Genetic Analysis of Functional Rare Germline Variants across Nine Cancer Types from an Electronic Health Record Linked Biobank

Manu Shivakumar[1,2], Jason E. Miller[3,4], Venkata Ramesh Dasari[5], Yanfei Zhang[6], Ming Ta Michael Lee[6], David J. Carey[7], Radhika Gogoi[5], and Dokyoon Kim[1,2,4]; for the DiscovEHR collaboration

ABSTRACT

Background: Rare variants play an essential role in the etiology of cancer. In this study, we aim to characterize rare germline variants that impact the risk of cancer.

Methods: 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 variants in 5,538 patients with cancer and 7,286 matched controls in a discovery set and 1,991 patients with cancer and 2,504 matched controls in a validation set across nine cancer types. Further, The Cancer Genome Atlas (TCGA) germline data were used to replicate the findings.

Results: We identified 133 significant pathway–cancer pairs (85 replicated) and 90 significant gene–cancer pairs (12 replicated). In addition, we identified 18 genes and 3 pathways that were associated with survival outcome across cancers (Bonferroni $P < 0.05$).

Conclusions: In this study, we identified potential predisposition genes and pathways based on rare variants in nine cancers.

Impact: This work adds to the knowledge base and progress being made in precision medicine.

Introduction

Cancer is not a single disease. Even though cancers are alike in some ways, they can start in different parts of the body and the process by which they grow and spread can be very different. Cancer is caused by inherited germline variants and acquired somatic mutations. A recent twin study showed approximately 33% heritability (proportion of variance in a trait due to genetic differences among individuals) of cancer across 23 cancer types (1). To date, many genome-wide association studies (GWAS) have been conducted. Many variants and genes have been discovered, which are associated with various cancer types. However, a large portion of inherited genetic factors that result in carcinogenesis is still unknown. For instance, all variants discovered to date explain just one-third of total genetic contribution in prostate cancer and 30% in breast cancer (1). GWAS has been used as a fundamental tool to identify common variants [variants with minor allele frequency (MAF) ≥ 0.01] associated with traits, but inferring causal mechanisms have proved to be challenging (2). Loci are often identified but studies are not necessarily able to determine functional variants or genes that contribute to risk for disease.

Because common variants discovered to be associated with multiple cancers have only modest effect size, the missing heritability could be further explained by rare variants (variants with MAF < 0.01). Moreover, rare variants have been known to contribute to various complex diseases including cancer (3, 4). The aggregation of rare variants in a gene can lead to loss of function of the gene or change in expression (5). Similarly, because 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. 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 (4). To that effect, binning the rare variants into genes and pathways would help us increase statistical power to detect associations and infer biological mechanisms (6). Rare variant analyses often use variants likely to impact protein function such as in-frame insertion or deletion (indel), protein-truncating and canonical splice site (7). Annotating and filtering the variants to only include variants that change a protein's function is essential to reduce contamination from neutral background variation (7).

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, which is available for use in broad research (8). 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. The sequenced data are 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. Thus, large genetic data and linked rich clinical resources provide ample opportunities for genetic studies to discover variants associated with many phenotypes

[1]Biomedical & Translational Informatics Institute, Geisinger, Danville, Pennsylvania. [2]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania. [3]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania. [4]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, Pennsylvania. [5]Weis Center for Research, Geisinger Clinic, Danville, Pennsylvania. [6]Genomic Medicine Institute, Geisinger, Danville, Pennsylvania. [7]Department of Molecular and Functional Genomics, Geisinger, Danville, Pennsylvania.

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Current address for R. Gogoi: Karmanos Cancer Institute, Detroit, Michigan.

Corresponding Authors: Dokyoon Kim, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104-6116. Phone: 215-573-5336; Fax: 215-573-3111; E-mail: dokyoon.kim@pennmedicine.upenn.edu; and Radhika Gogoi, Karmanos Cancer Institute, 4100 John R., HP07GO, Harper Professional Building, Detroit, MI 48201; Phone: 313-576-9438; E-mail: radhikagogoi@wayne.edu

Cancer Epidemiol Biomarkers Prev 2021;30:1681–8

doi: 10.1158/1055-9965.EPI-21-0082

including cancer. Moreover, the sharing of genetic findings from the research 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.

Materials and Methods

Study population

As part of MyCode community initiative blood samples from Geisinger patients were collected and stored in a system wide biorepository. In phase I, whole exome sequence was generated for 60,000 samples and later in phase II further 30,000 samples were sequenced at Regeneron Genetics center. The study populations for the discovery set (∼60,000) and validation set (∼30,000) are from Geisinger patients who have exome sequencing data available and were consented to participate in the MyCode community initiative. The cases were identified from the cancer registry and were classified into different cancers using ICD-O site codes as defined in Supplementary Table S1. The cases with cancer at multiple primary sites were removed. Only cancer types with at least 300 cases in the discovery set were included; other cancers were excluded from the study due to low number of cases, which can result in a higher type 1 error and lower statistical power in association studies (9). Age and BMI-matched samples who did not have any ICD9/ICD10 code in their encounters diagnoses, inpatient hospitalization-discharge diagnoses or problem-list were selected as controls for all cancers (Supplementary Materials and Methods). The controls for sex-specific cancers including breast, uterine, and prostate cancer were selected separately to be sex-matched and having the same number of controls as the common control set. Thus, separate datasets for each cancer were created. Further, only samples with European ancestry were retained in the study population.

