also discovered novel genes and pathways associated with survival of patients which could be
334 potential targets for cancer therapy. The genes and pathways discovered in this study can be
335 used to screen for high-risk cancer patients and personalized therapy. In summary, results from
336 this study could help define a portion of the missing heritability associated with cancer and
337 have broad applications in precision medicine.

338

339 **Methods**

340

341 **Study population**

342 The study population consisted of Geisinger patients who consented to participate in the
343 MyCode community initiative. As part of MyCode initiative, individuals agreed to provide blood
344 and DNA samples for broad, future research, including genomic analyses as part of the
345 Regeneron-Geisinger DiscovEHR collaboration and linking to data in the Geisinger EHR under a
346 protocol approved by the Geisinger Institutional Review Board. The cases were a subset of
347 cancer patients from a cancer registry that were part of 90,000 patients sequenced as part of
348 the DiscovEHR project. The cases were classified into different cancers using ICD-O site codes as
349 defined in Table S3. Cases that were recorded as having cancer in multiple primary sites were
350 removed. Further, only cancers with at least 300 cases in 60,000 patients were sequenced in
351 phase 1 of the DiscovEHR study were included and other cancers were discarded due to a low
352 number of cases. A common control set was selected for all non-sex specific cancers using
353 matched age and BMI to cases from a pool of patients who did not have any ICD9/ICD10 code
354 related to cancer in a problem-list entry of the diagnosis code, an inpatient hospitalization-
355 discharge diagnosis code, or an encounter diagnosis code. The controls for breast, uterine and
356 prostate cancer were pulled separately to have the same number of controls as the common
357 control set. Age was calculated as age at diagnosis for cases and current age if alive or age at
358 death for controls. The median of BMI values recorded in the EHR from a year before diagnosis
359 date was used as BMI for cases. Furthermore, the median of BMI values from a year before

360 current date or date of death for alive and deceased patients respectively was used as BMI for
361 controls.

362

363 **Sequencing and quality control**

364 All the study population was sequenced as part of the DiscovEHR project at the Regeneron
365 Genetic Center. Initially, around 60,000 samples were sequenced using NimbleGen probe
366 target-capture (SeqCap VCRome) and further a separate batch of 30,000 samples were
367 sequenced using xGen capture (Integrated DNA Technologies) followed by sequencing on the
368 Illumina HiSeq 2500. The variant calling was done using GATK^{66; 67}. Further detailed description
369 of sequencing is available at Shivakumar et al.¹³ and Mirshahi et al.⁶⁸. Additional call rate quality
370 controls were applied. The markers and samples with a call rate below 90% were filtered out.
371 All related patients showing up to 3rd degree relatedness corresponding to IBD > 0.125 were
372 removed.

373

374 **Variant annotation and filtering**

375

376 All variants were annotated using VEP and ClinVar. All the loci that satisfied the following two
377 conditions were retained and the rest were filtered out.

- 378 1. Loci that were annotated with impact 'HIGH' using VEP.
- 379 2. Loci that were annotated as pathogenic and likely pathogenic with at least 1 star using
380 ClinVar.

381 Variants that satisfied the conditions were considered pathogenic and likely pathogenic and all
382 analysis were run using only these loci.

383

384 **Gene based rare variant association**

385

386 All the variants that were annotated as pathogenic and likely pathogenic were binned using
387 BioBin^{17; 30}. BioBin uses pre-compiled knowledge in a LOKI database, which is compiled using
388 information from various data sources including Entrez and KEGG. The variants were binned
389 into genes using Entrez annotations. Only variants with MAF < 0.05 were considered rare and
390 the rest of the variants were filtered out. Additionally, bins with less than 20 variants (MAC)
391 were filtered out due to low sample size. Further, the binned variants were weighed using
392 Madsen-browning weights⁶⁹. The statistical association tests were run using SKAT-O
393 implemented as R package³². Additionally, the association tests were adjusted using age, BMI
394 and first four principle components as covariates. The principle components were calculated
395 using EIGENSOFT⁷⁰, using common variants after LD pruning with indep-pairwise 50 5 0.5 and
396 Hardy-Weinberg equilibrium of 10^{-6} . The association test p-values were further adjusted using
397 Bonferroni correction to account for multiple testing correction.

398

399 **Pathway based rare variant association**

400 BioBin was used to bin rare variants with a MAF < 0.05 into KEGG pathways derived from
401 LOKI^{17; 30}. Any pathway bin that did not contain a total of at least 20 variants across case and
402 cancer were filtered out due to small sample size. Further all the bins were weighted using

403 Madsen-browning weighting⁶⁹. Statistical association was run using SKAT-O implemented as R
404 package³². The association tests were adjusted using age, BMI and four principle components
405 as covariates. Further, the association p-values were adjusted using Bonferroni correction.

406

407 **Survival analysis**

408

409 Survival analysis was run using cox regression adjusting for age and BMI. Specifically, the BioBin
410 bin-phe output files which contains a weighted burden of variants were used to get the
411 weighted burden of each patient for the bin (gene/pathway) and cox regression was run on the
412 bin adjusting for age and BMI. Survival analysis was performed on each cancer using gene-
413 based bins and pathway-based bins. Further, survival p-values were adjusted for multiple
414 testing using Bonferroni correction.

415

416 **Author contributions**

417

418 MS, JEM, VRD, RG and DK designed and conceived the project. MS and JEM carried out the
419 methodology and implementation. DK and RG helped supervise the project. MS wrote the
420 paper in consultation with JEM, VRD, RG and DK.

421

422 **Conflicts of Interest**

423

424 The authors declare that they have no competing interests.

425

426 **Acknowledgement**

427

428 This work was supported by NLM R01 NL012535. This project was also funded, in part, under a
429 grant with the Pennsylvania Department of Health (#SAP 4100070267). The Department
430 specifically disclaims responsibility for any analyses, interpretations or conclusions. We
431 gratefully acknowledge the funding support from Geisinger Medical Center (SRC-075) (RG) and
432 Rice Women’s Cancer Research Fund (RG and VRD). Support for this work was also provided by
433 NHGRI T32HG009495-01 (JEM). The funders specifically disclaim responsibility for the study
434 design, data collection, analyses, interpretation, conclusions, and writing of the manuscript.

