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

Manu Shivakumar[1,2], Jason E. Miller[2], Venkata Ramesh Dasari[3], David Carey[4], 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

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

Introduction

Cancer is the second most lethal disease in United States with an estimated 1,735,350 new cases and 609,640 deaths in 2018[1]. Cancer is caused by inherited germline variants and acquired somatic mutations. A recent twin study showed ~33% heritability of cancer across 23 cancer types with a high estimate of 57% for prostate (MIM: 176807), 31% for breast (MIM: 114480), 38% for kidney (MIM: 144700), and 58% for skin melanoma (MIM: 155600)[2]. Germline genetic markers for cancers have been widely studied leading to the discovery of many heritable predisposition genes such as *BRCA1* (MIM: 113705), *BRCA2* (MIM: 600185) and *PALB2* (MIM: 610355) in breast cancer, *RB1* (MIM: 614041) in retinoblastoma and *MLH1* (MIM: 120436), *MSH2* (MIM: 609309), *MSH6* (MIM: 600678), and *PMS2* (MIM: 600259) in Lynch syndrome (MIM: 120435). To date, many genome-wide association studies (GWAS) have been conducted and many more variants and genes have been discovered as associated with various cancer types[3-8]. However, a large portion of inherited genetic factors that result in carcinogenesis is still unknown and many studies are being undertaken to discover these

genetic variants. For instance, the genetic contribution explained by all variants discovered to date is about 39% in prostate cancer[2; 9] and 30% in breast cancer[2; 10].

Since common variants discovered to be associated with multiple cancers have only modest effect size, the missing heritability could be further explained by rare variants. Moreover, rare variants have been known to contribute to various complex diseases including cancer[11-13]. The aggregation of rare variants in a gene can lead to loss of function of the gene or change in expression[14]. Similarly, since pathways perform a sequence of biochemical actions leading to a cellular function or product, changes in the expression of genes involved within a pathway can lead to cancer[15; 16]. Previous studies have also indicated that cancer is caused by an accumulation of a number of singular or rare variants in particular genes or pathways[12]. To that effect, binning the pathogenic and likely pathogenic rare variants into genes and pathways would help us increase statistical power to detect associations and also infer biological mechanisms[13; 17; 18].

The MyCode community initiative is a precision medicine project, launched at Geisinger in 2007, which enabled the storage of blood, serum, and DNA samples in a system-wide biorepository that is available for use in broad research[19]. To date, over 244,000 patients have signed up for the MyCode initiative and over 90,000 patient blood samples have been sequenced as part of the DiscovEHR project in collaboration with the Regeneron Genetics Center[20]. The sequenced data can be easily linked to the electronic health record (EHR) of the patient, allowing access to rich longitudinal data. Apart from the EHR, Geisinger also maintains

a cancer registry that contains all the patients diagnosed or treated for cancer at any Geisinger medical facility. Further, as part of the MyCode program, the genetic data is also being used to detect increased risk of developing one or more of 21 medically actionable conditions, including breast cancer, ovarian cancer (MIM: 167000), Marfan syndrome (MIM: 154700), Lynch syndrome, etc., and the results are returned to the patients though a "Return of Results" program[21]. Moreover, many similar programs around the world are helping to integrate genomics into clinical practice[22]. Manolio et al. emphasize that the integration of genomic findings to clinical practice has been relatively slow and insist on the need to have an openly accessible knowledge base of variants, phenotypes and clinically actionable variants[23]. The sharing of genetic findings is likely to help the scientific research community to improve our understanding of the phenotype of interest and propel precision medicine by bringing more genomics into clinical practice.

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

Results

Study design and population characteristics

This study was based on a subset of 7,449 cancer cases and 9,792 controls selected by matching age, BMI and gender from ~90,000 sequenced samples from the DiscovEHR study. The samples were sequenced in two phases using different platforms as described in the Methods section. In phase 1, 60,000 samples were sequenced, and 5,538 cancer patients across nine cancers and 7,286 matched controls were pulled. In phase 2, 30,000 samples were sequenced which included 1,991 cancer patients and 2,504 matched controls. Consequently, the phase 2 dataset was used to replicate results from phase 1. Cancer patient IDs that were retrieved from the cancer registry were classified into particular cancers using International Classification of Diseases for Oncology (ICD-O) codes. After classifying the cancer patients to their respective cancers, only nine cancers, including bladder (MIM: 109800), breast, colorectal (MIM: 114500), kidney, lung (MIM: 211980), melanoma, prostate, thyroid (MIM: 188550), and uterine cancer, had more than 300 samples in the discovery set. A low number of samples in association studies results in a higher type 1 error rate and lower statistical power to detect associations[24]. Thus, the rest of the cancers were excluded from this study. The distribution and basic demographics of patients across these cancers are shown in Table 1. A common set of controls were used for all the cancers except breast, uterine and prostate cancer as they are sex-specific

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

Table 1 provides age, BMI, female ratio, and vital status (alive or deceased) information across cancers. Among all cancers, breast cancer had the largest number of cases (N = 1,214) followed by prostate cancer (N = 1,146). Further, uterine cancer had a significantly higher average BMI as compared to other cancers. The average BMI for uterine cancer patients was 38.33 kg/m^2 in the discovery dataset and 36.76 kg/m^2 in the replication dataset. Additionally, lung cancer had the highest number of cases who are deceased, which is expected as lung cancer is by far the leading cause of death due to cancer[1]. Further, observing the female ratio across the cancers also shows a gender disparity in some cancers. Specifically, the incidence rate in bladder cancer was found to be 4.46 fold higher in male than female, and in thyroid cancer it was 3.95 fold higher in female than male. The difference in incidence rates have been well documented in other studies with a 3-4 times increased risk of bladder cancer risk in men[25] and 2.9 times increased risk of thyroid cancer in women[26].