TCGA germline data were used as an external independent validation set to replicate the findings from the DiscovEHR dataset. The TCGA germline data only contains cancer samples ($N = 10{,}389$), so the data were combined with the control dataset from the discovery dataset (Supplementary Materials and Methods).

All the data used in this study were de-identified, thus were exempt from the IRB review. We received approvals from Geisinger MyCode Governing Board and TCGA to conduct the study.

Sequencing and quality control

All the MyCode samples were sequenced as part of the DiscovEHR project at the Regeneron Genetic Center. Initially, approximately 60,000 samples (used as the discovery dataset) were sequenced using NimbleGen probe target-capture (SeqCap VCRome) and further a separate batch of approximately 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 (10). Further detailed description of sequencing is available at Shivakumar and colleagues (11) and Supplementary Materials and Methods. The phase I and phase II sequencing, variant calling, and quality control (QC) were performed separately. The TCGA dataset was generated by TCGA PanCanAtlas Germline Working Group. The variant calls from GATK, VarScan2, and Pindel were merged and filtered resulting in 10,389 samples that passed stringent QC criteria. More detailed information on the QC criteria and method is available at Haung and colleagues (5).

Variant annotation and filtering

Variants were annotated using VEP (12) and NCBI ClinVar (13). The variants were filtered to reduce the neutral background variation. VEP is one of the tools that can be used to annotate the consequence of the variant, which can be further used to filter the variants. ClinVar is a database of relationships among human variations and phenotypes, with supporting evidence. The variants that satisfy at least one of the following conditions were retained:

1. Annotated as impact "HIGH" using VEP.
2. Annotated as pathogenic or likely pathogenic with at least 1 star in ClinVar.

Because there is supporting evidence for the variants to be classified as "pathogenic" or "likely pathogenic" in ClinVar, they were included even if they did not fall in the VEP filtration criteria (HIGH).

Gene-based rare variant association

All the variants were binned using BioBin (6, 14, 15), which uses precompiled knowledge in a LOKI (6, 14) database containing information from various data sources including Entrez and KEGG. Only variants with MAF < 0.01 were considered rare and the rest of the variants were filtered out. In addition, bins with total minor allele count (MAC) less than 20 variants were filtered out. Further, the binned variants were weighed using Madsen-browning weights (16). The statistical association tests were run using SKAT-O implemented as R package (17), and were adjusted using age, sex, BMI and first four principal components (PC) as covariates. Sex was not adjusted in breast, prostate, and uterine cancers. Some previous studies have used PCs to adjust for population substructure, as it could lead to inflated type 1 error rates (18, 19). The PCs were calculated using EIGENSOFT (20), 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. The Bonferroni correction was performed separately for each cancer type as they have different total numbers of tests, which is defined as the number of bins tested for association (Supplementary Table S2).

Pathway-based rare variant association

The same method described for gene-based rare variant association analysis was used except the variants were binned into KEGG (21) pathways derived from LOKI (6, 14) in place of genes. The LOKI database used in this study was created on April 15, 2017. The pathway information was integrated into the database using KEGG API. The rare variants were binned into 317 KEGG pathways. The Bonferroni correction was performed separately for each cancer and 317 was considered as total number of tests for the correction, as none of the pathways had zero variants binned in them. Because there may be correlation structure of genes in a given pathway, permutation testing was also performed by permuting the phenotype 100 times to show results were not by chance. (22, 23).

Survival analysis

Survival analysis was performed using Cox regression adjusting for age, sex, and BMI. Sex was not included as a covariate in breast, prostate, and uterine cancer. Specifically, the weighted burden of each patient for the bin (gene/pathway) was obtained from BioBin bin-phe output files, Cox regression was run on the bin adjusting for covariates. Survival analysis was performed on each cancer using gene-based bins and pathway-based bins. The bins with a number of samples with rare variant burden <10 were excluded due to low sample size for survival analysis. Further, survival P values were adjusted for multiple testing using Bonferroni correction. The Bonferroni correction was performed separately for each cancer and the number of bins on which

Cox regression was conducted was considered as the total number of tests for the correction.