435

436 **Web resources**

437

438 BioBin, <https://ritchielab.org/software/biobin-download>

439 VEP, <https://ensembl.org/info/docs/tools/vep/index.html>

440 ClinVar, <https://www.ncbi.nlm.nih.gov/clinvar/>

441 EIGENSOFT, <https://www.hsph.harvard.edu/alkes-price/software/>

442 OMIM, <http://www.omim.org/>

443 R statistical software, <https://www.r-project.org/>

444 NCBI, <https://www.ncbi.nlm.nih.gov/gene/4049>

445 DiscovEHR, <http://www.discovehrshare.com/>

446

447 References

448

- 449 1. Siegel, R.L., Miller, K.D., and Jemal, A. (2018). Cancer statistics, 2018. CA: A Cancer Journal for
450 Clinicians 68, 7-30.
- 451 2. Mucci, L.A., Hjelmborg, J.B., Harris, J.R., Czene, K., Havelick, D.J., Scheike, T., Graff, R.E., Holst,
452 K., Möller, S., Unger, R.H., et al. (2016). Familial Risk and Heritability of Cancer Among
453 Twins in Nordic Countries. JAMA 315, 68-76.
- 454 3. Ghousaini, M., Fletcher, O., Michailidou, K., Turnbull, C., Schmidt, M.K., Dicks, E., Dennis, J.,
455 Wang, Q., Humphreys, M.K., Luccarini, C., et al. (2012). Genome-wide association
456 analysis identifies three new breast cancer susceptibility loci. Nature genetics 44, 312-
457 318.
- 458 4. Haiman, C.A., Chen, G.K., Vachon, C.M., Canzian, F., Dunning, A., Millikan, R.C., Wang, X.,
459 Ademuyiwa, F., Ahmed, S., Ambrosone, C.B., et al. (2011). A common variant at the
460 TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer.
461 Nature genetics 43, 1210-1214.
- 462 5. Kote-Jarai, Z., Olama, A.A.A., Giles, G.G., Severi, G., Schleutker, J., Weischer, M., Campa, D.,
463 Riboli, E., Key, T., Gronberg, H., et al. (2011). Seven prostate cancer susceptibility loci
464 identified by a multi-stage genome-wide association study. Nature genetics 43, 785-791.
- 465 6. Peters, U., Hutter, C.M., Hsu, L., Schumacher, F.R., Conti, D.V., Carlson, C.S., Edlund, C.K.,
466 Haile, R.W., Gallinger, S., Zanke, B.W., et al. (2012). Meta-analysis of new genome-wide
467 association studies of colorectal cancer risk. Human genetics 131, 217-234.
- 468 7. Al Olama, A.A., Kote-Jarai, Z., Berndt, S.I., Conti, D.V., Schumacher, F., Han, Y., Benlloch, S.,
469 Hazelett, D.J., Wang, Z., Saunders, E., et al. (2014). A meta-analysis of 87,040 individuals
470 identifies 23 new susceptibility loci for prostate cancer. Nature genetics 46, 1103-1109.
- 471 8. Artomov, M., Stratigos, A.J., Kim, I., Kumar, R., Lauss, M., Reddy, B.Y., Miao, B., Daniela
472 Robles-Espinoza, C., Sankar, A., Njauw, C.-N., et al. (2017). Rare Variant, Gene-Based
473 Association Study of Hereditary Melanoma Using Whole-Exome Sequencing. JNCI:
474 Journal of the National Cancer Institute 109, djx083-djx083.
- 475 9. Amin Al Olama, A., Dadaev, T., Hazelett, D.J., Li, Q., Leongamornlert, D., Saunders, E.J.,
476 Stephens, S., Cieza-Borrella, C., Whitmore, I., Benlloch Garcia, S., et al. (2015). Multiple
477 novel prostate cancer susceptibility signals identified by fine-mapping of known risk loci
478 among Europeans. Human molecular genetics 24, 5589-5602.
- 479 10. Complexo, Southey, M.C., Park, D.J., Nguyen-Dumont, T., Campbell, I., Thompson, E.,
480 Trainer, A.H., Chenevix-Trench, G., Simard, J., Dumont, M., et al. (2013). COMPLEXO:
481 identifying the missing heritability of breast cancer via next generation collaboration.
482 Breast cancer research : BCR 15, 402-402.
- 483 11. Rahman, N. (2014). Realizing the promise of cancer predisposition genes. Nature 505, 302.
- 484 12. Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. (2009). Common vs. rare allele
485 hypotheses for complex diseases. Current opinion in genetics & development 19, 212-
486 219.

- 487 13. Shivakumar, M., Miller, J.E., Dasari, V.R., Gogoi, R., and Kim, D. (2019). Exome-Wide Rare
488 Variant Analysis From the DiscovEHR Study Identifies Novel Candidate Predisposition
489 Genes for Endometrial Cancer. *Frontiers in Oncology* 9, 574.
- 490 14. Huang, K.-l., Mashl, R.J., Wu, Y., Ritter, D.l., Wang, J., Oh, C., Paczkowska, M., Reynolds, S.,
491 Wyczalkowski, M.A., Oak, N., et al. (2018). Pathogenic Germline Variants in 10,389 Adult
492 Cancers. *Cell* 173, 355-370.e314.
- 493 15. Li, D., Dong, X., Duell, E.J., Arslan, A.A., Zeleniuch-Jacquotte, A., Bueno-de-Mesquita, H.B.,
494 Gallinger, S., Gross, M., Holly, E.A., Bracci, P.M., et al. (2012). Pathway analysis of
495 genome-wide association study data highlights pancreatic development genes as
496 susceptibility factors for pancreatic cancer. *Carcinogenesis* 33, 1384-1390.
- 497 16. Cantor, R.M., Lange, K., and Sinsheimer, J.S. (2010). Prioritizing GWAS results: A review of
498 statistical methods and recommendations for their application. *American Journal of*
499 *Human Genetics* 86, 6-22.
- 500 17. Moore, C.B., Wallace, J.R., Frase, A.T., Pendergrass, S.A., and Ritchie, M.D. (2013). BioBin: a
501 bioinformatics tool for automating the binning of rare variants using publicly available
502 biological knowledge. *BMC medical genomics* 6 Suppl 2, S6-S6.
- 503 18. Kim, D., Li, R., Dudek, S.M., Wallace, J.R., and Ritchie, M.D. (2015). Binning somatic
504 mutations based on biological knowledge for predicting survival: an application in renal
505 cell carcinoma. *Pacific Symposium on Biocomputing Pacific Symposium on*
506 *Biocomputing*, 96-107.
- 507 19. Carey, D.J., Fetterolf, S.N., Davis, F.D., Faucett, W.A., Kirchner, H.L., Mirshahi, U., Murray,
508 M.F., Smelser, D.T., Gerhard, G.S., and Ledbetter, D.H. (2016). The Geisinger MyCode
509 community health initiative: an electronic health record-linked biobank for precision
510 medicine research. *Genetics In Medicine* 18, 906.
- 511 20. Dewey, F.E., Murray, M.F., Overton, J.D., Habegger, L., Leader, J.B., Fetterolf, S.N.,
512 O'Dushlaine, C., Van Hout, C.V., Staples, J., Gonzaga-Jauregui, C., et al. (2016).
513 Distribution and clinical impact of functional variants in 50,726 whole-exome sequences
514 from the DiscovEHR study. *Science* 354, aaf6814.
- 515 21. Schwartz, M.L.B., McCormick, C.Z., Lazzeri, A.L., Lindbuchler, D.A.M., Hallquist, M.L.G.,
516 Manickam, K., Buchanan, A.H., Rahm, A.K., Giovanni, M.A., Frisbie, L., et al. (2018). A
517 Model for Genome-First Care: Returning Secondary Genomic Findings to Participants
518 and Their Healthcare Providers in a Large Research Cohort. *The American Journal of*
519 *Human Genetics* 103, 328-337.
- 520 22. Stark, Z., Dolman, L., Manolio, T.A., Ozenberger, B., Hill, S.L., Caulfield, M.J., Levy, Y., Glazer,
521 D., Wilson, J., Lawler, M., et al. (2019). Integrating Genomics into Healthcare: A Global
522 Responsibility. *The American Journal of Human Genetics* 104, 13-20.
- 523 23. Manolio, T.A., Chisholm, R.L., Ozenberger, B., Roden, D.M., Williams, M.S., Wilson, R., Bick,
524 D., Bottinger, E.P., Brilliant, M.H., Eng, C., et al. (2013). Implementing genomic medicine
525 in the clinic: the future is here. *Genetics In Medicine* 15, 258.
- 526 24. Zhang, X., Basile, A.O., Pendergrass, S.A., and Ritchie, M.D. (2019). Real world scenarios in
527 rare variant association analysis: the impact of imbalance and sample size on the power
528 in silico. *BMC Bioinformatics* 20, 46.