Variant filtering based on functional annotation and scores improves power and has been successfully used in many association studies[14; 27]. In this study, the variants from whole exome sequence data were annotated using Variant Effect Predictor (VEP)[28] and ClinVar[29]. Subsequently, only the variants categorized as pathogenic and likely-pathogenic based on the annotations were retained for further analysis. A strategy for classifying variants as pathogenic and likely pathogenic is elaborated in the Methods section. Additionally, all common variants

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

Pathway-based rare variant analysis

Association analysis using rare variants usually suffers from a lack of power as very large datasets are required. Therefore, rare variants are often binned into a biologically informed-unit such as a gene or pathway to improve the power[17]. In this study, pathogenic and likely pathogenic rare variants with minor allele frequency (MAF) < 0.05 were binned into genes followed by KEGG pathways using BioBin[17; 30; 31]. Next, an association test was performed to determine if the gene/pathway is significantly associated with the phonotype. SKAT-O is an optimal unified approach that combines a burden and non-burden sequence kernel association test (SKAT) test, and maintains power regardless of the direction of effect and causality of the variants[32]. After determining the association p-values, they were adjusted for multiple testing in each cancer type separately using a Bonferroni correction. Any genes/pathways with a Bonferroni-corrected $P < 0.05$ were considered as significant results. The same procedure was followed to conduct rare variant analysis in discovery and replication datasets. Figure 2 shows all the pathways that were Bonferroni significant across all cancers and the same are listed in Table S4-S6.

In total, 106 pathways were found to be significantly associated across all cancers (Figure 2). However, no significant pathways were found in prostate cancer. Further, 26 pathways: 12 in

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

Gene-based rare variant analysis

All pathogenic and likely pathogenic variants below MAF < 0.05 were binned into gene boundaries defined by Entrez annotations derived from Library of Knowledge Integration (LOKI) using BioBin[17; 30]. The total number of genes that the variants were binned across all cancers is listed in Table S2. The bar plot in Figure 3 shows the total number of loci binned for a given gene and variant types as annotated by VEP. In total, there were 87 genes that were significantly associated with a specific cancer (Bonferroni-corrected $P < 0.05$) (Figure 3). Furthermore, seven genes - *MLNR* (MIM: 602885), *CPAMD8* (MIM: 608841), *CHRNE* (MIM: 100725), *HOXB13* (MIM: 604607), *SCML4* (HGNC: 256380), *BST1* (MIM: 600387) and *TMEM186* (HGNC: 25880), marked in red in Figure 3, were replicated in the phase 2 dataset ($P < 0.05$) (Table 4).

Additionally, four genes, including *MAPK12* (MIM: 602399)*, ECE2* (MIM: 610145)*, DNMT3A* (MIM: 602769) *and CHIA* (MIM: 606080)*,* were significantly associated in multiple cancers. The gene *MAPK12* was significantly associated with bladder cancer and colorectal cancer, *ECE2* and *CHIA* for melanoma and colorectal cancer, and *DNMT3A* for bladder and lung cancer. Further association test statistics on these genes are described in Table 5. The PhenoGram[33] plot in Figure 4 shows all the genes found to be significantly associated across all cancers. Additionally, the lollipop[34] plots in Figure 5 and Figure S1 shows the type of variants – frameshift, missense, stop gained, stop lost, splice acceptor, splice donor, start lost, and their relative position in the gene. Variants that were found in the Catalog of Somatic Mutation in Cancer (COSMIC) database were marked with COSMIC ids. The nonsense stop gained variants marked in yellow usually results in a truncated protein, which are non-functional and the frameshift variants marked in red usually cause a loss of function due to a shift in the reading frame.

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

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

help validate cancer susceptibility genes and even discover new gene associations that were not statistically significant in Huang et al. As such, the list of gene-cancer pairs were filtered to the nine cancer types under consideration in this study and TFT was run on the genes to validate the enrichment of pathogenic and likely pathogenic variants in cancer patients against control group. Seven genes – *ATM* (lung, MIM: 607585), *CHECK2* (breast, HGNC: 11200), *MSH2* (colorectal), *BRCA2* (thyroid), *POLE* (kidney, MIM: 174762), *PALB2* (breast), *MLH1* (colorectal) were found to be significant at a TFT p-value < 0.05. Moreover, two of the genes – *ATM* (lung) and *BRCA2* (thyroid) found to be previously significant [14] were replicated in this study. Further, three genes – *ATM* (lung), *CHECK2* (breast) and *MSH2* (colorectal) were significant at FDR < 0.15. The carrier frequency of these genes and distribution of pathogenic and likely pathogenic variants across oncogenes and tumor suppressor genes is shown in Figure 6 (Table S8). In summary, we were able to validate two genes (TFT p-value < 0.05) and we discovered two more genes with a suggestive association (FDR < 0.15).

Survival analysis

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

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

Discussion

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

Many KEGG pathways identified in this study have already been implicated in cancer, such as "pathways in cancer", "GnRH signaling pathway" [36], "bladder cancer", "FoxO signaling pathway"

[37], "metabolic pathways", "gap junction" [38], "apoptosis", "base excision repair", "melanoma", "choline metabolism in cancer" and "basal cell carcinoma". One pathway of interest that is associated with bladder cancer is the "insulin secretion pathway". Previous studies have shown diabetes mellitus increases the risk of bladder cancer[39] and is possibly due to administration of the anti-diabetic drug Pioglitazone[40]. Another pathway "HTLV-I infection" was found to be associated with kidney cancer and was also replicated. HTLV-I is a known oncovirus that causes cancer[41]. Further studies on the "HTLV-I infection" pathway could elucidate the role of germline variants in cancers. Another pathway "Legionellosis" was found to be associated with kidney and bladder cancer, Legionella pneumonia in cancer has a very high mortality rate ~31%[42], and variants in pathway could play a role in susceptibility or recovery of the patients. Another pathway, the "Hippo signaling pathway" was found to be associated with uterine cancer. The "Hippo tumor suppressor pathway" is known to phosphorylate YAP and TAZ which are critical for cell growth, reprogramming and development[43]. The Hippo pathway also interacts with the PI3K/AKT pathway which is commonly involved in cancer[43]. Additionally, Hippo pathway is known to affect the survival of cancer patients[44] and in this study, one of the genes *DCHS2*, which is part of Hippo pathway, was also associated with survival in uterine cancer and it was also replicated in the TCGA data.