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 nearly 90,000 sequenced samples from the DiscovEHR study. The samples were sequenced in two phases using different platforms as described in the Materials and Methods section. In phase I, 5,538 patients with cancer across nine cancers and 7,286 matched controls were identified as discovery dataset and in phase II, 1,991 patients with cancer and 2,504 matched controls were identified as replication dataset. Cancer patient IDs retrieved from the cancer registry were classified into particular cancers using International Classification of Diseases for Oncology (ICD-O) codes (Supplementary Table S1). After classifying the patients with cancer to their respective cancers, only nine cancers had more than 300 samples in the discovery set, including bladder, breast, colorectal, kidney, lung, melanoma, prostate, thyroid, and uterine cancer.

Further, TCGA germline data obtained from Genomic Data Commons (GDC) was also used to replicate the findings from the discovery dataset. The TCGA germline data contains only cancer samples, so the dataset was merged with controls from the discovery dataset (Supplementary Materials and Methods). The distribution and basic demographics of controls and cases across the datasets are shown in Supplementary Table S2.

Variant filtering based on functional annotation and scores improves power and has been successfully used in many association studies (5, 24). In this study, the variants from whole exome sequence data were annotated using variant effect predictor (VEP; ref. 12) and ClinVar (13). Subsequently, only the variants categorized as pathogenic or likely pathogenic (PLP) based on the annotations were retained for further analysis. Additionally, all common variants were removed, and only rare variants (MAF < 0.01) were retained. The number of rare variants available after filtering in each cancer cohort is listed in Supplementary Table S3. The statistical power analysis was performed using SKAT package in R (Supplementary Materials and Methods; Supplementary Table S4). **Figure 1** shows the schematic overview of the analysis.

Pathway-based rare variant analysis

In this study, PLP rare variants were binned into KEGG pathways derived from Library of Knowledge Integration (LOKI) using BioBin (6,14,15,). An association test was performed on the bin to determine if the gene/pathway is significantly associated with the phenotype. We found 133 significant pathway–cancer pairs (106 unique pathways) that were significantly associated across all cancers after adjusting for multiple testing using Bonferroni correction. Of the 133 significant pathway–cancer pairs in the discovery dataset, 27 were replicated in the replication dataset and 68 in the TCGA dataset (**Fig. 2**; Supplementary Table S5). Further, 85 pairs that were replicated in either the replication dataset or the TCGA dataset are listed in Supplementary Table S6. In addition, 21 pathways were found to be associated with multiple cancers, 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 (Supplementary Tables S7 and S8). Further, none of the pathways were found to be significant (permutation $P < 0.05$) with permutation testing (Supplementary Table S5). The number of significant pathways across datasets is summarized in **Table 1**.

Gene-based rare variant analysis

All PLP rare variants were binned into genes defined by Entrez annotations derived from LOKI using BioBin. The total number of genes that the variants were binned across all cancers is listed in Supplementary Table S9. The bar plot in **Fig. 3** shows the total number of loci binned for a given gene and variant types as annotated by VEP. In total, there were 90 gene–cancer pairs (86 unique genes, Supplementary Table S10) that were identified significantly associated with a specific cancer (Bonferroni $P < 0.05$) in the discovery dataset. Of the 90 cancer–gene pairs 12 were replicated ($P < 0.05$) in the replication or TCGA dataset (**Fig. 3**). In addition, four genes, including *MAPK12, ECE2, DNMT3A*, and *CHIA*, were significantly associated with multiple cancers (Supplementary Table S11). The PhenoGram plot in Supplementary Fig. S1 shows all the genes found to be significantly associated across all cancers. In addition, the lollipop plots in Supplementary Figs. S2 and S3 illustrate 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 identifier.

Significant genes within significant pathways

Many significant genes discovered in this study were found within significantly associated pathways. Of the 90 significant gene–cancer pairs, 15 were found within 37 significantly associated pathways in given cancers (Supplementary Table S12). Genes *DNMT3A* and *ATM* are part of the KEGG pathway "MicroRNAs in cancer." The pathway and both genes were significantly associated with lung cancer. Similarly, "Metabolic pathways" with genes *HEXB* and *DNMT3A* was significant in bladder cancer. The rest of the pathways only had one significant gene. Some of the genes also spanned into multiple significant pathways—*ATM* in six pathways in lung cancer, *BST1* in two pathways in thyroid cancer, *HEXB* and *KCNU1* in two pathways in bladder cancer, *MAPK12* in five pathways in bladder cancer, and six pathways in colorectal cancer, *RAP1B* in four pathways in uterine cancer, and *RRAS* in five pathways in lung cancer.

Survival analysis

The survival analysis was carried out using all the gene bins and pathway bins in the discovery dataset, which had more than 10 samples with any rare variant burden (Supplementary Table S13). We identified 18 genes and 3 pathways that were significantly associated with survival across nine cancers at Bonferroni $P < 0.05$. Gene *OR6C70* was replicated in prostate cancer using the replication dataset and *FAM166A* was replicated in thyroid cancer in the TCGA dataset. The statistics are listed in Supplementary Table S14 for significant genes and Supplementary Table S15 for significant pathways.