- 529 25. Dobruch, J., Daneshmand, S., Fisch, M., Lotan, Y., Noon, A.P., Resnick, M.J., Shariat, S.F.,
530 Zlotta, A.R., and Boorjian, S.A. (2016). Gender and Bladder Cancer: A Collaborative
531 Review of Etiology, Biology, and Outcomes. *European Urology* 69, 300-310.
- 532 26. Rahbari, R., Zhang, L., and Kebebew, E. (2010). Thyroid cancer gender disparity. *Future*
533 *oncology* (London, England) 6, 1771-1779.
- 534 27. Esteban-Jurado, C., Vila-Casadesús, M., Garre, P., Lozano, J.J., Pristoupilova, A., Beltran, S.,
535 Muñoz, J., Ocaña, T., Balaguer, F., López-Cerón, M., et al. (2014). Whole-exome
536 sequencing identifies rare pathogenic variants in new predisposition genes for familial
537 colorectal cancer. *Genetics In Medicine* 17, 131.
- 538 28. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and
539 Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology* 17, 122.
- 540 29. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J.,
541 Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of
542 clinically relevant variants. *Nucleic Acids Research* 44, D862-D868.
- 543 30. Basile, A.O., Byrska-Bishop, M., Wallace, J., Frase, A.T., and Ritchie, M.D. (2018). Novel
544 features and enhancements in BioBin, a tool for the biologically inspired binning and
545 association analysis of rare variants. *Bioinformatics* (Oxford, England) 34, 527-529.
- 546 31. Moore, C.C.B., Basile, A.O., Wallace, J.R., Frase, A.T., and Ritchie, M.D. (2016). A biologically
547 informed method for detecting rare variant associations. *BioData mining* 9, 27-27.
- 548 32. Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Team,
549 N.G.E.S.P.E.L.P., Christiani, D.C., Wurfel, M.M., and Lin, X. (2012). Optimal unified
550 approach for rare-variant association testing with application to small-sample case-
551 control whole-exome sequencing studies. *American Journal of Human Genetics* 91, 224-
552 237.
- 553 33. Wolfe, D., Dudek, S., Ritchie, M.D., and Pendergrass, S.A. (2013). Visualizing genomic
554 information across chromosomes with PhenoGram. *BioData mining* 6, 18-18.
- 555 34. Skidmore, Z.L., Wagner, A.H., Lesurf, R., Campbell, K.M., Kunisaki, J., Griffith, O.L., and
556 Griffith, M. (2016). GenVisR: Genomic Visualizations in R. *Bioinformatics* (Oxford,
557 England) 32, 3012-3014.
- 558 35. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A.,
559 Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and
560 clinical profiles using the cBioPortal. *Science signaling* 6, pl1-pl1.
- 561 36. Gründker, C., and Emons, G. (2017). The Role of Gonadotropin-Releasing Hormone in
562 Cancer Cell Proliferation and Metastasis. *Frontiers in endocrinology* 8, 187-187.
- 563 37. Farhan, M., Wang, H., Gaur, U., Little, P.J., Xu, J., and Zheng, W. (2017). FOXO Signaling
564 Pathways as Therapeutic Targets in Cancer. *International journal of biological sciences*
565 13, 815-827.
- 566 38. Aasen, T., Mesnil, M., Naus, C.C., Lampe, P.D., and Laird, D.W. (2016). Gap junctions and
567 cancer: communicating for 50 years. *Nature reviews Cancer* 16, 775-788.
- 568 39. Fang, H., Yao, B., Yan, Y., Xu, H., Liu, Y., Tang, H., Zhou, J., Cao, L., Wang, W., Zhang, J., et al.
569 (2013). Diabetes mellitus increases the risk of bladder cancer: an updated meta-analysis
570 of observational studies. *Diabetes technology & therapeutics* 15, 914-922.
- 571 40. Tuccori, M., Filion, K.B., Yin, H., Yu, O.H., Platt, R.W., and Azoulay, L. (2016). Pioglitazone use
572 and risk of bladder cancer: population based cohort study. *BMJ* 352, i1541.

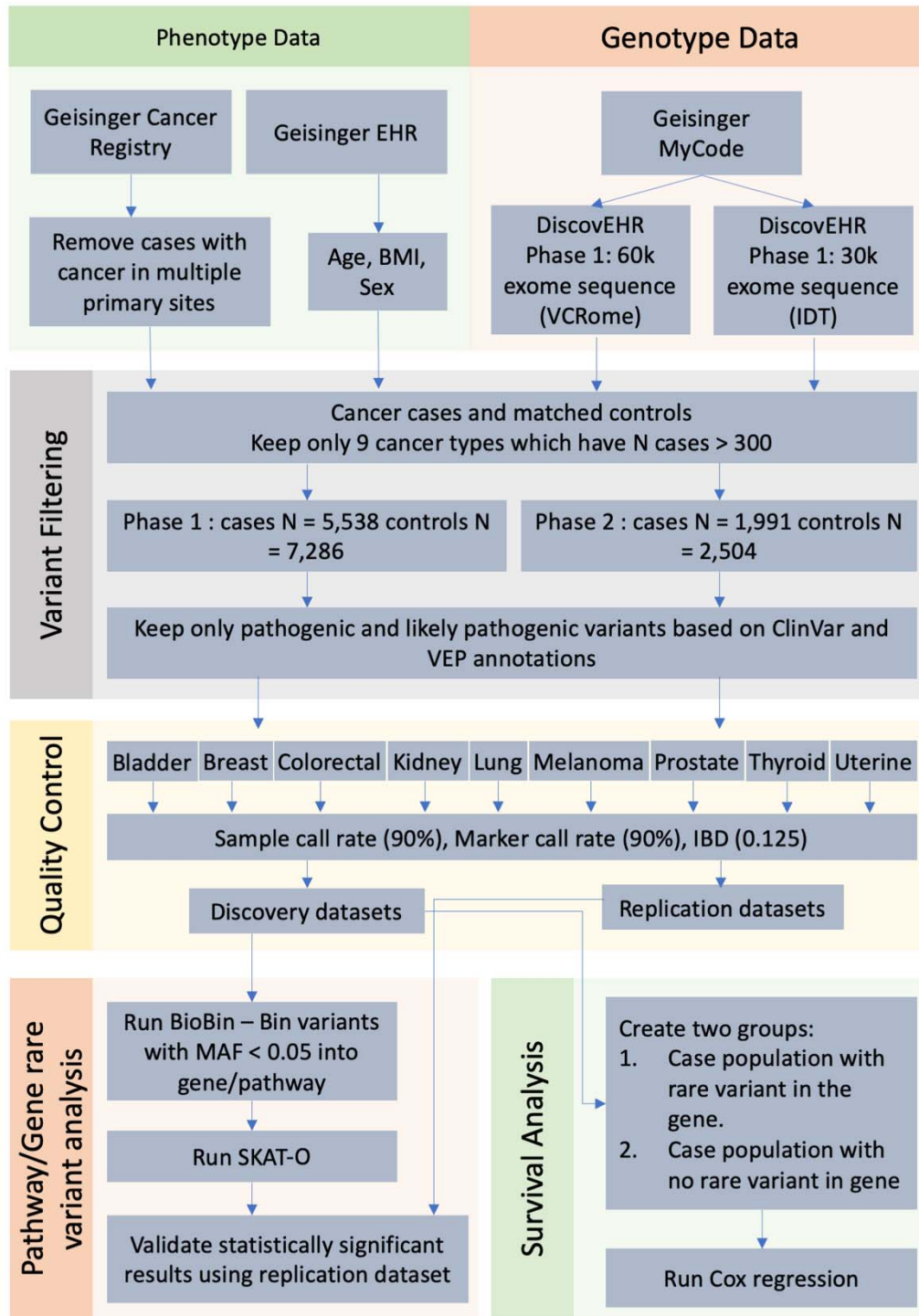
- 573 41. Ahmadi Ghezeldasht, S., Shirdel, A., Assarehzadegan, M.A., Hassannia, T., Rahimi, H., Miri,
574 R., and Rezaee, S.A.R. (2013). Human T Lymphotropic Virus Type I (HTLV-I) Oncogenesis:
575 Molecular Aspects of Virus and Host Interactions in Pathogenesis of Adult T cell
576 Leukemia/Lymphoma (ATL). *Iranian journal of basic medical sciences* 16, 179-195.
- 577 42. Jacobson, K.L., Miceli, M.H., Tarrand, J.J., and Kontoyiannis, D.P. (2008). Legionella
578 Pneumonia in Cancer Patients. *Medicine* 87.
- 579 43. Wang, C., Gu, C., Jeong, K.J., Zhang, D., Guo, W., Lu, Y., Ju, Z., Panupinthu, N., Yang, J.Y.,
580 Gagea, M., et al. (2017). YAP/TAZ-Mediated Upregulation of GAB2 Leads to Increased
581 Sensitivity to Growth Factor–Induced Activation of the PI3K Pathway. *Cancer Research*
582 77, 1637.
- 583 44. Poma, A.M., Torregrossa, L., Bruno, R., Basolo, F., and Fontanini, G. (2018). Hippo pathway
584 affects survival of cancer patients: extensive analysis of TCGA data and review of
585 literature. *Scientific reports* 8, 10623-10623.
- 586 45. Brechka, H., Bhanvadia, R.R., VanOpstall, C., and Vander Griend, D.J. (2017). HOXB13
587 mutations and binding partners in prostate development and cancer: Function, clinical
588 significance, and future directions. *Genes & diseases* 4, 75-87.
- 589 46. Ewing, C.M., Ray, A.M., Lange, E.M., Zuhlke, K.A., Robbins, C.M., Tembe, W.D., Wiley, K.E.,
590 Isaacs, S.D., Johng, D., Wang, Y., et al. (2012). Germline Mutations in HOXB13 and
591 Prostate-Cancer Risk. *New England Journal of Medicine* 366, 141-149.
- 592 47. Pilie, P.G., Giri, V.N., and Cooney, K.A. (2016). HOXB13 and other high penetrant genes for
593 prostate cancer. *Asian journal of andrology* 18, 530-532.
- 594 48. Xu, J., Lange, E.M., Lu, L., Zheng, S.L., Wang, Z., Thibodeau, S.N., Cannon-Albright, L.A.,
595 Teerlink, C.C., Camp, N.J., Johnson, A.M., et al. (2013). HOXB13 is a susceptibility gene
596 for prostate cancer: results from the International Consortium for Prostate Cancer
597 Genetics (ICPCG). *Human genetics* 132, 5-14.
- 598 49. Li, Z.-F., Wu, X.-h., and Engvall, E. (2004). Identification and characterization of CPAMD8, a
599 novel member of the complement 3/ α 2-macroglobulin family with a C-terminal Kazal
600 domain. *Genomics* 83, 1083-1093.
- 601 50. Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J., and Godzik, A. (2015). A Pan-Cancer
602 Catalogue of Cancer Driver Protein Interaction Interfaces. *PLOS Computational Biology*
603 11, e1004518.
- 604 51. Singh, S., Pillai, S., and Chellappan, S. (2011). Nicotinic acetylcholine receptor signaling in
605 tumor growth and metastasis. *Journal of oncology* 2011, 456743-456743.
- 606 52. Felix, A.S., Yang, H.P., Gierach, G.L., Park, Y., and Brinton, L.A. (2014). Cigarette smoking and
607 endometrial carcinoma risk: the role of effect modification and tumor heterogeneity.
608 *Cancer causes & control : CCC* 25, 479-489.
- 609 53. Schmit, K., and Michiels, C. (2018). TMEM Proteins in Cancer: A Review. *Frontiers in*
610 *pharmacology* 9, 1345-1345.
- 611 54. Wang, W., Qin, J.-J., Voruganti, S., Nag, S., Zhou, J., and Zhang, R. (2015). Polycomb Group
612 (PcG) Proteins and Human Cancers: Multifaceted Functions and Therapeutic
613 Implications. *Medicinal research reviews* 35, 1220-1267.
- 614 55. Slattery, M.L., Lundgreen, A., and Wolff, R.K. (2012). MAP kinase genes and colon and rectal
615 cancer. *Carcinogenesis* 33, 2398-2408.