A number of previous studies have shown *HOXB13* to be associated with prostate cancer[45-48], and in this study as well, *HOXB13* was found to be associated with prostate cancer in the discovery dataset and was replicated. Another gene, *CPAMD8,* which is involved in broad-spectrum protease inhibition, innate immunity and damage control was found to be associated

with kidney cancer in the discovery and replication datasets[49]. *CPAMD8* is known to be substantially expressed in kidney[49; 50] and given its functional role, rare gene-disruptive variants in *CPAMD8* could lead to carcinogenesis. We also identified two genes associated with uterine cancer that replicated - *CHRNE* which is a subunit of nicotinic acetylcholine receptors (nAChRs) and *TMEM186* which is a member of the transmembrane protein family. Nicotine, a compound present in cigarettes, mediates cell proliferation and angiogenesis though nicotinic acetylcholine receptors (nAChRs) and its subunits[51]. Still, its mechanism of action is not well understood for uterine cancer where some studies have shown smoking to reduce the risk of uterine cancer contrary to other cancers[51; 52]. Again, the exact role of *TMEM186* in uterine cancer is also unexplored. TMEMs are differentially regulated in many types of cancers and some TMEMs are known to act as tumor suppressors while others as oncogenes[53]. Further, in bladder cancer, the Putative Polycomb group (PcG) protein gene (*SCML4*), is involved in the regulation of crucial developmental and physiological processes and is known to promote proliferation and inhibit apoptosis. *SCML4* was replicated in bladder cancer[54].

Different cancer types share some pathways and genes, which generally include common tumor suppressor genes and oncogenes[14]. In this study, we identified 21 pathways and four genes that were associated with multiple cancers. Two of the genes *MAPK12* and *DNMT3A* are well known genes involved in cancer with *MAPK12* acting as *p38 MAPK,* which is involved in cell differentiation, apoptosis and autophagy, whereas *DNMT3A* is involved in DNA methylation and its disruption leads to tumorigenesis[55; 56]. Gene *ECE2* cleaves endothelin-1 (ET-1) which is a potent vasoconstrictor peptide and ET-1 is known to be involved in angiogenesis, apoptosis and

294 growth in colorectal cancer and melanoma[57]. Elevated levels of plasma levels and increased
295 immunopositivity of ET-1 has been observed in colorectal cancer[57].
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 cancer[58], *PCDHB8* whose downregulation is known to result in poor prognosis in bladder
303 cancer[59] , *ATXN3* (MIM: 607047) whose downregulation increases expression of tumor
304 suppressor *PTEN* (MIM: 601728)[60], and *TPTE*2 (MIM: 606791) , a homolog of *PTEN*, whose
305 upregulation suppresses metastasis and/or tumorigenesis[61]. Other genes were associated with
306 survival in this study - *ANO5* (MIM: 608662) is known to regulate cell migration and invasion[62],
307 the *HOGA1* (MIM: 613597) gene is involved in metabolism, *CSH2* (MIM: 118820) is involved in
308 postnatal and intrauterine growth[63] 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 presentation[64]. 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

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

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

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

Sequencing and quality control

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

Variant annotation and filtering

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

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

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

Gene based rare variant association

All the variants that were annotated as pathogenic and likely pathogenic were binned using BioBin[17; 30]. BioBin uses pre-compiled knowledge in a LOKI database, which is compiled using information from various data sources including Entrez and KEGG. The variants were binned into genes using Entrez annotations. Only variants with MAF < 0.05 were considered rare and the rest of the variants were filtered out. Additionally, bins with less than 20 variants (MAC) were filtered out due to low sample size. Further, the binned variants were weighed using Madsen-browning weights[69]. The statistical association tests were run using SKAT-O implemented as R package[32]. Additionally, the association tests were adjusted using age, BMI and first four principle components as covariates. The principle components were calculated using EIGENSOFT[70], using common variants after LD pruning with indep-pairwise 50 5 0.5 and Hardy-Weinberg equilibrium of 10^{-6}. The association test p-values were further adjusted using Bonferroni correction to account for multiple testing correction.

Pathway based rare variant association

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

Madsen-browning weighting[69]. Statistical association was run using SKAT-O implemented as R package [32]. The association tests were adjusted using age, BMI and four principle components as covariates. Further, the association p-values were adjusted using Bonferroni correction.

Survival analysis

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

Author contributions

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

Conflicts of Interest

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

References

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

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

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

41. Ahmadi Ghezeldasht, S., Shirdel, A., Assarehzadegan, M.A., Hassannia, T., Rahimi, H., Miri, R., and Rezaee, S.A.R. (2013). Human T Lymphotropic Virus Type I (HTLV-I) Oncogenesis: Molecular Aspects of Virus and Host Interactions in Pathogenesis of Adult T cell Leukemia/Lymphoma (ATL). Iranian journal of basic medical sciences 16, 179-195.
42. Jacobson, K.L., Miceli, M.H., Tarrand, J.J., and Kontoyiannis, D.P. (2008). Legionella Pneumonia in Cancer Patients. Medicine 87.
43. Wang, C., Gu, C., Jeong, K.J., Zhang, D., Guo, W., Lu, Y., Ju, Z., Panupinthu, N., Yang, J.Y., Gagea, M., et al. (2017). YAP/TAZ-Mediated Upregulation of GAB2 Leads to Increased Sensitivity to Growth Factor–Induced Activation of the PI3K Pathway. Cancer Research 77, 1637.
44. Poma, A.M., Torregrossa, L., Bruno, R., Basolo, F., and Fontanini, G. (2018). Hippo pathway affects survival of cancer patients: extensive analysis of TCGA data and review of literature. Scientific reports 8, 10623-10623.
45. Brechka, H., Bhanvadia, R.R., VanOpstall, C., and Vander Griend, D.J. (2017). HOXB13 mutations and binding partners in prostate development and cancer: Function, clinical significance, and future directions. Genes & diseases 4, 75-87.
46. Ewing, C.M., Ray, A.M., Lange, E.M., Zuhlke, K.A., Robbins, C.M., Tembe, W.D., Wiley, K.E., Isaacs, S.D., Johng, D., Wang, Y., et al. (2012). Germline Mutations in HOXB13 and Prostate-Cancer Risk. New England Journal of Medicine 366, 141-149.
47. Pilie, P.G., Giri, V.N., and Cooney, K.A. (2016). HOXB13 and other high penetrant genes for prostate cancer. Asian journal of andrology 18, 530-532.
48. Xu, J., Lange, E.M., Lu, L., Zheng, S.L., Wang, Z., Thibodeau, S.N., Cannon-Albright, L.A., Teerlink, C.C., Camp, N.J., Johnson, A.M., et al. (2013). HOXB13 is a susceptibility gene for prostate cancer: results from the International Consortium for Prostate Cancer Genetics (ICPCG). Human genetics 132, 5-14.
49. Li, Z.-F., Wu, X.-h., and Engvall, E. (2004). Identification and characterization of CPAMD8, a novel member of the complement 3/α2-macroglobulin family with a C-terminal Kazal domain. Genomics 83, 1083-1093.
50. Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J., and Godzik, A. (2015). A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. PLOS Computational Biology 11, e1004518.
51. Singh, S., Pillai, S., and Chellappan, S. (2011). Nicotinic acetylcholine receptor signaling in tumor growth and metastasis. Journal of oncology 2011, 456743-456743.
52. Felix, A.S., Yang, H.P., Gierach, G.L., Park, Y., and Brinton, L.A. (2014). Cigarette smoking and endometrial carcinoma risk: the role of effect modification and tumor heterogeneity. Cancer causes & control : CCC 25, 479-489.
53. Schmit, K., and Michiels, C. (2018). TMEM Proteins in Cancer: A Review. Frontiers in pharmacology 9, 1345-1345.
54. Wang, W., Qin, J.-J., Voruganti, S., Nag, S., Zhou, J., and Zhang, R. (2015). Polycomb Group (PcG) Proteins and Human Cancers: Multifaceted Functions and Therapeutic Implications. Medicinal research reviews 35, 1220-1267.
55. Slattery, M.L., Lundgreen, A., and Wolff, R.K. (2012). MAP kinase genes and colon and rectal cancer. Carcinogenesis 33, 2398-2408.