Discussion

In this study, we performed exome-wide rare variant analysis across nine cancers using a cohort of 7,449 cancer cases and 9,792 controls from a single hospital system with further independent external validation using germline TCGA exome sequence data. We filtered and retained only PLP variants to reduce the background noise and conducted rare variant association analysis and rare variant burden survival analysis in all three cohorts. A total of 133 significant pathway-cancer pairs (85 replicated) and 90 significant gene–cancer pairs

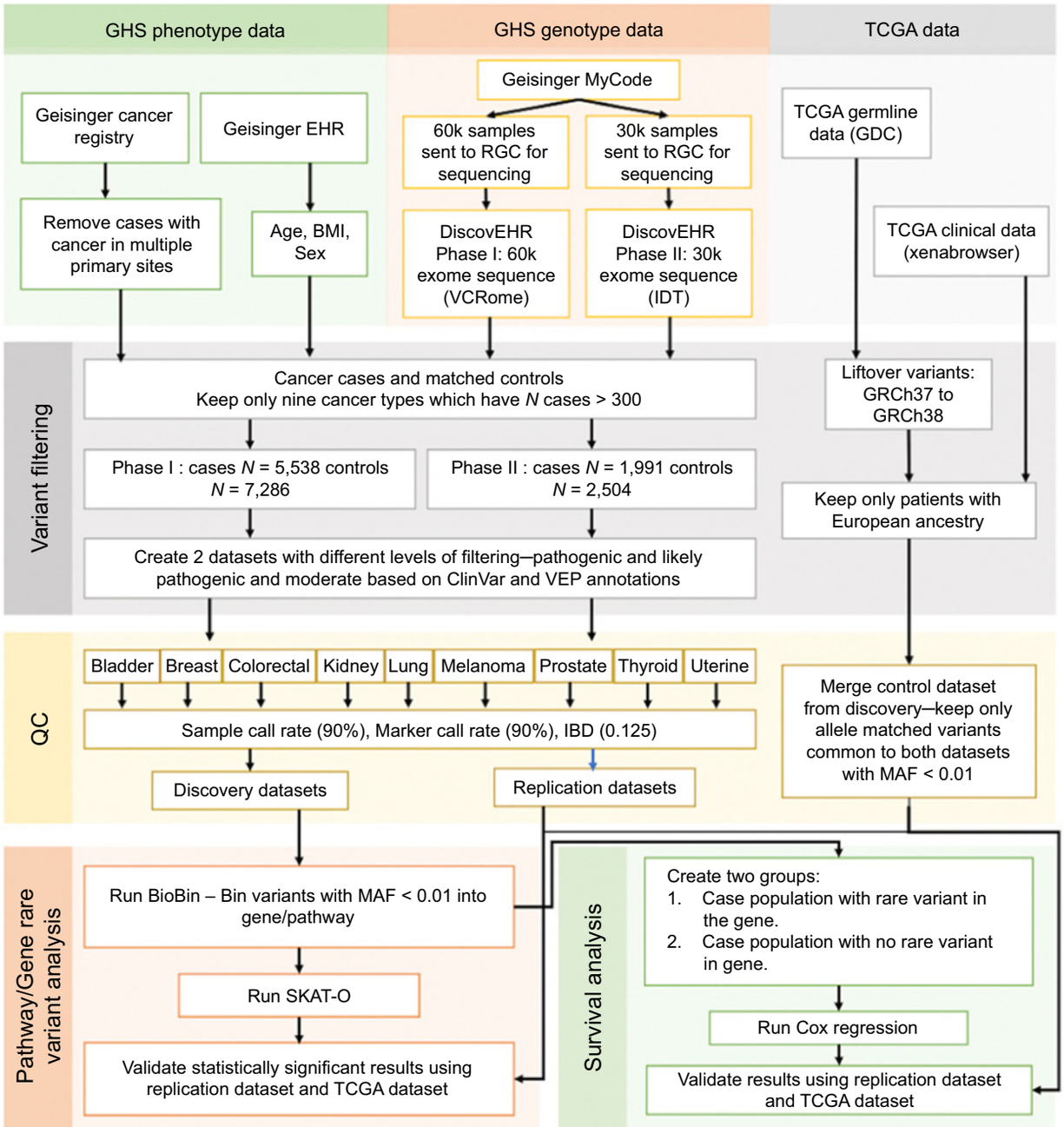

Figure 1.