- 616 56. Zhang, W., and Xu, J. (2017). DNA methyltransferases and their roles in tumorigenesis.
617 Biomarker research 5, 1-1.
- 618 57. Grant, K., Loizidou, M., and Taylor, I. (2003). Endothelin-1: a multifunctional molecule in
619 cancer. British journal of cancer 88, 163-166.
- 620 58. Mukhtar, E., Adhami, V.M., and Mukhtar, H. (2014). Targeting microtubules by natural
621 agents for cancer therapy. Molecular cancer therapeutics 13, 275-284.
- 622 59. Ma, J.-G., He, Z.-K., Ma, J.-H., Li, W.-P., and Sun, G. (2013). Downregulation of
623 protocadherin-10 expression correlates with malignant behaviour and poor prognosis in
624 human bladder cancer. Journal of International Medical Research 41, 38-47.
- 625 60. Sacco, J.J., Yau, T.Y., Darling, S., Patel, V., Liu, H., Urbé, S., Clague, M.J., and Coulson, J.M.
626 (2014). The deubiquitylase Ataxin-3 restricts PTEN transcription in lung cancer cells.
627 Oncogene 33, 4265-4272.
- 628 61. Lusche, D.F., Buchele, E.C., Russell, K.B., Soll, B.A., Vitolo, M.I., Klemme, M.R., Wessels, D.J.,
629 and Soll, D.R. (2018). Overexpressing TPTE2 (TPIP), a homolog of the human tumor
630 suppressor gene PTEN, rescues the abnormal phenotype of the PTEN(-/-) mutant.
631 Oncotarget 9, 21100-21121.
- 632 62. Chang, Z., Cai, C., Han, D., Gao, Y., Li, Q., Feng, L., Zhang, W., Zheng, J., Jin, J., Zhang, H., et
633 al. (2017). Anoctamin5 regulates cell migration and invasion in thyroid cancer.
634 International Journal of Oncology 51, 1311-1319.
- 635 63. Männik, J., Vaas, P., Rull, K., Teesalu, P., Rebane, T., and Laan, M. (2010). Differential
636 expression profile of growth hormone/chorionic somatomammotropin genes in
637 placenta of small- and large-for-gestational-age newborns. The Journal of clinical
638 endocrinology and metabolism 95, 2433-2442.
- 639 64. Seliger, B., and Schlaf, G. (2007). Structure, expression and function of HLA-G in renal cell
640 carcinoma. Seminars in Cancer Biology 17, 444-450.
- 641 65. Chen, Q.-R., Hu, Y., Yan, C., Buetow, K., and Meerzaman, D. (2014). Systematic genetic
642 analysis identifies Cis-eQTL target genes associated with glioblastoma patient survival.
643 PloS one 9, e105393-e105393.
- 644 66. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytzsky, A., Garimella, K.,
645 Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A
646 MapReduce framework for analyzing next-generation DNA sequencing data. Genome
647 Research 20, 1297-1303.
- 648 67. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine,
649 A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high
650 confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current
651 protocols in bioinformatics / editorial board, Andreas D Baxevanis [et al] 11, 11.10.11-
652 11.10.33.
- 653 68. Mirshahi, U.L., Luo, J.Z., Manickam, K., Wardeh, A.H., Mirshahi, T., Murray, M.F., and Carey,
654 D.J. (2018). Trajectory of exonic variant discovery in a large clinical population:
655 implications for variant curation. Genetics in Medicine.
- 656 69. Madsen, B.E., and Browning, S.R. (2009). A Groupwise Association Test for Rare Mutations
657 Using a Weighted Sum Statistic. PLOS Genetics 5, e1000384.

658 70. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006).
659 Principal components analysis corrects for stratification in genome-wide association
660 studies. *Nature Genetics* 38, 904.
661