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

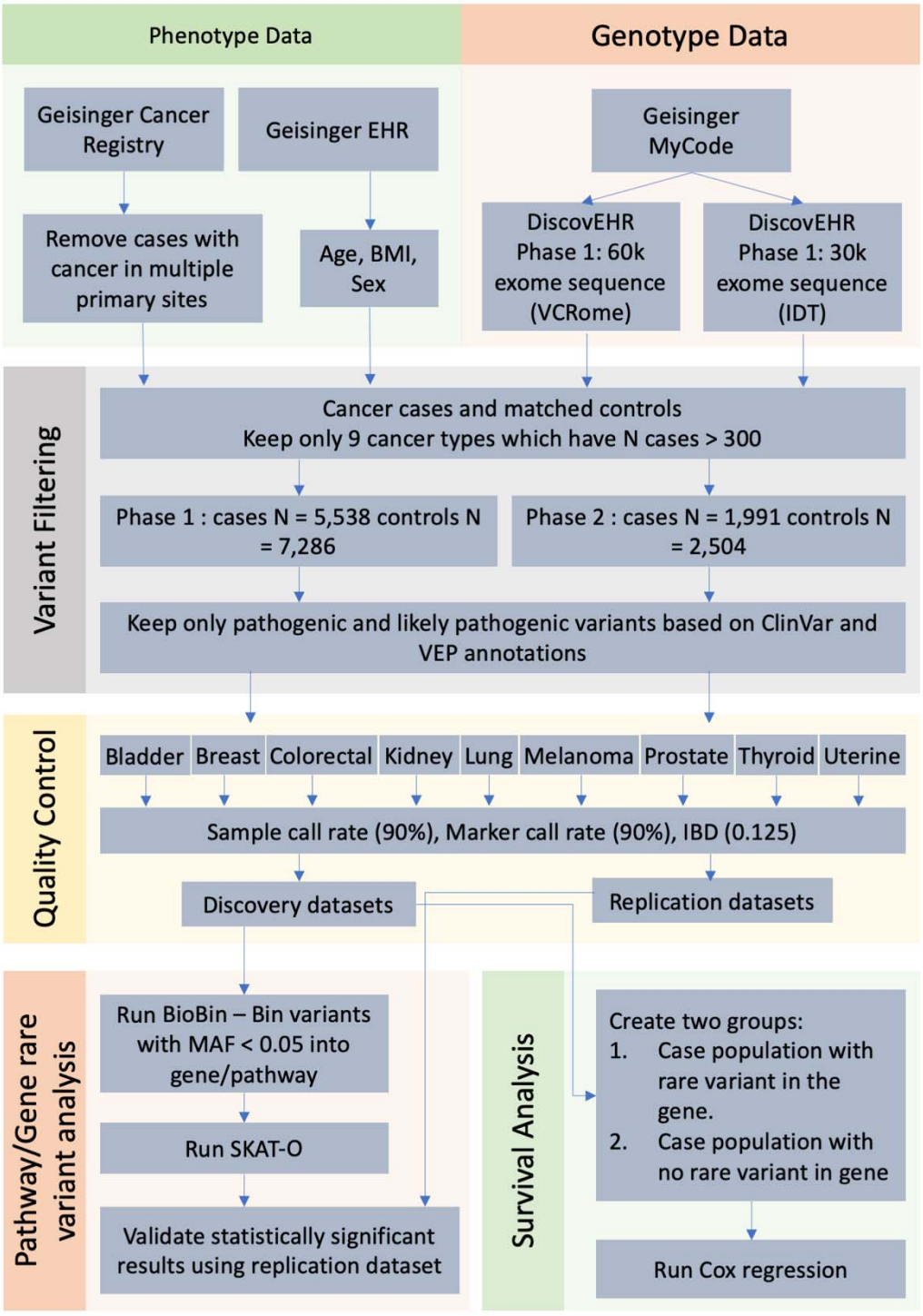


Figure 1. Schematic overview of Pan-caner analysis. The phenotype data was obtained from the Geisinger cancer registry and EHR and the genotype data was obtained from DiscovEHR study. Figure shows multiple steps involved in the analysis – variant filtering, quality control, rare variant analysis, and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