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

(12 replicated) were identified as associated with cancers. Furthermore, 21 pathways and four genes were associated with multiple cancer types. In addition, we identified 18 genes and three 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" (25), "bladder cancer," "FoxO signaling pathway" (26), "metabolic pathways," "gap junction" (27), "apoptosis," "base excision repair," "melanoma," "choline metabolism in cancer," and "basal cell carcinoma." 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 (28). Further studies on the "HTLV-I infection" pathway could elucidate the role of germline variants in cancers. Another pathway, the "Hippo signaling pathway" was found

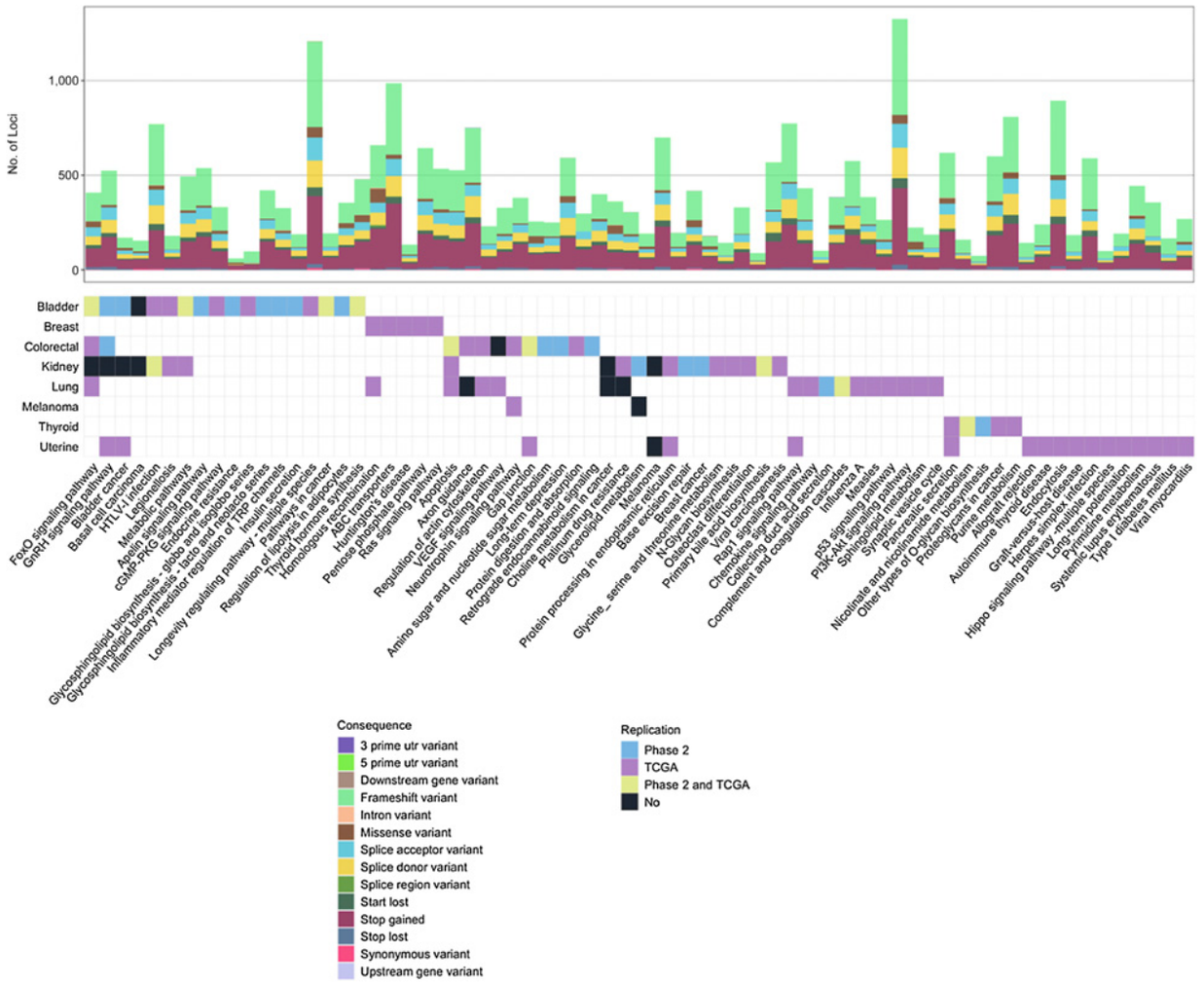

Figure 2.

Waterfall plot with pathways (*x*-axis) that were significantly associated with cancer (*y*-axis) (Bonferroni $P < 0.05$) and were replicated (SKAT-O $P < 0.05$) in either replication (phase II) or TCGA. The pathways that were significant in multiple cancers but not replicated are also included. They are marked in "black" as replication "No." The top bar plot shows the distribution of variant types as annotated by VEP across each pathway.

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 (29). The Hippo pathway also interacts with the PI3K/AKT pathway, which is commonly involved in cancer (29).

A number of previous studies have shown *HOXB13* to be associated with prostate cancer (30, 31), and in this study as well, *HOXB13* was found to be associated with prostate cancer in the discovery dataset and was replicated in phase II dataset and TCGA dataset. 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 (32). *CPAMD8* is known to be substantially expressed in the kidney (32, 33) 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

Table 1. Summary of results for pathway and gene based rare variant analysis.