662

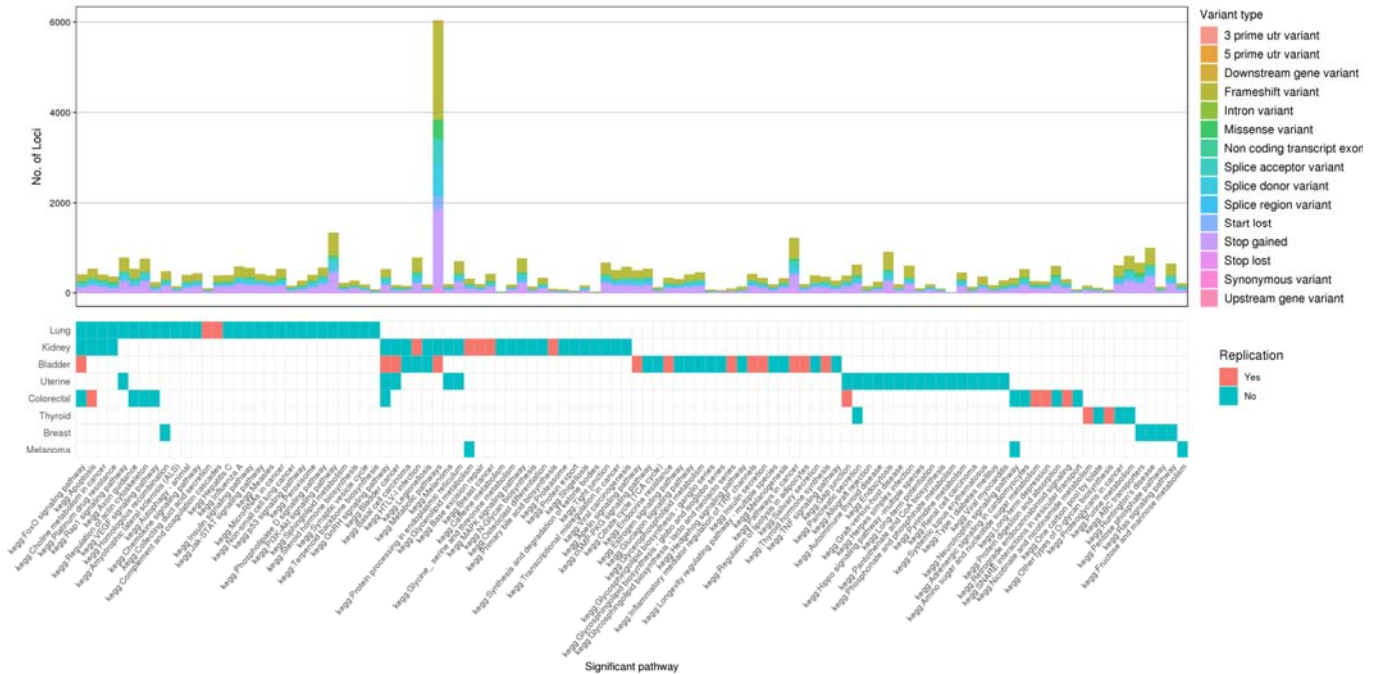
663



664
665

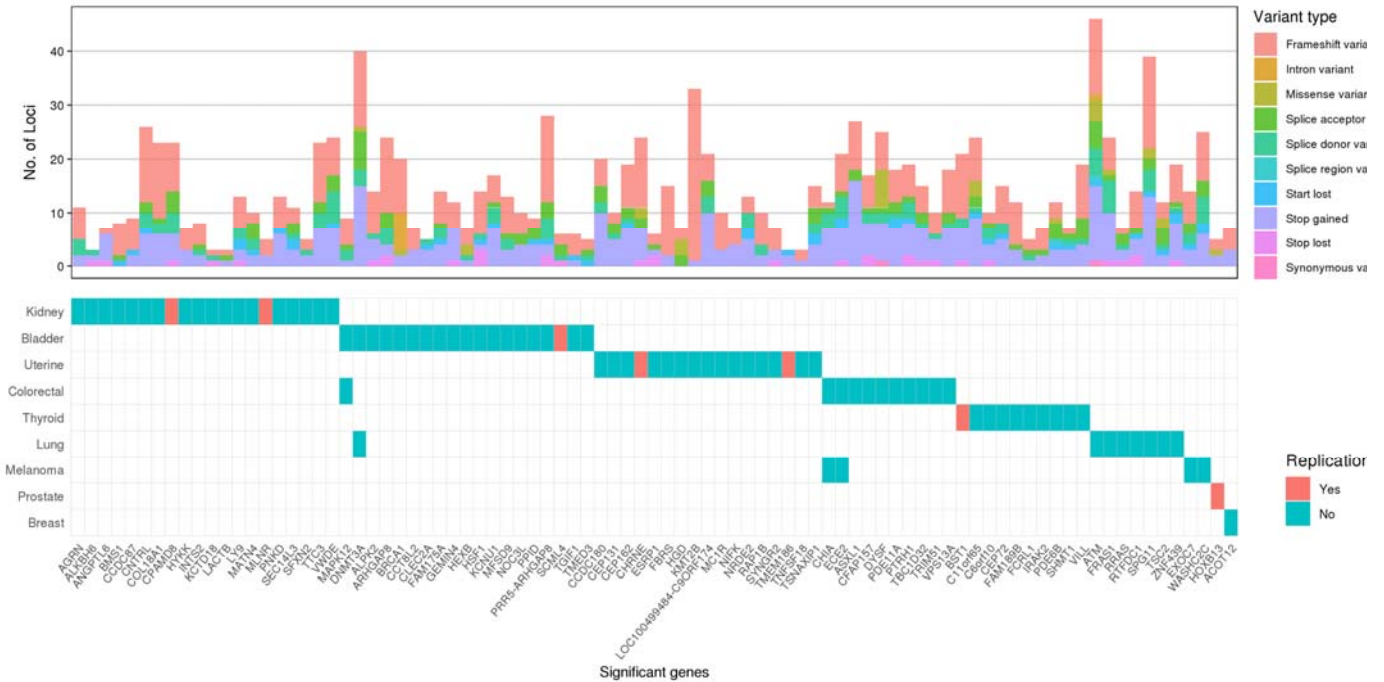
666 **Figure 1.** Schematic overview of Pan-cancer analysis. The phenotype data was obtained from the
667 Geisinger cancer registry and EHR and the genotype data was obtained from DiscovEHR study.
668 Figure shows multiple steps involved in the analysis – variant filtering, quality control, rare
669 variant analysis, and survival analysis.

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/) .



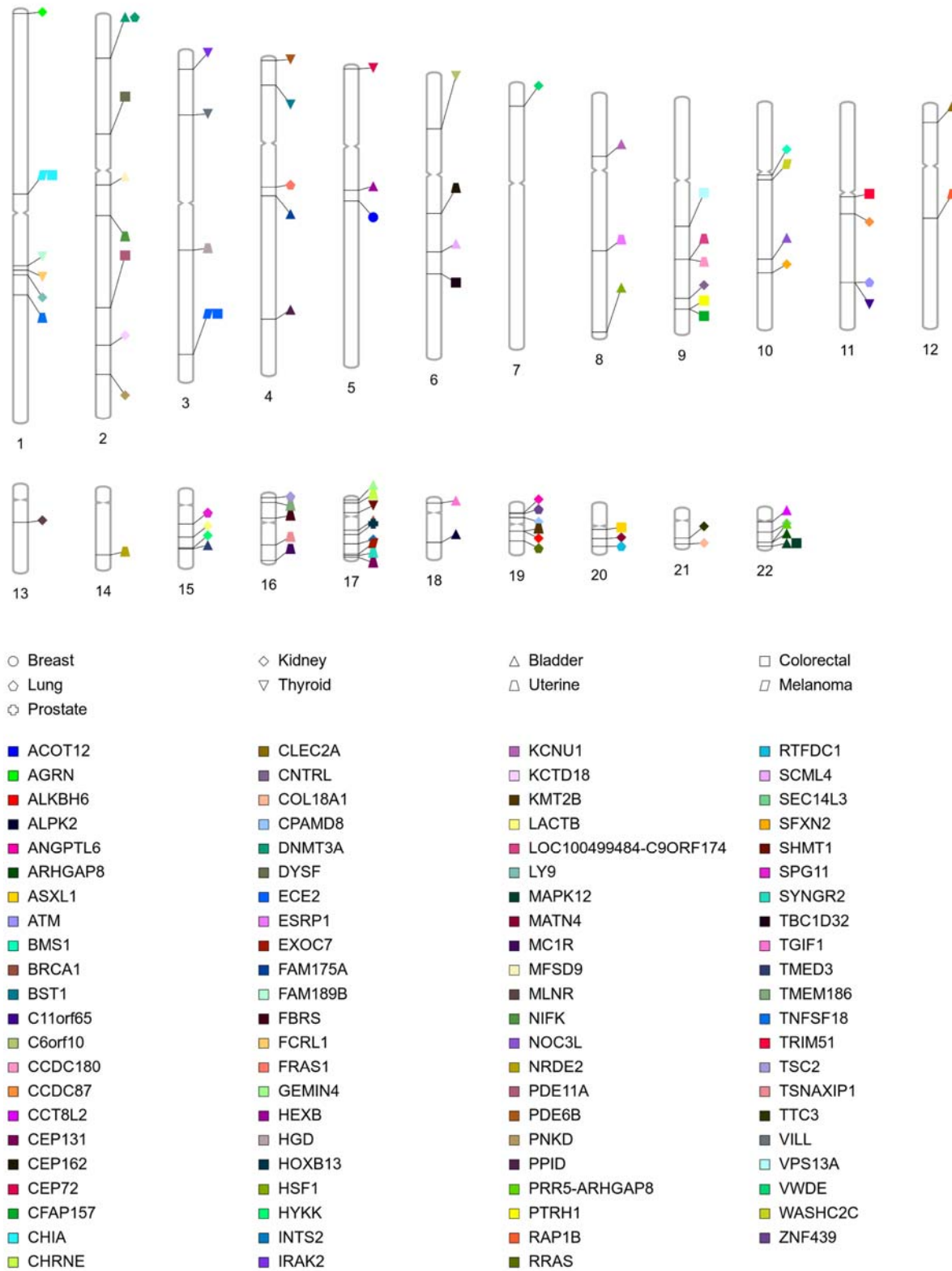
670
671 **Figure 2.** Pathways (x-axis) that were significantly associated with cancer (y-axis) (Bonferroni-
672 corrected $P < 0.05$). The pathways marked as replication 'Yes' were significant in discovery
673 (Bonferroni-corrected $P < 0.05$) and replication (SKAT-O $P < 0.05$) datasets. The top bar plot
674 shows the distribution of variant types as annotated by VEP across each pathway.