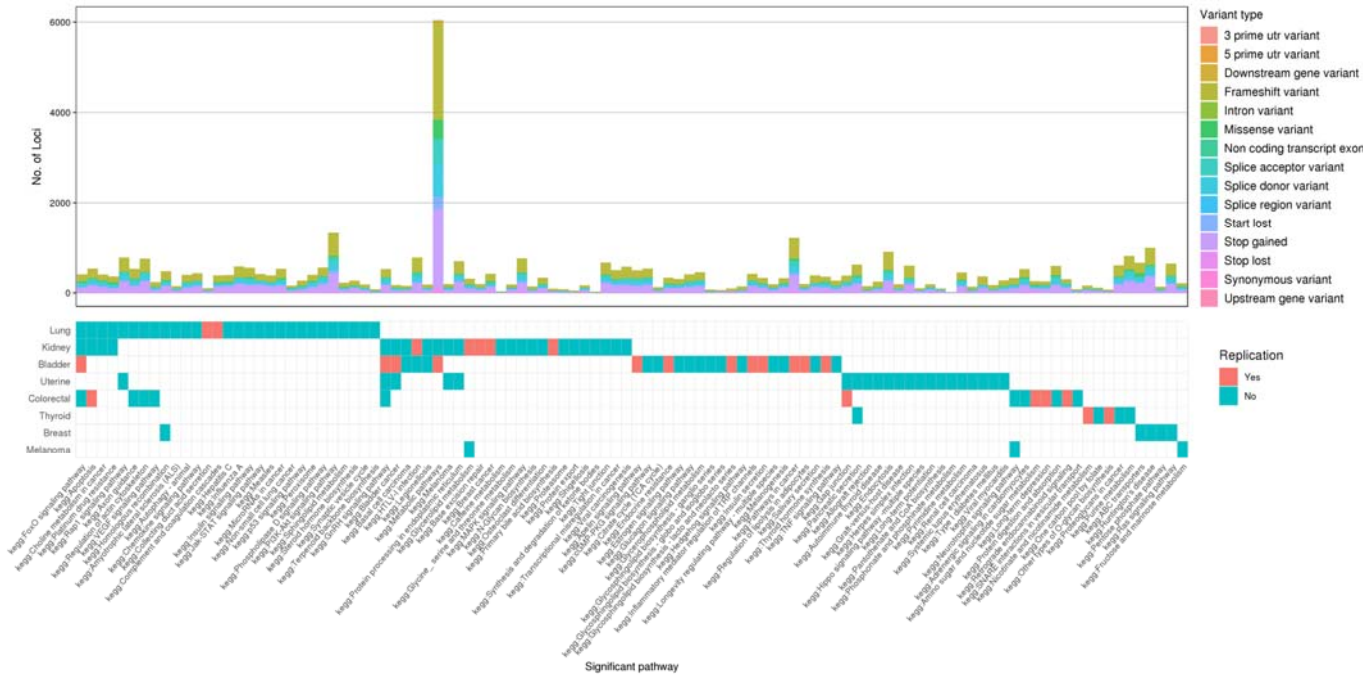


Figure 2. Pathways (x-axis) that were significantly associated with cancer (y-axis) (Bonferroni-corrected $P < 0.05$). The pathways marked as replication 'Yes' were significant in discovery (Bonferroni-corrected $P < 0.05$) and replication (SKAT-O $P < 0.05$) datasets. The top bar plot shows the distribution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

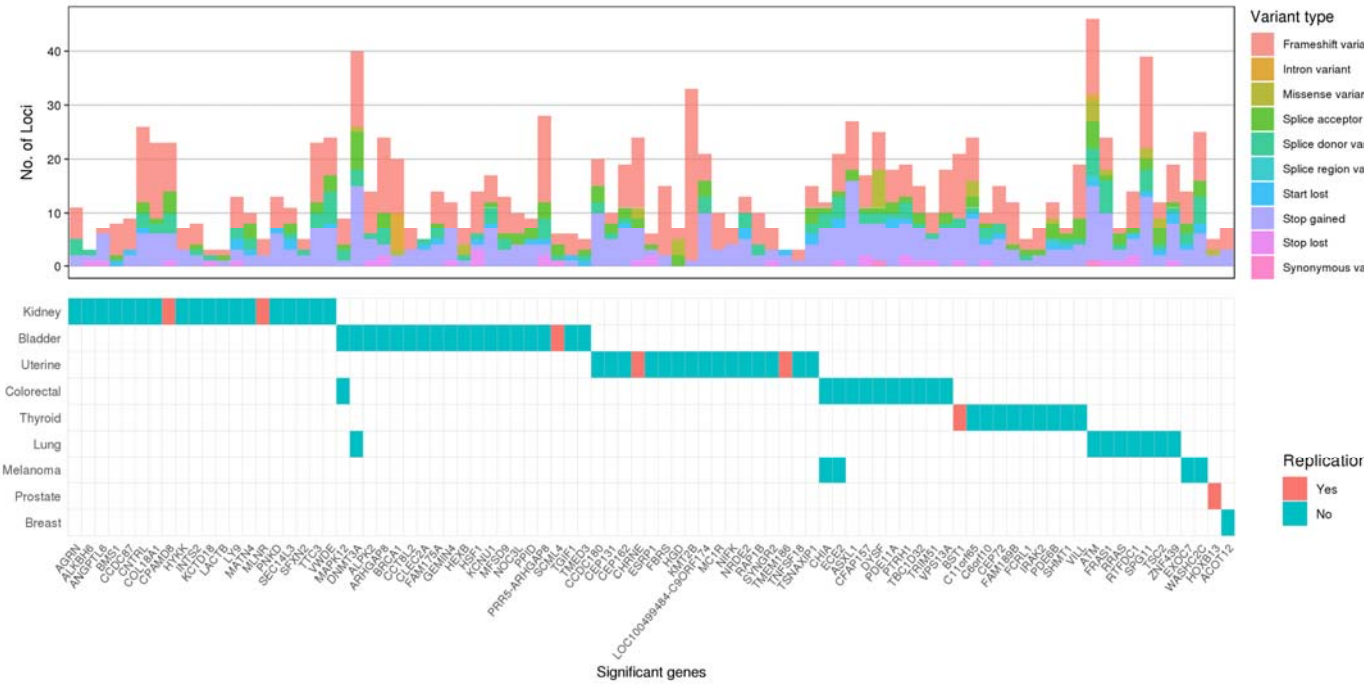


Figure 3. Genes (x-axis) that were significantly associated with cancer (y-axis) using Bonferroni < 0.05. The genes marked as replication 'Yes' were significant in discovery (Bonferroni < 0.05) and replication (association p-value < 0.05) datasets. The top bar plot shows the distribution of variant types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