	Discovery[a]	Replication[b]	TCGA[b]	Replication + TCGA[b]
Gene	90	7 (7.78%)	6 (6.67%)	12 (13.33%)
Pathway	133	27 (20.30%)	68 (51.13%)	85 (63.91%)

Note: The values in the brackets are the percentage of genes/pathways that were replicated.
[a]Significant Bins (gene/pathway) Bonferroni <0.05.
[b]Significant Bins (gene/pathway) SKAT-O <0.05.

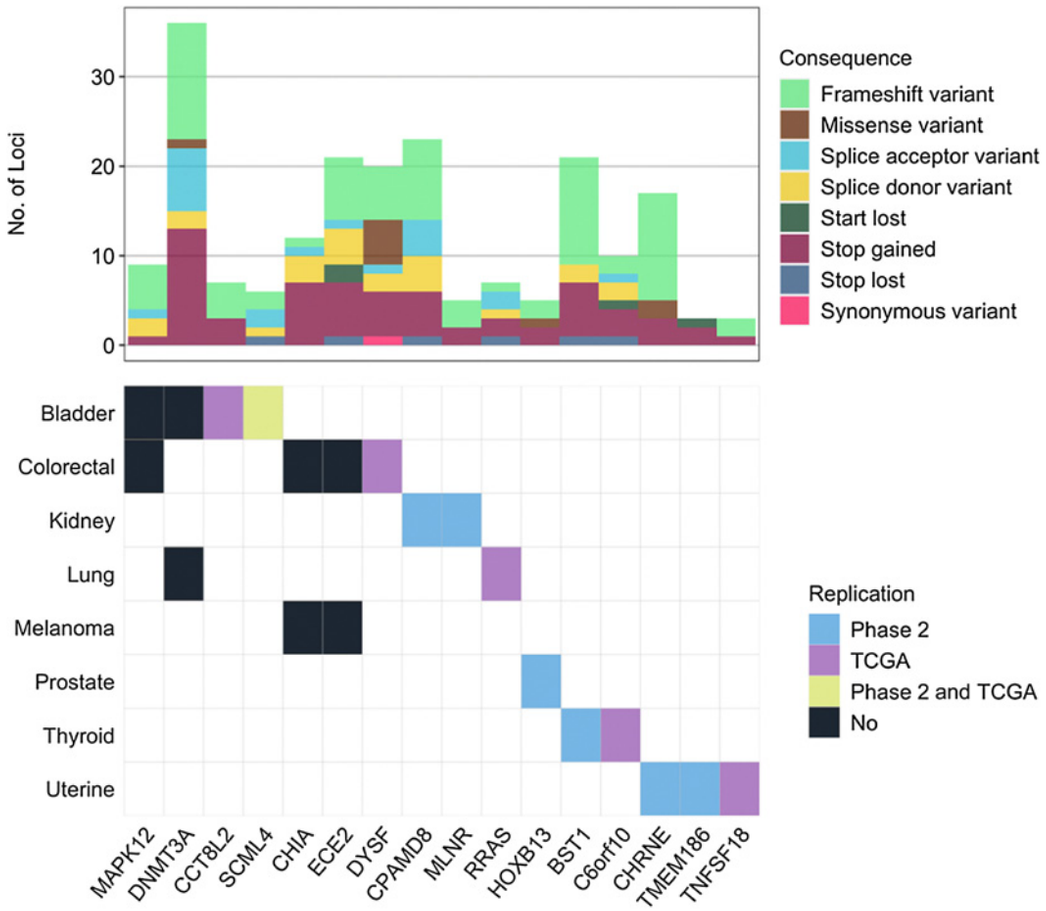

Figure 3.

Waterfall plot with genes (*x*-axis) that were significantly associated with cancer (*y*-axis; Bonferroni $P < 0.05$) and were replicated (SKAT-O $P < 0.05$) in either replication (phase II) or TCGA. The genes that were significant in multiple cancers but not replicated are also included. They are marked in "black" as replication "No." The top bar plot shows the distribution of variant types as annotated by VEP across each gene.

acetylcholine receptors (nAChR) and *TMEM186* which is a member of the transmembrane protein family. Nicotine, a compound present in cigarettes, mediates cell proliferation and angiogenesis through nicotinic acetylcholine receptors (nAChRs) and its subunits (34). 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 (34, 35). 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 whereas others as oncogenes (36). Further, in bladder cancer, the Putative Polycomb group (PcG) protein gene (*SCML4*), is involved in the regulation of crucial developmental and physiologic processes and is known to promote proliferation and inhibit apoptosis (37). *SCML4* was significantly associated with bladder cancer in replication and TCGA dataset.