675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690



691
692 **Figure 3.** Genes (x-axis) that were significantly associated with cancer (y-axis) using Bonferroni <
693 0.05. The genes marked as replication 'Yes' were significant in discovery (Bonferroni < 0.05) and
694 replication (association p-value < 0.05) datasets. The top bar plot shows the distribution of
695 variant types as annotated by VEP across each gene.
696

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).



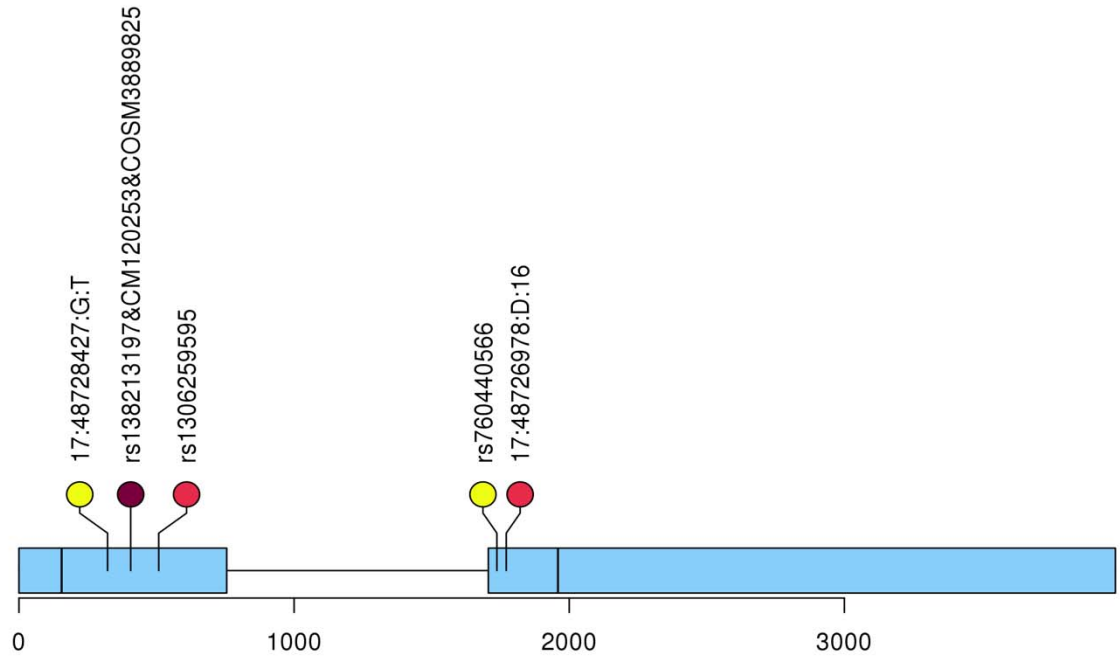
697
 698
 699
 700

Figure 4. PhenoGram plot of all significant genes across all cancers.

701

A.

● frameshift_variant ● missense_variant ● stop_gained

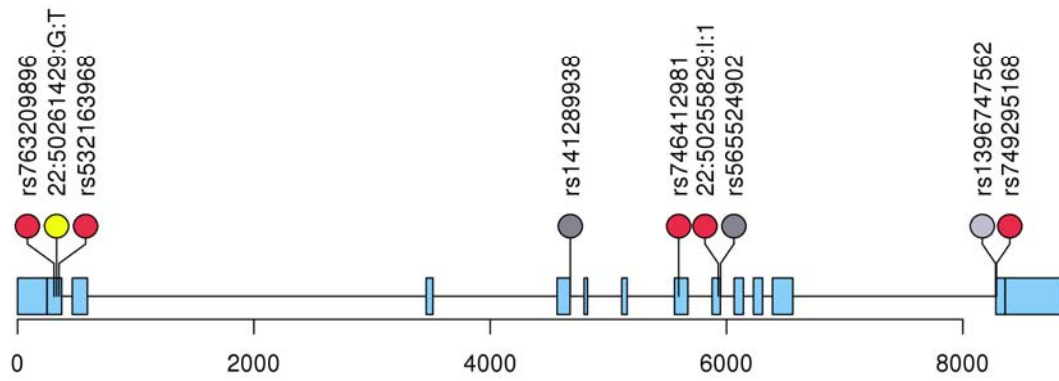


702

703

B.

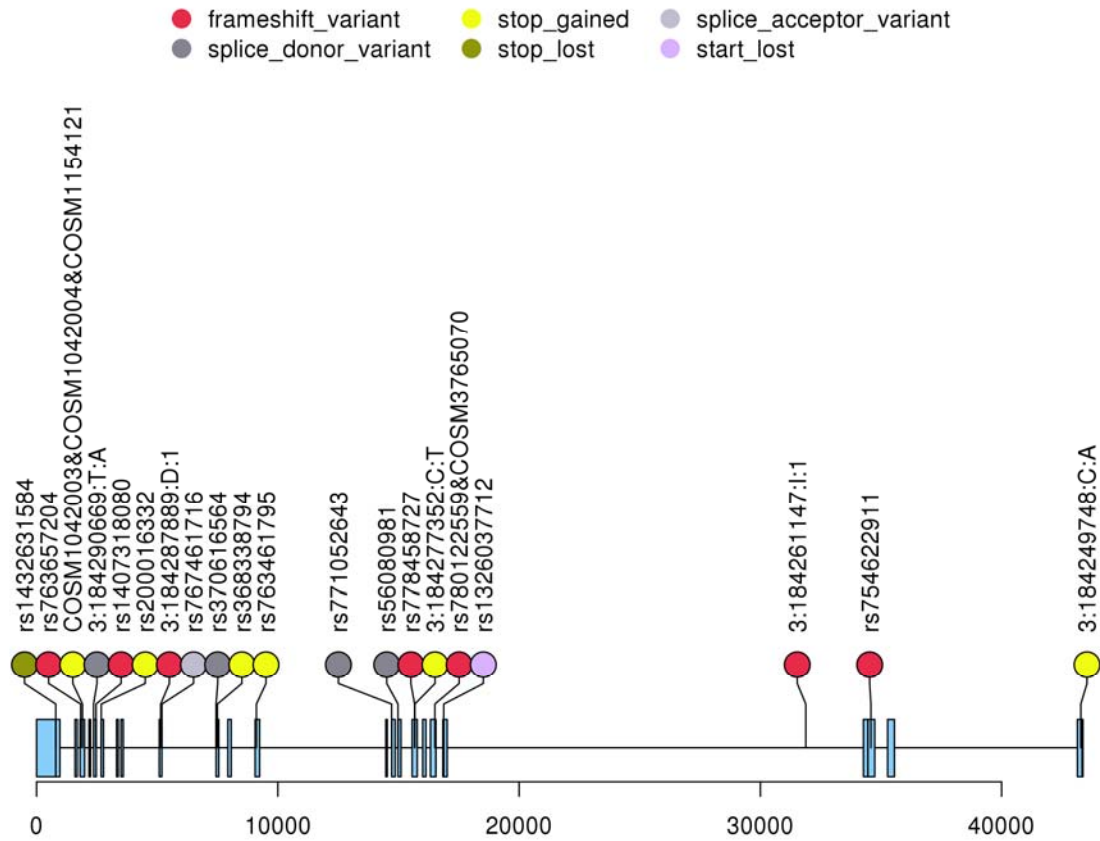
● frameshift_variant ● stop_gained
● splice_acceptor_variant ● splice_donor_variant



704

705

C.



706

707

708 **Figure 5.** Lollipop plot of genes with all loci binned in them. The color represents different types

709 of variants as assonated by VEP. A. Lollipop plot for HOXB13 gene; B. Lollipop plot for MAPK12

710 gene; C. Lollipop plot for ECE2 gene.

711

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/) .



712

713 **Figure 6.** Carrier frequency of pathogenic and likely pathogenic variants across known
 714 oncogenes and tumor suppressors.