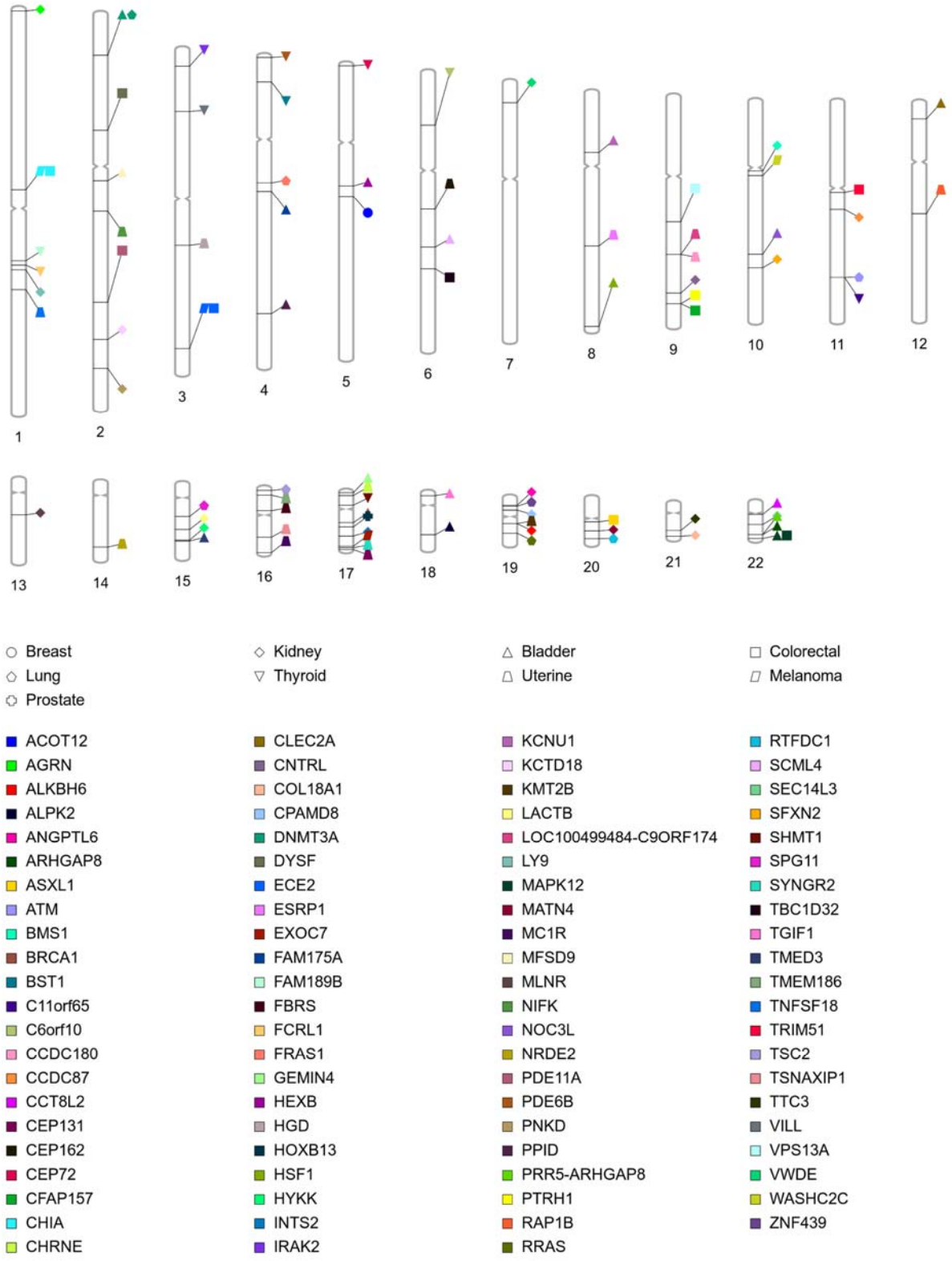
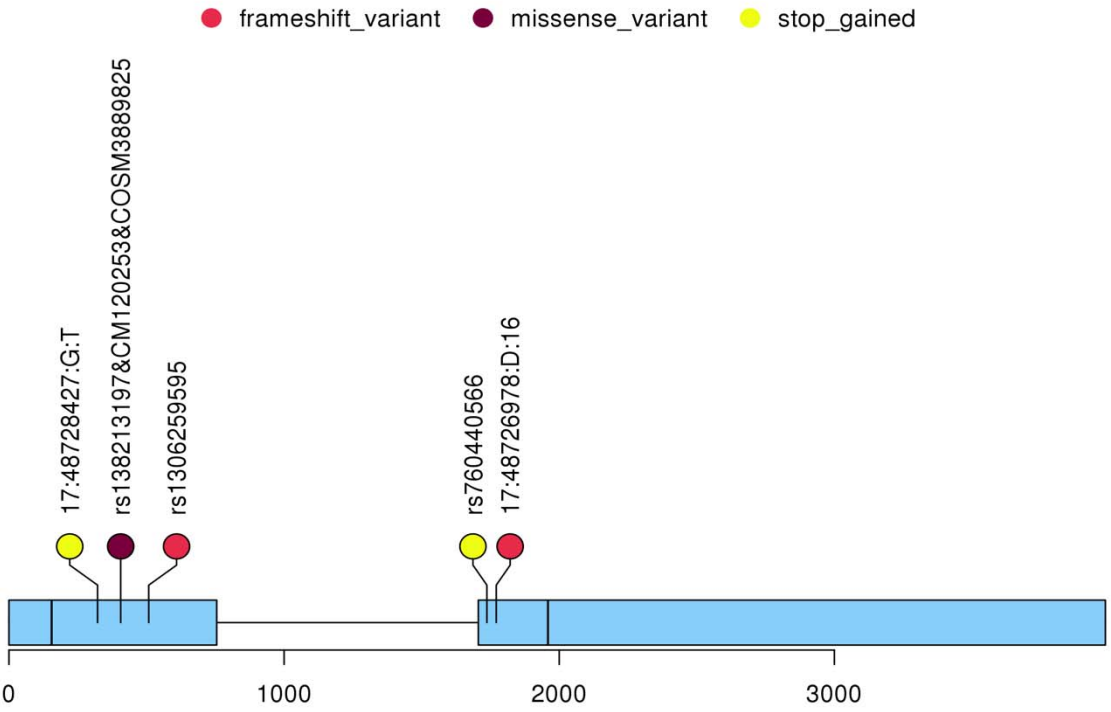