Different cancer types share some pathways and genes, which generally include common tumor suppressor genes and oncogenes (5). 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 (38, 39). 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 growth in colorectal cancer and melanoma (40). Further, about 27.8% of the significant pathways had at least one significantly associated gene in that specific cancer. Interestingly, the pathway with the highest number of significant genes was "MicroRNAs in cancer" in lung cancer and "Metabolic pathways" in bladder cancer, both sharing gene *DNMT3A*.

Even though many associations were identified in this study, further studies would be required to elucidate the molecular mechanisms. Depending on the variant filtering criteria, there is possibility of increased false positive- or false-negative rate. The inclusion of a lot of variants which are not causal usually leads to lower true positive rates and higher false-positive rates (41). On the contrary, restricting the number of variants like in PLP variants set, could filter out some loss of function variants which could lead to false negative results. In addition, the participants in the replication cohort were derived from participants who enrolled in the MyCode program more recently than the discovery dataset. Therefore, most of the patients in the replication

set were alive and in some cases like thyroid cancer, where all patients were alive, it was not possible to run survival analysis. In addition, for some of the gene–cancer pairs the power was low in the replication dataset. Furthermore, the TCGA data only contains cancer cases, so we merged it with controls from the discovery dataset. Because the cases and controls were on different sequencing platforms, there might be some batch effects that could increase false-positive rates. 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 an exome-wide rare-variant analysis to find novel genes and pathways associated across nine cancers. We replicated many genes and pathways that were known to be associated with cancers. Some of the significant genes in this study were linked to the pathways that were also significantly associated with cancers, which could potentially aid in understanding the mechanism of gene action. The genes and pathways discovered in this study could eventually be used to screen for high-risk patients with cancer and personalized therapy.

Data availability

The raw data supporting the conclusions of this manuscript will be made available by the authors to any qualified researcher subject to a data use agreement.

Authors' Disclosures

No disclosures were reported.

Authors' Contributions

M. Shivakumar: Data curation, formal analysis, validation, investigation, visualization, methodology, writing–original draft, writing–review and editing. **J.E. Miller:** Investigation, visualization, writing–review and editing. **V.R. Dasari:** Investigation, writing–review and editing. **Y. Zhang:** Investigation, writing–review and editing. **M.T.M. Lee:** Investigation, writing–review and editing. **D.J. Carey:** Investigation, writing–review and editing. **R. Gogoi:** Conceptualization, resources, supervision, investigation, project administration, writing–review and editing. **D. Kim:** Conceptualization, resources, supervision, funding acquisition, investigation, project administration, writing–review and editing.

Acknowledgments

This work was supported by NLM R01 NL012535 and NIGMS R01 GM138597. This project was also funded, in part, under a grant with the Pennsylvania Department of Health (#SAP 4100070267). The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions. We gratefully acknowledge the funding support from Geisinger Medical Center (SRC-075 to R. Gogoi) and Rice Women's Cancer Research Fund (to R. Gogoi and V.R. Dasari). Support for this work was also provided by NHGRI T32HG009495-01 (to J.E. Miller). The funders specifically disclaim responsibility for the study design, data collection, analyses, interpretation, conclusions, and writing of the manuscript.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received January 18, 2021; revised February 15, 2021; accepted June 17, 2021; published first July 8, 2021.

References

1. Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, Scheike T, et al. Familial risk and heritability of cancer among twins in nordic countries. JAMA 2016;315:68–76.
2. Giral H, Landmesser U, Kratzer A. Into the wild: GWAS exploration of noncoding RNAs. Front Cardiovasc Med 2018;5:181.
3. Rahman N. Realizing the promise of cancer predisposition genes. Nature 2014; 505:302.
4. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 2009;19:212–9.
5. Huang K-l, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, et al. Pathogenic germline variants in 10,389 adult cancers. Cell 2018;173:355–70.
6. Moore CB, Wallace JR, Frase AT, Pendergrass SA, Ritchie MD. BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge. BMC Med Genet 2013;6:S6.
7. Povysil G, Petrovski S, Hostyk J, Aggarwal V, Allen AS, Goldstein DB. Rare-variant collapsing analyses for complex traits: guidelines and applications. Nat Rev Genet 2019;20:747–59.
8. Carey DJ, Fetterolf SN, Davis FD, Faucett WA, Kirchner HL, Mirshahi U, et al. The Geisinger MyCode community health initiative: an electronic health record–linked biobank for precision medicine research. Genet Med 2016;18:906.
9. Zhang X, Basile AO, Pendergrass SA, Ritchie MD. Real world scenarios in rare variant association analysis: the impact of imbalance and sample size on the power in silico. BMC Bioinformatics 2019;20:46.
10. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491–8.
11. Shivakumar M, Miller JE, Dasari VR, Gogoi R, Kim D. Exome-wide rare variant analysis from the discovEHR study identifies novel candidate predisposition genes for endometrial cancer. Front Oncol 2019;9:574.
12. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biol 2016;17:122.
13. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 2016;44:D862–D8.
14. Basile AO, Byrska-Bishop M, Wallace J, Frase AT, Ritchie MD. Novel features and enhancements in BioBin, a tool for the biologically inspired binning and association analysis of rare variants. Bioinformatics 2018;34:527–9.
15. Moore CCB, Basile AO, Wallace JR, Frase AT, Ritchie MD. A biologically informed method for detecting rare variant associations. BioData mining 2016;9: 27.
16. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLos Genet 2009;5:e1000384.
17. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet 2012;91:224–37.
18. Zhang Y, Guan W, Pan W. Adjustment for population stratification via principal components in association analysis of rare variants. Genet Epidemiol 2013;37: 99–109.
19. Verma SS, Josyula N, Verma A, Zhang X, Veturi Y, Dewey FE, et al. Rare variants in drug target genes contributing to complex diseases, phenome-wide. Sci Rep 2018;8:4624.
20. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904.
21. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 2016;44: D457-D62.
22. Backes C, Rühle F, Stoll M, Haas J, Frese K, Franke A, et al. Systematic permutation testing in GWAS pathway analyses: identification of genetic networks in dilated cardiomyopathy and ulcerative colitis. BMC Genomics 2014;15: 622.
23. Richardson TG, Timpson NJ, Campbell C, Gaunt TR. A pathway-centric approach to rare variant association analysis. Eur J Hum Genet 2017;25:123–9.
24. Esteban-Jurado C, Vila-Casadesús M, Garre P, Lozano JJ, Pristoupilova A, Beltran S, et al. Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer. Genet Med 2014;17: 131.