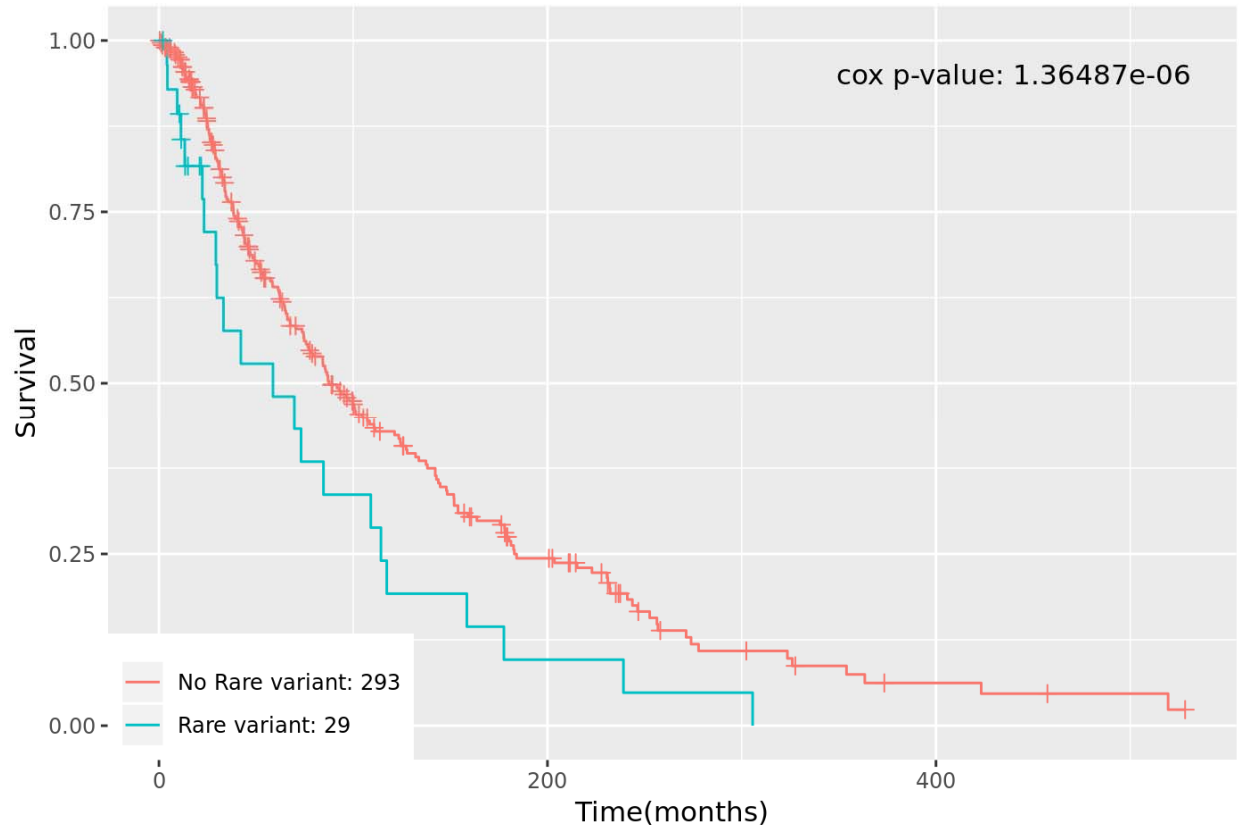
715 The color of the boxes represents the carrier frequency of pathogenic and likely pathogenic
 716 variants as denoted by the frequency legend.

717 The 7 genes significant at TFT p-value < 0.05 are marked with blue or green border.

718 The 2 genes replicated from Huang et al ¹⁴ are marked with green border.

719

720



721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743

Figure 7. Kaplan Meier survival plot for gene *PCDHB8* which was significantly associated with survival in bladder cancer. *PCDHB8* was also significantly associated with survival in bladder cancer in TCGA dataset.

744 **Table 1.** Characteristics for the cancer patients.

Cancer	Phase	Case sample size	Diagnosis age [#]	Female ratio	BMI [*]	Alive	Deceased
Bladder	1	322	67.27 +/- 11.83	18.32%	29.72 +/- 6.01	215	107
	2	93	65.72 +/- 12.65	25.81%	30.76 +/- 6.13	86	7
Breast	1	1214	59.37 +/- 11.91	100%	31.08 +/- 7.2	1036	178
	2	472	56.37 +/- 12.16	100%	30.86 +/- 6.94	454	18
Colorectal	1	477	64.37 +/- 12.29	46.54%	30.85 +/- 6.81	317	160
	2	163	59.42 +/- 12.74	44.79%	31.36 +/- 7.22	144	19
Kidney	1	309	62.19 +/- 11.53	37.22%	32.64 +/- 7.4	244	65
	2	94	60.27 +/- 11.43	39.36%	32.73 +/- 6.53	86	8
Lung	1	512	67.47 +/- 10.52	43.36%	28.82 +/- 6.61	175	336
	2	146	63.51 +/- 10.19	53.42%	29.1 +/- 7.11	92	53
Melanoma	1	730	61.32 +/- 14.8	42.05%	30.42 +/- 6.27	614	116
	2	252	57.22 +/- 15.19	45.63%	30.31 +/- 6.6	243	9
Prostate	1	1146	65.24 +/- 8.05	0%	30.01 +/- 5.25	935	211
	2	369	63.86 +/- 8.74	0%	30.16 +/- 5.36	356	13
Thyroid	1	441	48.64 +/- 14.65	79.82%	31.79 +/- 7.69	414	26
	2	101	46.08 +/- 16.14	80.20%	31.62 +/- 7.88	101	0
Uterine	1	387	60.26 +/- 11.69	100%	38.33 +/- 9.82	342	45
	2	221	61.07 +/- 10.41	100%	36.76 +/- 10.44	209	12
All	1	5538	61.85 +/- 12.71	51.97%	31.12 +/- 7.18	4292	1244
Combined	2	1911	59.38 +/- 12.77	57.61%	31.37 +/- 7.45	1771	139

745 Table shows case distribution and characteristics from Phase 1 and Phase 2.

746 [#]Average Age of patients at diagnosis in years +/- standard deviation

747 ^{*}Average BMI of the patients in kg/m² +/- standard deviation

748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763

764 **Table 2.** Pathways significantly associated with cancers in both discovery and replication
765 datasets.

Cancer	KEGG Pathway	Discovery					Replication			
		N locus	MAC Case	MAC Control	SKAT-O <i>P</i>	Bonferro ni <i>P</i>	N locus	MA C Case	MAC Contro l	SKAT-
Bladder	Insulin secretion	325	56	822	2.72E-12	8.57E-10	160	19	435	2.07E
Bladder	Bladder cancer	124	34	550	4.28E-07	1.35E-04	55	13	432	1.84E
Bladder	GnRH signaling pathway	358	81	1728	1.99E-06	6.28E-04	187	32	761	1.47E
Bladder	FoxO signaling pathway	341	63	1298	2.37E-06	7.46E-04	179	23	653	1.11E
Bladder	Inflammatory mediator regulation of TRP channels	419	87	1770	3.47E-06	1.09E-03	212	31	696	2.22E
Bladder	Pathways in cancer	1203	256	5138	1.11E-05	3.49E-03	565	100	2402	2.55E
Bladder	Metabolic pathways	5806	1225	27356	2.23E-05	7.04E-03	2906	425	11816	3.59E
Bladder	Regulation of lipolysis in adipocytes	191	62	1215	3.45E-05	1.09E-02	94	23	427	3.08E
Bladder	Glycosphingolipid biosynthesis - lacto and neolacto series	95	28	601	3.88E-05	1.22E-02	45	8	196	8.40E
Bladder	Apelin signaling pathway	492	92	1630	4.65E-05	1.47E-02	230	39	743	1.70E
Bladder	Endocrine resistance	325	70	1435	7.95E-05	2.50E-02	148	37	984	2.92E
Bladder	Thyroid hormone synthesis	352	60	1100	9.94E-05	3.13E-02	174	20	473	2.05E
Colorectal	Gap junction	269	64	875	1.36E-07	4.27E-05	128	27	402	8.70E
Colorectal	Retrograde endocannabinoid signaling	296	72	960	1.06E-06	3.33E-04	150	36	603	7.34E
Colorectal	Amino sugar and nucleotide sugar metabolism	252	66	947	1.06E-05	3.35E-03	121	23	321	5.19E
Colorectal	Long-term depression	250	76	999	2.25E-05	7.10E-03	126	42	570	1.27E
Colorectal	Apoptosis	474	132	1930	1.06E-04	3.33E-02	246	85	953	1.00E
Kidney	Base excision repair	191	49	865	9.83E-10	3.10E-07	88	10	242	4.25E
Kidney	Primary bile acid biosynthesis	88	17	199	1.60E-05	5.03E-03	38	5	70	6.18E
Kidney	HTLV-I infection	728	248	5636	2.19E-05	6.90E-03	342	101	2260	4.55E
Kidney	Glycerolipid metabolism	282	74	1333	7.03E-05	2.22E-02	152	26	841	1.50E
Kidney	Breast cancer	413	60	1116	8.80E-05	2.77E-02	184	55	1137	2.62E
Lung	Collecting duct acid secretion	98	23	166	2.29E-06	7.22E-04	47	6	104	1.11E
Lung	Complement and coagulation cascades	383	164	2092	1.05E-04	3.31E-02	188	22	748	2.26E
Thyroid	Nicotinate and nicotinamide metabolism	158	64	942	5.73E-05	1.80E-02	75	15	389	9.20E
Thyroid	Other types of O-glycan biosynthesis	75	10	134	1.22E-04	3.85E-02	45	5	57	3.10E