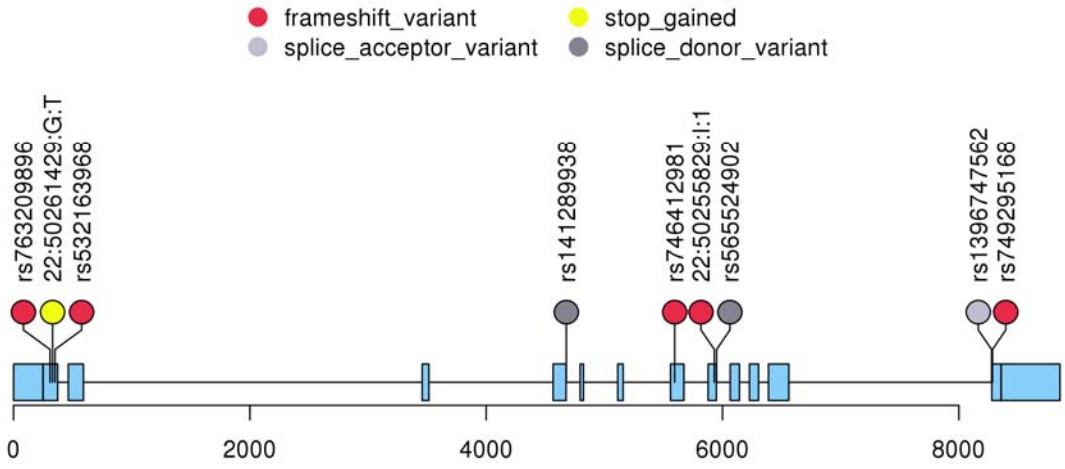
Figure 4. PhenoGram plot of all significant genes across all cancers.

701 A.



702

703 B.



704

C.

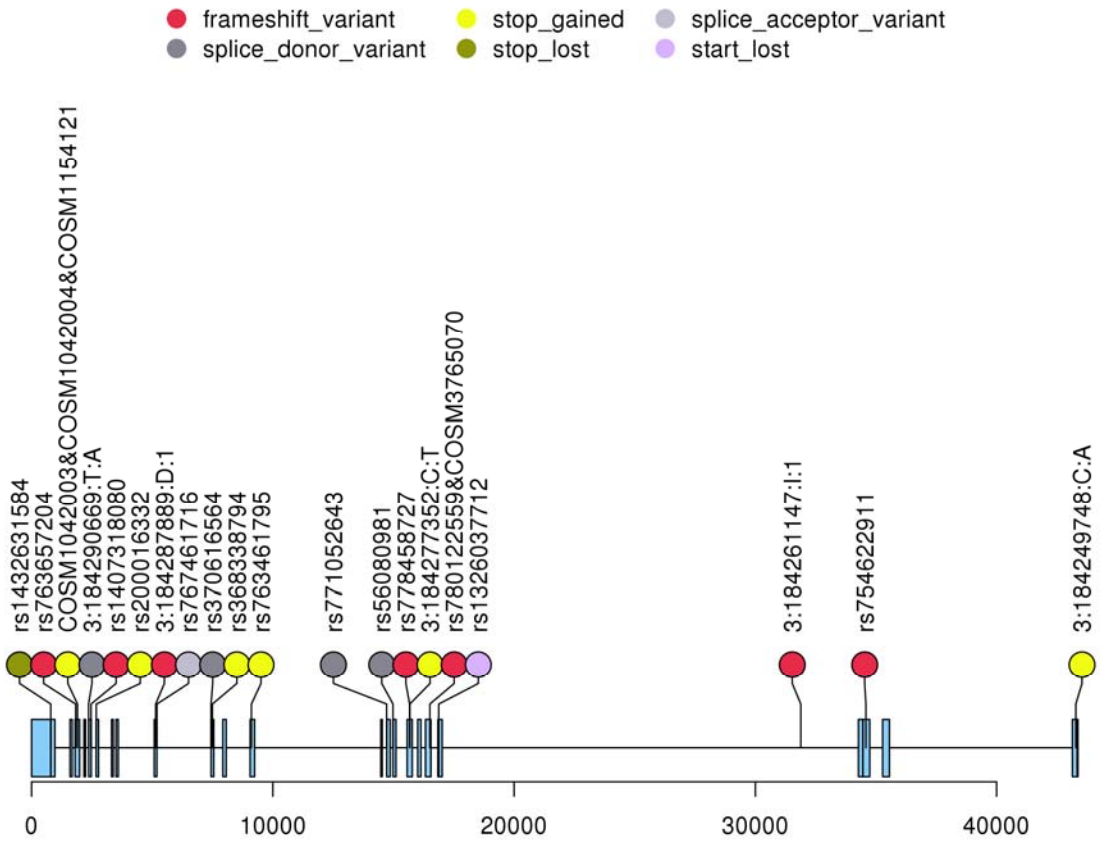


Figure 5. Lollipop plot of genes with all loci binned in them. The color represents different types of variants as assonated by VEP. A. Lollipop plot for HOXB13 gene; B. Lollipop plot for MAPK12 gene; C. Lollipop plot for ECE2 gene.

Figure 6. Carrier frequency of pathogenic and likely pathogenic variants across known oncogenes and tumor suppressors.
The color of the boxes represents the carrier frequency of pathogenic and likely pathogenic variants as denoted by the frequency legend.
The 7 genes significant at TFT p-value < 0.05 are marked with blue or green border.
The 2 genes replicated from Huang et al [14] are marked with green border.