25. Gründker C, Emons G. The role of gonadotropin-releasing hormone in cancer cell proliferation and metastasis. Front Endocrinol 2017;8:187.
26. Farhan M, Wang H, Gaur U, Little PJ, Xu J, Zheng W. FOXO signaling pathways as therapeutic targets in cancer. Int J Biol Sci 2017;13:815–27.
27. Aasen T, Mesnil M, Naus CC, Lampe PD, Laird DW. Gap junctions and cancer: communicating for 50 years. Nat Rev Cancer 2016;16:775–88.
28. Ahmadi Ghezeldasht S, Shirdel A, Assarehzadegan MA, Hassannia T, Rahimi H, Miri R, et al. 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). Iran J Basic Med Sci 2013;16:179–95.
29. Wang C, Gu C, Jeong KJ, Zhang D, Guo W, Lu Y, et al. YAP/TAZ-mediated upregulation of GAB2 leads to increased sensitivity to growth factor–induced activation of the PI3K pathway. Cancer Res 2017;77:1637.
30. Xu J, Lange EM, Lu L, Zheng SL, Wang Z, Thibodeau SN, et al. HOXB13 is a susceptibility gene for prostate cancer: results from the International Consortium for Prostate Cancer Genetics (ICPCG). Hum Genet 2013;132:5–14.
31. Ewing CM, Ray AM, Lange EM, Zuhlke KA, Robbins CM, Tembe WD, et al. Germline mutations in HOXB13 and prostate-cancer risk. N Engl J Med 2012; 366:141–9.
32. Li Z-F, Wu X-h, Engvall E. Identification and characterization of CPAMD8, a novel member of the complement 3/α2-macroglobulin family with a C-terminal Kazal domain. Genomics 2004;83:1083–93.
33. Porta-Pardo E, Garcia-Alonso L, Hrabe T, Dopazo J, Godzik A. A Pan-cancer catalogue of cancer driver protein interaction interfaces. PLoS Comput Biol 2015;11:e1004518.
34. Singh S, Pillai S, Chellappan S. Nicotinic acetylcholine receptor signaling in tumor growth and metastasis. J Oncol 2011;2011:456743.
35. Felix AS, Yang HP, Gierach GL, Park Y, Brinton LA. Cigarette smoking and endometrial carcinoma risk: the role of effect modification and tumor heterogeneity. Cancer Causes Control 2014;25:479–89.
36. Schmit K, Michiels C. TMEM proteins in cancer: a Review. Front Pharmacol 2018;9:1345.
37. Wang W, Qin J-J, Voruganti S, Nag S, Zhou J, Zhang R. Polycomb Group (PcG) proteins and human cancers: multifaceted functions and therapeutic implications. Med Res Rev 2015;35:1220–67.
38. Slattery ML, Lundgreen A, Wolff RK. MAP kinase genes and colon and rectal cancer. Carcinogenesis 2012;33:2398–408.
39. Zhang W, Xu J. DNA methyltransferases and their roles in tumorigenesis. Biomark Res 2017;5:1.
40. Grant K, Loizidou M, Taylor I. Endothelin-1: a multifunctional molecule in cancer. Br J Cancer 2003;88:163–6.
41. Lin W-Y. Beyond rare-variant association testing: pinpointing rare causal variants in case-control sequencing study. Sci Rep 2016;6: 21824.