766 N locus: Total number of genomic loci binned in pathway.

767 MAC Case: Total minor allele count of variants in pathway in case population.

768 MAC Control: Total minor allele count of variants in pathway in control population.

769

770

771

772

773

774 **Table 3.** Pathways significantly associated with more than one cancer.

KEGG Pathway	N Cancers	Cancers
Apoptosis	3	Colorectal, Kidney, Lung
Axon guidance	2	Colorectal, Lung
Basal cell carcinoma	2	Bladder, Kidney
Bladder cancer	3	Bladder, Kidney, Uterine
Choline metabolism in cancer	2	Kidney, Lung
FoxO signaling pathway	4	Bladder, Colorectal, Kidney, Lung
Gap junction	2	Colorectal, Uterine
Glycerolipid metabolism	2	Kidney, Melanoma
GnRH signaling pathway	4	Bladder, Colorectal, Kidney, Uterine
Homologous recombination	2	Breast, Lung
HTLV-I infection	2	Bladder, Kidney
Legionellosis	2	Bladder, Kidney
Melanoma	2	Kidney, Uterine
Metabolic pathways	2	Bladder, Kidney
Neurotrophin signaling pathway	2	Colorectal, Melanoma
Pancreatic secretion	2	Thyroid, Uterine
Platinum drug resistance	2	Kidney, Lung
Protein processing in endoplasmic reticulum	2	Kidney, Uterine
Rap1 signaling pathway	2	Lung, Uterine
Regulation of actin cytoskeleton	2	Colorectal, Lung
VEGF signaling pathway	2	Colorectal, Lung

775 All pathways were significant in the cancers provided in column 3 with Bonferroni-corrected $P < 0.05$

776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792

793 **Table 4.** Genes associated with cancers that were replicated.

Cancer	Gene information		Discovery					Replication			
	Gene	Chr: Build 38 position	N locus	MAC Case	MAC Control	SKAT-O p-value	Bonferroni corrected p-value	N locus	MAC Case	MAC Control	SKAT-O p-value
Kidney	<i>MLNR</i>	13: 49220338-49222377	5	6	39	1.07E-07	3.70E-04	3	7	48	7.89E-04
Kidney	<i>CPAMD8</i>	19: 16892947-17026818	23	9	75	2.24E-07	7.70E-04	14	3	43	1.29E-04
Uterine	<i>CHRNE</i>	17: 4801064-4806369	17	9	95	7.92E-07	2.70E-03	17	8	27	1.29E-04
Prostate	<i>HOXB13</i>	17: 48724763-48728749	5	19	37	2.86E-06	1.06E-02	2	8	14	5.54E-04
Bladder	<i>SCML4</i>	6: 107697297-107845959	6	5	116	2.26E-07	7.70E-04	4	3	18	1.13E-04
Thyroid	<i>BST1</i>	4: 15701866-15774178	21	5	41	2.36E-07	8.20E-04	13	4	49	2.56E-04
Uterine	<i>TMEM186</i>	16: 8795180-8797648	3	13	101	8.66E-08	3.00E-04	2	10	47	4.49E-04

794 N locus: Total number of genomic loci binned in gene.

795 MAC Case: Total minor allele count of variants in gene in case population.

796 MAC Control: Total minor allele count of variants in gene in control population.

797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813

814 **Table 5. Genes significantly associated with more than one cancer.**

Cancer	Gene information		Discovery				
Cancer	Gene	Chr: Build 38 position	N locus	MAC Case	MAC Control	SKAT-O p-value	Bonferroni corrected p-value
Bladder	<i>MAPK12</i>	22: 50252901-50261810	7	6	42	1.18E-06	4.02E-03
Colorectal	<i>MAPK12</i>	22: 50252901-50261810	9	8	42	1.18E-05	4.08E-02
Colorectal	<i>ECE2</i>	3: 184276011-184293031	17	4	17	7.23E-06	2.50E-02
Melanoma	<i>ECE2</i>	3: 184276011-184293031	20	6	17	2.93E-06	1.04E-02
Lung	<i>DNMT3A</i>	2: 25232961-25342590	30	10	28	6.62E-10	2.00E-06
Bladder	<i>DNMT3A</i>	2: 25232961-25342590	25	6	27	2.21E-06	7.53E-03
Colorectal	<i>CHIA</i>	1: 111290852-111320566	9	9	70	4.14E-10	1.00E-06
Melanoma	<i>CHIA</i>	1: 111290852-111320566	9	8	70	1.38E-05	4.89E-02

815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843

844 **Table 6. Genes associated with survival**

Cancer	Gene	N cases [#]	N controls [*]	Cox p-value	FDR	Bonferroni	Permutator p-value
Thyroid	<i>NAA38</i>	10	430	1.67E-05	9.80E-03	9.80E-03	2.98E-03
Melanoma	<i>THNSL1</i>	18	712	1.75E-06	1.60E-03	1.60E-03	8.40E-04
Uterine	<i>SAXO2</i>	17	370	7.56E-05	1.95E-02	3.89E-02	6.20E-04
	<i>DCHS2</i> [^]	19	368	9.71E-07	5.00E-04	5.00E-04	1.50E-04
	<i>DBF4B</i>	13	374	1.77E-04	3.04E-02	9.15E-02	5.84E-03
Bladder	<i>PCDHB8</i> [^]	29	293	1.36E-06	6.00E-04	6.00E-04	4.50E-04
Breast	<i>ANO5</i>	15	1199	1.75E-04	4.92E-02	2.46E-01	1.56E-03
	<i>HOGA1</i>	10	1204	4.23E-05	1.98E-02	5.94E-02	2.39E-03
	<i>CSH2</i>	22	1192	3.71E-07	9.80E-03	9.80E-03	2.98E-03
	<i>ATXN3</i>	10	1204	3.89E-05	1.60E-03	1.60E-03	8.40E-04
	<i>FAM186A</i>	35	1179	1.64E-04	1.95E-02	3.89E-02	6.20E-04
Prostate	<i>TPTE2</i>	15	1131	3.25E-05	5.00E-04	5.00E-04	1.50E-04
Kidney	<i>HLA-G</i>	10	299	5.54E-06	3.04E-02	9.15E-02	5.84E-03

845 [#] Number of cancer patients who have rare variants in given gene

846 ^{*} Number of cancer patients who do not have rare variants in given gene

847 [^] Genes significantly associated with survival in TCGA data- *PCDHB8* (Logrank P = 9.22E-03) and *DCHS2* (Logrank P =
848 0.036)

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866 **Table 7.** Pathways associated with survival

Cancer	KEGG Pathway	N cases [#]	N controls [*]	Cox p-value	FDR	Bonferroni	Permutation p-value
Melanoma	Phenylalanine tyrosine and tryptophan biosynthesis	19	711	6.77E-05	2.05E-02	2.05E-02	5.20E-04
Melanoma	Phenylalanine metabolism	31	699	1.57E-04	2.39E-02	4.77E-02	3.90E-04

867 [#] Number of cancer patients who have rare variants in given gene

868 ^{*} Number of cancer patients who do not have rare variants in given gene

869