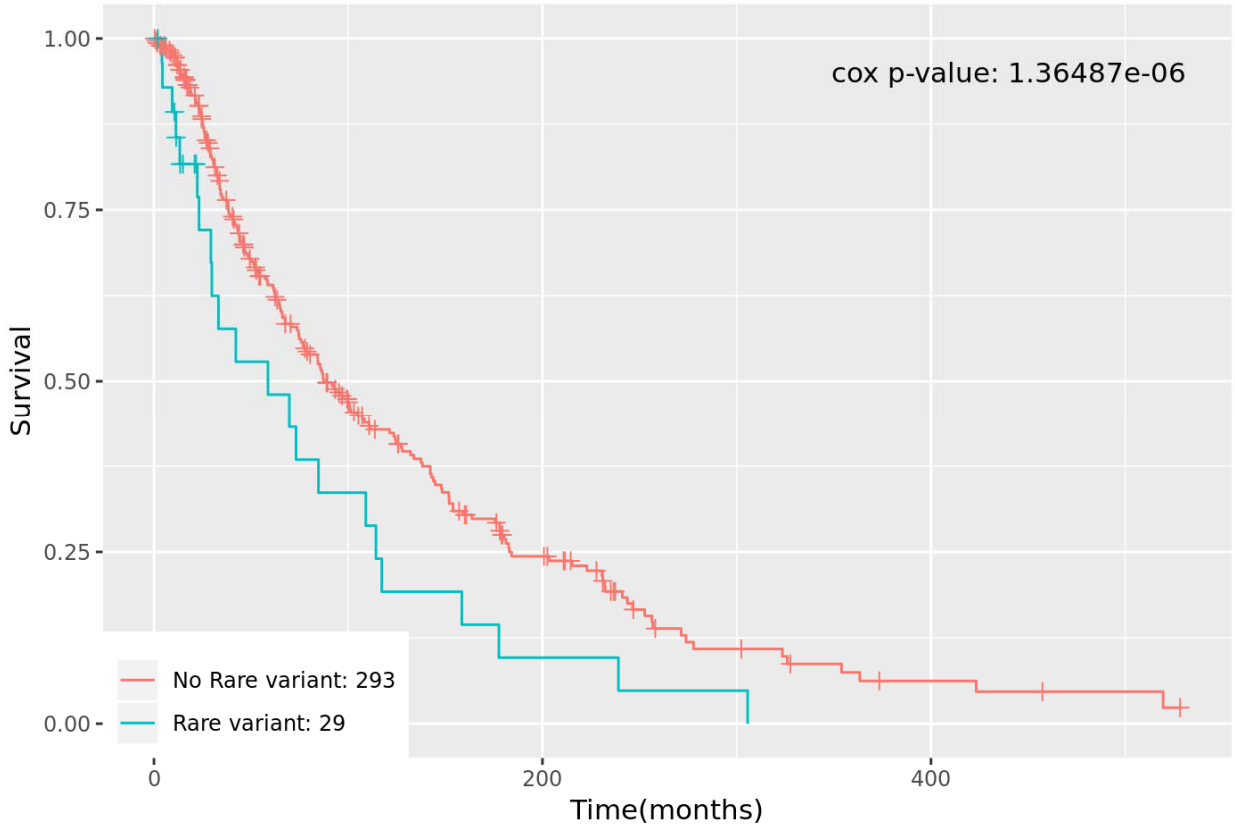


Figure 7. Kaplan Meier survival plot for gene *PCDHB8* which was significantly associated with survival in bladder cancer. *PCDHB8* was also significantly associated with survival in bladder cancer in TCGA dataset.

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 Combined	1	5538	61.85 +/- 12.71	51.97%	31.12 +/- 7.18	4292	1244
	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/m^2 +/- 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 *P*	Bonferroni *P*	N locus	MAC Case	MAC Control	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

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-C p-value
Kidney	*MLNR*	13: 49220338-49222377	5	6	39	1.07E-07	3.70E-04	3	7	48	7.89E-
Kidney	*CPAMD8*	19: 16892947-17026818	23	9	75	2.24E-07	7.70E-04	14	3	43	1.29E-
Uterine	*CHRNE*	17: 4801064-4806369	17	9	95	7.92E-07	2.70E-03	17	8	27	1.29E-
Prostate	*HOXB13*	17: 48724763-48728749	5	19	37	2.86E-06	1.06E-02	2	8	14	5.54E-
Bladder	*SCML4*	6: 107697297-107845959	6	5	116	2.26E-07	7.70E-04	4	3	18	1.13E-
Thyroid	*BST1*	4: 15701866-15774178	21	5	41	2.36E-07	8.20E-04	13	4	49	2.56E-
Uterine	*TMEM186*	16: 8795180-8797648	3	13	101	8.66E-08	3.00E-04	2	10	47	4.49E-

N locus: Total number of genomic loci binned in gene.
MAC Case: Total minor allele count of variants in gene in case population.
MAC Control: Total minor allele count of variants in gene in control population.

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	*MAPK12*	22: 50252901-50261810	7	6	42	1.18E-06	4.02E-03
Colorectal	*MAPK12*	22: 50252901-50261810	9	8	42	1.18E-05	4.08E-02
Colorectal	*ECE2*	3: 184276011-184293031	17	4	17	7.23E-06	2.50E-02
Melanoma	*ECE2*	3: 184276011-184293031	20	6	17	2.93E-06	1.04E-02
Lung	*DNMT3A*	2: 25232961-25342590	30	10	28	6.62E-10	2.00E-06
Bladder	*DNMT3A*	2: 25232961-25342590	25	6	27	2.21E-06	7.53E-03
Colorectal	*CHIA*	1: 111290852-111320566	9	9	70	4.14E-10	1.00E-06
Melanoma	*CHIA*	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	Permutation p-value
Thyroid	*NAA38*	10	430	1.67E-05	9.80E-03	9.80E-03	2.98E-03
Melanoma	*THNSL1*	18	712	1.75E-06	1.60E-03	1.60E-03	8.40E-04
Uterine	*SAXO2*	17	370	7.56E-05	1.95E-02	3.89E-02	6.20E-04
	DCHS2[^]	19	368	9.71E-07	5.00E-04	5.00E-04	1.50E-04
	DBF4B	13	374	1.77E-04	3.04E-02	9.15E-02	5.84E-03
Bladder	*PCDHB8*[^]	29	293	1.36E-06	6.00E-04	6.00E-04	4.50E-04
Breast	*ANO5*	15	1199	1.75E-04	4.92E-02	2.46E-01	1.56E-03
	HOGA1	10	1204	4.23E-05	1.98E-02	5.94E-02	2.39E-03
	CSH2	22	1192	3.71E-07	9.80E-03	9.80E-03	2.98E-03
	ATXN3	10	1204	3.89E-05	1.60E-03	1.60E-03	8.40E-04
	FAM186A	35	1179	1.64E-04	1.95E-02	3.89E-02	6.20E-04
Prostate	*TPTE2*	15	1131	3.25E-05	5.00E-04	5.00E-04	1.50E-04
Kidney	*HLA-G*	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